Argumentative Zoning:
Information Extraction from Scientific Text

Simone Teufel

PhD

University of Edinburgh

1999

Acknowledgements

Let me tell you, writing a thesis is not always a barrel of laughs—and strange things can happen, too. For example, at the height of my thesis paranoia, I had a recurrent dream in which my cat Amy gave me detailed advice on how to restructure the thesis chapters, which was awfully nice of her. But I also had a lot of human help throughout this time, whether things were going fine or beserk.

Most of all, I want to thank Marc Moens: I could not have had a better or more knowledgable supervisor. He always took time for me, however busy he might have been, reading chapters thoroughly in two days. He both had the calmness of mind to give me lots of freedom in research, and the right judgement to guide me away, tactfully but determinedly, from the occasional catastrophe or other waiting along the way. He was great fun to work with and also became a good friend.

My work has profitted from the interdisciplinary, interactive and enlightened atmosphere at the Human Communication Centre and the Centre for Cognitive Science (which is now called something else). The Language Technology Group was a great place to work in, as my research was grounded in practical applications developed there.

Jean Carletta helped me design the annotation experiment and interpret the results. Her keen eye and sharp mind helped me nip many errors in the bud. I have also enjoyed many exciting discussions with Chris Brew, about statistics, scientific writing and many other things. Katja Markert and Michael Strube read and commented on versions of chapters.

My annotators, Vasilis Karaiskos, Anne Wilson and Anders Bowers, did meticulous work; their critical comments on the task, the guidelines, and their observations about the articles were extremely valuable. Thanks also to the subjects who participated in the smaller annotation experiment.

David McKelvie and Claire Grover from the Language Technology Group expertly helped me with problems to do with XML encoding and transformation, fsgmatch grammar debugging, cascading style sheets and the like. Andrei Mikheev also helped me with statistical and practical problems. I was often thankful for his programs which have proven very useful for my task. Frank Keller and David Sterrat singlehandedly solved the one or two LaTeXproblems I might have had :-).

I was also blessed with the finest crowd of friends in Edinburgh: Frank and

Mirella, Katja, Vasilis, Janet, Claire, Marc, Scott, Elina, Mercè, Karen, Ash, Line, Zelal, Massimo, Roberto, Ira and Jesse. My youngest friends in Edinburgh, Thomas and Catrin, have occupied a special place in my heart. Richard, for reading the damn thing twice, for staying up (remotely, in Berlin!) during the frantic night of last corrections and printing, for sorting out my misbehavin' linux—but even more so for dragging me away from my desk from time to time to lovely places like Gullane beach et al., and for being there.

Last, and most importantly, I want to thank the poor people who are contractually bound not to give up on me. Liebe Eltern, ich danke euch für euren unerschütterlichen Glauben an mich, für eure Engelsgeduld, eure Liebe und Unterstützung, und eure Toleranz in allen Dingen. Ohne euch hätte ich das nie geschafft.

Abstract

We present a new type of analysis for scientific text which we call *Argumentative Zoning*.

We demonstrate that this type of text analysis can be used for generating user-tailored and task-tailored summaries and for performing more informative citation analyses.

We also demonstrate that our type of analysis can be applied to unrestricted text, both automatically and by humans. The corpus we use for the analysis (80 conference papers in computational linguistics) is a difficult test bed; it shows great variation with respect to subdomain, writing style, register and linguistic expression. We present reliability studies which we performed on this corpus and for which we use two unrelated trained annotators.

The definition of our seven categories (argumentative zones) is not specific to the domain, only to the text type; it is based on the typical argumentation to be found in scientific articles. It reflects the attribution of intellectual ownership in scientific articles, expressions of authors' stance towards other work, and typical statements about problem-solving processes.

On the basis of sentential features, we use two statistical models (a Naive Bayesian model and an ngram model operating over sentences) to estimate a sentence's argumentative status, taking the hand-annotated corpus as training material. An alternative, symbolic system uses the features in a rule-based way.

The general working hypothesis of this thesis is that empirical discourse studies can contribute to practical document management problems: the analysis of a significant amount of naturally occurring text is essential for discourse linguistic theories, and the application of a robust discourse and argumentation analysis can make text understanding techniques for practical document management more robust.

Contents

Chapter 1

Introduction

The topic of this thesis is information management for researchers. Information management is a task that has attracted the attention of researchers in information retrieval and recently also researchers in artificial intelligence and natural language processing. The management of information contained in *scientific articles* poses specific problems. This introduction will set the scene by elaborating what is special about scientific articles. Before we describe the specific goal of this thesis, we will introduce the data we work with: a corpus of "real-life" computational linguistics conference articles. We will also discuss why we find this topic interesting, both from a research perspective as well as from a practical one.

This discussion will result in our general hypotheses for this work. We will argue for the application of empirical discourse studies when tackling document management problems. We believe that the argumentative analysis of naturally occurring text can provide subject-matter independent information which can fulfil many searchers' information needs, particularly the needs of less experienced searchers.

1.1. Information Foraging in Science

In today's fast moving academic world, new conferences, journals and other publications are springing into existence and are expanding the already huge repository of scientific knowledge at an alarming rate. Cleverdon (1984) estimates an annual output of 400,000 papers from the most important journals covering the natural sciences and technology. Kircz (1998) states that Physics Abstracts, the major bibliographic abstracting service in physics and the manufacturer of the INSPEC database, indexed 174,000 items in one year alone (1996), of which about 146,500 are journal articles. However, these already impressive numbers exclude less important journals, workshop proceedings, conference papers and non-English material. Indeed, the growth rate is probably exponential—Maron and Kuhns (1960) estimated that the indexed scientific material doubles in volume every 12 years.

The masses of information the researcher is exposed to make it hard for her to find the needle in the haystack as it is impossible to skim-read even a portion of the potentially relevant material. The information access and search problem is particularly acute for researchers in interdisciplinary subject areas like computational linguistics or cognitive science, as they must in principle be aware of articles in a whole range of neighbouring fields, such as computer science, theoretical linguistics, psychology, philosophy and formal logic.

Apart from keeping abreast of developments in scientific fields in general, more practical requirements emerge when researchers who are experienced in one scientific field start getting interested in a *new* scientific field, in which they have no prior knowledge. Their information needs have suddenly changed: Kircz (1991) states that such readers seek understanding instead of a firm, formal answer. The exact information need is not known beforehand; the questions they pose are not precise (Kircz' example is the question *"what are they doing in high-temperature super-conductivity?"* (p. 357)). Belkin (1980) refers to their situation as an "anomalous knowledge state". We think that researchers in a new field initially need answers to the following questions:

What are the main problems and main approaches? Knowledge of a number of important concepts in the field needs to be acquired: the current problems and the standard methodologies in the field. For the main approaches, the researcher needs to know their strengths and weaknesses. The searcher also needs to gain an overview of the evaluation methodology and typical numerical results in the field.

Which researchers and groups are connected with which concepts? Researchers' names—and the institutions where they work—must be associated with seminal approaches and seminal papers. The searcher must determine *schools of thought*: clusters of people working together, sharing premises and building on each others work.

If researchers read a paper in a new field, they are particularly interested in the general approaches described, the relation to other work, and its conclusions, instead of specialist details (Kircz, 1991). Oddy et al. (1992) and Shum (1998) argue that what such readers particularly need is an embedding of the particular piece of work within a broader context and in relation to other works.

The preferred information source at that stage of knowledge is an experienced colleague. Another standard technique for gaining a deeper overview of a field is to find a recent review article, to follow up the bibliographic links and to read however many of those papers one's time permits.

But sometimes neither of these useful aids is available, and a full-blown bibliographic search using an electronic document retrieval system is necessary, e.g. BIDS, FirstSearch or MEDLINE. This is typically done by a keyword search, where the keywords can be combined with Boolean operators.

In most commercial bibliographic data bases, keyword search is still performed on *document surrogates*, rather than on the full text of the document, as the full text is not always available in electronic form. Typical document surrogates used in document retrieval environments are bibliographic information (i.e. title, authors, date of publication, journal name), a list of index terms, or a human-written summary. The assumption is that these document surrogates capture an important aspect of the meaning of the document, i.e. that they are able to give the searcher a characterization of the contents of the paper, and that they can thus be used as a search ground. Mathematically sophisticated matching procedures between the document surrogates and the user's query measure how appropriate the document is for a certain query (*query-document similarity*). Document surrogates are also used to present the search result to the searcher, typically as an unordered list. The user can then perform *relevance assessment* on the basis of the document surrogates, i.e., she can filter out the obviously irrelevant documents from the search results.

There is a wide range of empirical studies about users of online data bases (Bates, 1998; Borgman, 1996; Fidel, 1985, 1991; Saracevic et al., 1988; Ellis, 1992; Ingwersen, 1996). These studies look at many different factors like searching experience, task training, educational level, type of search questions and user goals. The few of these studies which include inexperienced users conclude that the state of the art in document retrieval systems puts less experienced users at a disadvantage: those who have less well-defined queries and information needs (Clove and Walsh, 1988).

As they know neither the basic concepts nor the terminology of the new field, such searchers cannot possibly do well on keyword searches. The search terms they choose are often too unspecific and produce too many hits (Ellis, 1989a,b), hits where the term has another meaning, or no hits at all. As most search engines for bibliographic search rely on Boolean search and return the search results as an unranked list, they are at risk of getting lost in the returned list of document surrogates. Kircz (1991) calls this phenomenon the "frustrating circularity of the Boolean search process": clean, relevant information can only be retrieved from a data base if the searcher already knows what she is looking for.

Inexperienced searchers also have problems with the relevance decision itself. They cannot be sure that the retrieved articles are relevant to them or if they contain so-called *false negatives*. On the other side, and even more frustratingly, they must suspect that a myriad of relevant articles are in the database which their search has *not* found (*false positives*). (False negatives and positives are a normal phenomenon in free-text search; they are caused by polysemy and synonymy and by more complex features of unrestricted language.) To have access to high-quality document surrogates would be very important to the searchers—good abstracts are essential, as these are often the first detailed indication of the document's contents that they see. Titles alone are typically not informative enough for them.

However, even with imperfect search there is typically a convergence towards a few seminal papers which are frequently cited—even if the searcher was unlucky enough to start the search with peripheral, controversial or weak papers (along with the outright irrelevant ones). However, this is a more or less random process which might require a long time.

There are many ways in which this situation could be ameliorated, e.g. by better search methods or by better presentation of the search results. Best match (i.e. ranking) search algorithms rely on the intuition that it is crucial to get the right papers to the user in the right order, e.g. Salton's (1971) SMART system, or Robertson et al.'s (1993) OKAPI system.

The retrieved items can also be displayed by *document–document* similarity rather than by *query–document* similarity, e.g. VIBE (Olsen et al., 1993), Scatter/Gather (Hearst and Pedersen, 1996), Vineta (Krohn, 1995), Bead (Chalmers and Chitson, 1992), TileBars (Hearst, 1995) and Envision (Nowell et al., 1996).

In this thesis we will choose a different route: in the line of automatic abstracting approaches, we aim to improve the document surrogates returned to the searcher. We believe that better document surrogates will not only support the searchers in their relevance decision but it should also improve search itself. We believe that it is particularly important to design document surrogates which represent information needs that are typical for new searchers. In order to generate such document surrogates, the *right* kind of information must be extracted from the articles. This thought is one of the starting points for the present thesis.

1.2. Scientific Articles

One of the reasons why we chose to work with scientific articles is the practical value of better document retrieval environments for scientists. Scientific research articles are the main source of current leading-edge information for researchers, rather than text books or other sources of scientific information. In a library setting, there is a realistic demand for better summaries, or better document surrogates in general, cf. the recent interest in digital libraries.

The other motivation is more theoretical. Scientific papers are different from other text types with respect to their overall structure, an aspect we are particularly interested in. For a start, they are not organized in a time-linear manner. Assumptions about time linearity might help with the processing and summarization of simple narrative and newspaper text. Even though scientific articles are reports of intellectual work which was conducted within a certain time frame, their presentation follows the chronological order only in exceptional cases. Instead, the article structure usually mirrors the internal problem space and the scientific argumentation. The clear communicative function of scientific articles and the text-type specific expectations based on this function can provide a possible handle for subject matter-inspecific information extraction from such articles.

The writing style in scientific articles shows a considerable level of variation. Some articles are overtly argumentative, arguing against another author's views; others present empirical work such as a linguistic survey or corpus study in a more objective manner; some describe practical work like an implementation for a given problem. In interdisciplinary fields, articles might combine research methodologies from more than one discipline, e.g. a computational simulation of human behaviour originally observed in a psychological experiment. The linguistic expressions occurring in the articles mirror this variety.

Scientific articles are also biased; they describe the author's work from her own viewpoint. This bias is an integral part of the communicative function of scientific articles: they were written to convince the reader of the validity of a given research. The texts thus typically contain explicit markup of this rhetorical information (*meta-discourse*). In contrast, news stories have a supposedly neutral news anchor, and narrations are often told by an omniscient, neutral narrator. We are interested in the author's bias and aim to exploit it for our task.

Scientific text is harder to analyze than the texts typically used in discourse linguistic approaches. The reason for this is that it is not trivial *which* kind of document structure underlies scientific articles. Grosz and Sidner (1986) analyze apprentice–experts dialogues with an obvious task-structure; Iwanska's (1985) procedural texts are similarly structured. Other texts used for discourse analysis are short and well-edited; cf. Marcu's (1997b) popular science texts. Our texts, in contrast, are more difficult.

We chose *computational linguistics* (CL) as a domain for a number of reasons. One reason is that it is a domain we are familiar with. This makes an intermediate evaluation of our work possible without requiring the judgement of external subject experts. The more theoretically interesting reason is that computational linguistics is a *heterogeneous* domain due to its multidisciplinarity: the papers in our collection cover a wide range of subject matters, such as logic programming, statistical language modelling, theoretical semantics and computational psycholinguistics. This results in large differences in document structure and forces us to choose a more domain independent approach to document structure. In sum, our collection is an exciting and challenging test bed for discourse analysis.

1.3. Empirical Natural Language Research

Corpus-based or empirical natural language research is the study of language based on examples of real life language use. It is a general methodology which has come back into fashion recently, and which is now applied in several tasks in theoretical linguistics and natural language processing, e.g. lexicography, syntax and lexical semantics (Manning and Schütze, 1999). The general idea is that a linguist's or system developer's introspection alone cannot predict the unexpected turns of real language use. Rather than dealing with invented or artificially simplified examples, a large sample of naturally occurring language should be used instead. Empirical linguists aim to describe as much of the data as possible, but accept the fact that it is not normally the case that 100% of the data can be accounted for.

It is generally accepted that large corpora are a reliable source of frequency-based data. Additionally, a corpus is a more powerful scientific methodology than introspection as it is open to verification of results (Leech, 1992).

We subscribe to this general methodology: if one is planning to develop a practical system for unrestricted and thus unpredictable text, it is indispensable to base the design of this system on some kind of corpus analysis.

Whereas the Message Understanding Conferences (e.g. MUC-7 1998) have provided several corpora of newspaper articles with answer keys which are readily used in the field (cf. section 2.2.2), researchers wanting to work on scientific articles are at a disadvantage. At the time when research on this thesis started, there was no corpus of scientific articles available, so we collected our own corpus. It was also generally agreed at the AAAI Spring Symposium 1998 for Intelligent Text Summarization (Radev and Hovy, 1998) that there is a real lack of corpora of scientific articles. A version of our corpus is now distributed by TIPSTER as part of the SUMMAC program (Tipster SUMMAC, 1999).

We are interested in naturally occurring, unrestricted text, and we wanted to choose data which is as representative of the field as possible. We chose the Computation and Language Archive (CMP_LG, 1994) as our source, which is part of the CoRR (Computing Research Repository), a large preprint archive.

The idea of a preprint archive is the rapid dissemination of work: researchers can make their results available to the community early, e.g. before the conference where the paper is presented. The preprint version can later be replaced with the published version. Preprint archives, if widely used within a community, are perhaps the best way to track new work, although there is not necessarily a guarantee that the work is peer reviewed.

Between its beginnings in April 1994 and the submission date of this thesis, 968 articles have been put into the CMP_LG archive. The archive seems to be commonly used in the field: for example, researchers in computational linguistics use CMP_LG numbers as a standard way of identifying their papers.

We collected *all* documents from CMP_LG deposited between 04/94 and 05/96 which fulfilled our selection criteria, e.g. they had to have an abstract and be available in LaTeX. All these criteria are formal and not content-based; they are described in full in sections 5.3.1 and 5.3.2, where details about the corpus collection work are given.

One of our selection criteria concerns where the papers were published. We chose what we perceived to be the most influential conferences in CL, namely the *Annual Meeting of the Association for Computational Linguistics* (ACL), the *Meeting of the European Chapter of the Association for Computational Linguistics* (EACL), the *Conference on Applied Natural Language Processing* (ANLP) and the *International Conference on Computational Linguistics* (COLING). As a result, we know that all our papers had been peer reviewed. Restriction to these conferences does not introduce a bias, as CL is a field with few journals, where conferences are very important, and as the chosen conferences are the most influential ones. We also included papers presented in the student sessions, and those published in the proceedings of ACL-sponsored or EACL-sponsored workshops.

The deposition of articles on a preprint archive is voluntary and not systematic; some researchers might choose not to contribute their articles at all, whereas others might deposit an unrepresentatively high number of their articles. It is therefore difficult to claim that our corpus is representative of the *field* of CL as such. However, due to the unbiased sampling procedure, our collection should be reasonably representative of computational linguistics conference articles published in the given time frame and deposited on the CMP_LG archive: there is no reason to believe that new articles which would fulfill our selection criteria should be systematically different from the articles in our collection.

80 papers passed our selection criteria. They constitute the final, closely inspected corpus used in this thesis; details of the corpus are listed in appendix A.2. Roughly, the largest part of articles (about 45%) describe implementational work, 25% describe theoretical-linguistic work, 20% experimental work (corpus studies or psycholinguistic experiments) and 10% report evaluation (i.e., no completely new method is introduced in these articles; instead, already known systems or theories are compared and evaluatively measured).

Following from the fact that we are using unrestricted, naturally occurring text coming from a prepring archive, our texts display large variability in writing style. Some articles in our collection which do not use fully grammatical English; typing errors abound, and the register varies between formal and extremely informal, as the following two sentences illustrate:

> Formal:
> *While these techniques can yield significant improvements in performance, the generality of unification-based grammar formalisms means that there are still cases where expensive processing is unavoidable.* (S-7, 9502021)

> Informal:
> *This paper represents a step toward getting as much leverage as possible out of work within that paradigm, and then using it to help determine relationships among word senses, which is really where the action is.* (S-158, 9511006)

The corpus contains 333,634 word tokens. Even though this is much smaller than the large scale corpora typically used in corpus-based NLP (natural language processing), it still provides an unbiased resource describing a substantial amount of scientific text in computational linguistics.

For comparative purposes, we also had access to two other corpora: a corpus of agriculture, from Chris Paice's group at the Computer Science department of the University of Lancaster, and a corpus of papers in cardiology, from Prof. Kathleen McKeown's group at the Computer Science Department of Columbia University, NYC. In some cases, we will compare properties of our texts to texts from these corpora.

1.4. Goal and Outline of this Thesis

This thesis aims to contribute towards the automatic generation of document surrogates in the framework of a document retrieval environment for scientific articles. The practical topic of this thesis is how document surrogates can help researchers in their scientific information foraging activities, particularly those researchers who are new in a given field.

The thesis is structured as follows: The next chapter will define the goal in more detail, after a look at summaries in today's document retrieval environments. It will show that traditional human-written summaries are not flexible toward user expertise and task requirements, which is particularly a problem for novice researchers in a field. We argue that document surrogates should capture similarities and differences between related articles, which summaries typically do not. Current methods for automatic abstracting, on the other hand, create summaries which are either too generic, containing too little information to adequately characterize the document, or too inflexible towards unexpected material in the text. To ameliorate these problems, a new document surrogate is introduced: the Rhetorical Document Profile (RDP). It encodes typical information needs of new readers, e.g. global level information like which SOLUTION was introduced in the article, or what the GOAL of the article was. We will argue that RDPs are useful for practical document retrieval applications: flexible summaries can be generated from them, and types of connections between articles can be expressed in a construct called a *citation map*. The rest of this thesis will explore the possibility of creating RDPs automatically by a process of robust text analysis and extraction.

Chapter 3 introduces a new document analysis called *Argumentative Zoning*. Argumentative Zoning concentrates on global discourse information: the rhetorical status of a sentence in relation to the discourse act of the overall paper. It turns out that some of these rhetorical states coincide with the information needs introduced in chapter 2; thus, this chapter also gives a justification for RDPs. Argumentative Zoning is independent of writing style, subject matter, and, to a certain degree, subdomain, but relies on text type specific expectations (communicative acts). Section 3.2 introduces our model of prototypical scientific argumentation. This model is operationalized in section 3.3 by introducing seven different information categories or argumentative zones.

Chapter 4 discusses our evaluation strategy for the new task of Argumentative Zoning, in view of similar tasks (fact extraction, text extraction and dialogue coding tasks). The annotation scheme developed in chapter 3 will be empirically validated with respect to human performance, i.e. we will measure to which degree human judgements of argumentative zones agree. This annotation experiment provides us with quantitative data about the reliability of the scheme, and it also gives us training material for our prototype implementation of Argumentative Zoning.

Chapter 5 documents an experiment in automatic Argumentative Zoning. First, we will describe a pool of sentential features which correlate with the sentence's rhetorical status. Then, we will describe the implementation of a prototype system for automatic annotation: the automatic determination of these features, the statistical classifiers used, and a rule-based alternative implementation. We will then present the results of an intrinsic evaluation of our system.

The conclusions will bring us back to the main working hypothesis of the thesis: that empirical discourse studies can contribute to practical document management problems. In this thesis, we use practical discourse studies (in our case, centered around argumentative zones) to help identify the kind of information in scientific texts which are crucial for searchers' information needs. We experimentally show that humans can be trained to perform Argumentative Zoning consistently, and that this behaviour can be simulated by an algorithm; we consider this as a proof of concept for RDPs and for Argumentative Zones.

In the course of the thesis, the following research questions will be addressed:

- *Discourse linguistics:* Is it possible to analyze the document structure of scientific articles in a subject matter-independent way? At which abstraction level should such an analysis define its units and relations? What are the linguistic signals of this structure?

- *Experimental psychology:* To which extent do humans share intuitions about information and document structure in scientific papers? Can people be trained to apply a fixed annotation scheme for the analysis? In which aspects do the humans' annotation differ and agree most?

- *Computational linguistics and artificial intelligence:* Can we identify algorithmically determinable signals of argumentation and document style in unrestricted text? Which of those can be used for system building and evaluation? How much "understanding" would such a system need to produce acceptable document characterizations?

Chapter 2

Motivation

In this chapter, we will define the goal of this thesis in more detail. We will start with a discussion of the most prominent document surrogates—summaries—and the state of the art in producing them, both manually and automatically.

In section 2.1 we focus on manual summarization. We argue that the current practice of abstracting is undergoing a big change because more and more scientific research text is available in electronic form. The high-quality human-written summaries, deeply rooted in the paper-based publishing world, cannot offer the flexibility towards task and user expertise that becomes more and more of a necessity. We will argue that one of the problems of current summaries is that they do not take connections between articles into account.

Section 2.2 will start with an overview of two current automatic summarization methods: text extraction and fact extraction methods. Both have advantages and drawbacks: inflexibility in the case of fact extraction method, the lack of context-sensitivity in the case of text extraction.

In section 2.3 we suggest an approach which synthesizes text and fact extraction methods by attaching global-level rhetorical information to extracted sentences. This results in *Rhetorical Document Profiles* (RDPs). We argue that RDPs combine the best of both worlds from fact extraction and text extraction methods, and that they have definite advantages in a document retrieval environment. We then show how the information contained in them could be used to generate tailored summaries and annotated *citation maps*.

2.1. Manual Abstracting

Humans are well-known to be good summarizers (Kintsch and van Dijk, 1978; Sherrard, 1985; Brown and Day, 1983), and summaries written by well-trained information specialists are of particularly high quality (Lancaster, 1998; Cremmins, 1996). However, as we will see, this is not enough to immediately solve all of the researchers' search problems introduced in the previous chapter.

2.1.1. Summary Tailoring

Information services (secondary publishers) like the *Institute for Science Information, Inc.* or *Chemical Abstracts Service* specialize in information management for scientists. In order to keep researchers informed of publications in their area of interest, these companies publish, amongst other things, journals with summaries of research material.

Such information services have made a huge investment in the production and dissemination of summaries. They employ information specialists (professional abstractors/indexers), highly qualified professionals who have been trained in the art of summarizing and indexing articles and books.

Professional summaries are written according to agreed guidelines and recommendations (McGirr, 1973; Borko and Chatman, 1963; ANSI, 1979; ISO, 1976). The guidelines are concerned with the informativeness and readability of the human-written summaries; they try to make sure that they are general, long-lived and high-quality accounts of the information contained in a scientific article. For example, the guidelines give a certain maximum and minimum number of words to be used in a summary. They recommend that summaries should be aimed at a particular kind of reader, a semi-expert: somebody who knows enough about the field to understand basic methodology and general goals but who would not understand all specialized detail. Also, the summaries are supposed to be self-contained (Lancaster, 1998, p. 108): the reader should be able to grasp the main goals and achievements of the full article without needing the source text for clarification.

In the literature on human summarization we find very little about the tasks that users are assumed to perform with the summaries. The only mention of summary use we find is at an abstract level (e.g. in Lancaster 1998):

1. Summaries can be used as substitutes for the whole document. If researchers

want to be kept *aware* of new publications in a field, it is often enough for them to read summaries in abstract journal (alerting function), instead of reading the full article.

2. Another example of substitutive use of summaries is when they are used to *refresh* a reader's memory of a previously read article.

3. Another situation is the use of summaries in parallel with the full text, e.g. when *previewing* of the structure of the source document. Here, the summary serves as orientation about the structure of a document that has already been chosen, similar to a table of contents.

4. Rarely, summaries are used for reasons having nothing to do with the original text. For example, when users need to decide if they have chosen the *right data base* for a search, they can looked at a random summary of that data base for mere seconds.

5. The most typical use of summaries in a document retrieval environment is for relevance decision, i.e., to judge whether or not the corresponding, as yet unknown, full article is relevant to searchers' current information need (Cremmins, 1996; Rowley, 1982). During this step, the reader might also recognize papers she has read before. The relevance decision process will determine a set of probably relevant papers, which can then be looked up in the library, requested in full from the author or ordered as paper copies. A similar use is the decision of whether or not the searcher has read an article already.

Typically, there is only one version of the summary. The only generally accepted dimensions of summary variance in the literature are compression (i.e. length of summary in comparison to the full text) and the distinction between *indicative* and *informative* summaries. Indicative summaries contain an indication about the topic of the text (i.e., they contain purpose, scope or methodology), whereas informative summaries also name the main findings and conclusions of the text (Rowley, 1982; Cremmins, 1996; Lancaster, 1998; Michaelson, 1980; Maizell et al., 1971). Indicative summaries are of use for relevance decision and all functions which assume that the full text is either available, or that an indication of the general contents is enough for the researcher. Informative summaries, on the other hand, are autonomous texts which can be used as full text substitutes.

Consider the following examples from Lancaster (1998, p. 95):

Indicative Summary:

> *Telephone interviews were conducted in 1985 with 655 Americans sampled probabilistically. Opinions are expressed on whether: (1) the establishment of a Palestinian state is essential for peace in the region; (2) U.S. aid to Israel and to Egypt should be reduced; (3) the U.S. should (a) participate in a peace conference that includes the PLO, (b) favor neither Israel nor the Arab nations, (c) maintain friendly relations with both. Respondents indicated whether or not they had sufficient information concerning various national groups in the region.*

Informative Summary:

> *Telephone interviews conducted in 1985 with 655 Americans, sampled probabilistically, brought these results: most (54–56%) think U.S. aid to Israel and Egypt should be reduced; most (65%) favor U.S. participation in a peace conference that includes the PLO; more than 80% consider it important that the U.S. should maintain friendly relations with both Israel and the Arab Countries; 70% believe that the U.S. should favor neither side; most (55%) think that the establishment of a Palestinian state is essential to peace in the region. The Israelis are the best known of the national groups and the Syrians the least known. The Arab-Israeli situation is second only to the conflict in Central America among the most serious international problems faced by the U.S.*

There is disagreement which type of abstract is easier to write. Rowley (1982) argues that indicative abstracts are more difficult to write, and (Manning, 1990) claims the opposite. Most authors distinguish the so-called *informative-indicative* summary, where some results are given (as would be in an informative summary), whereas other parts of the paper are treated only indicatively. Rowley (1982) states that this kind of summary is most commonly used nowadays; Lancaster (1998) (who does not recognize informative-indicative summaries) states that informative summaries are less common than indicative ones.

Informative summaries are further divided into purpose-oriented and findings-oriented summaries, which differ in the order of the information presented (Cremmins, 1996; ANSI, 1979). Findings-oriented summaries present findings (results and conclusions) first. The following examples from Cremmins (1996, p. 109) illustrate that the difference between them is not great.

Purpose-oriented indicative-informative summary:

> *Suggestibility was measured under indirect, auto-, hetero-, and conflicting forms of suggestion by using the Body Sway Test. Healthy and ill students and patients, with and without autogenic training, were tested. Equally strong effects occurred under all four forms of suggestion. Autogenic training affected positive behavior on the test in both healthy and ill students. Negative behavior in this test occurred when autogenic training was lacking. The behavior of female patients was more positive than that of males under conflicting suggestions.*

Findings-oriented indicative-informative summary:

> *Equally strong effects of suggestion occurred under indirect, auto-, hetero-, and conflicting forms when the Body Sway Test was given to healthy and ill students and patients, with and without autogenic training. The training affected positive behavior on the test in both healthy and ill students. Negative behavior in this test occurred when autogenic training was lacking. The behavior of female patients was more positive than that of males under conflicting suggestions.*

Even though Cremmins does not say so explicitly, it seems likely that the two types of summaries support (slightly) different kinds of tasks. For example, the findings-oriented summary might be more useful to a medical researcher trying to spot the kinds of experimental *results* she would need in support of an argument of her own. The difference in order seems to imply a model of summary use in which users sequentially read the summary from the start and stop reading when they have found what they need for their relevance decision (Borko and Bernier, 1975, p. 69). However, we found no empirical studies in the literature which focus on summary reading strategies or which measure the appropriateness of different kinds of summaries for a certain task. In sum, the assumptions in the literature about user tasks are minimal and do little more than support two uses of summaries: a) as texts that give an indication of the contents and b) as autonomous texts.

Another point is the question how to determine what is relevant for a given user at a given time. There are a myriad of reasons why a user would classify a given document as relevant at a given point in time during relevance decision (Rees, 1966). A vast experimental and theoretical literature in information science has been concerned with the slippery concept of relevance (Saracevic, 1975; Schamber et al., 1990). In principle, it is undisputed that the large-scale context influencing the interpretation of a text and the relative importance of a part of the text depends on and comprises the writer and reader of the text and their background, goals, and viewpoints. Even to the same reader, at different points in time, different aspects of the same text might be relevant. Spärck Jones (1990) describes the general problem by saying that pertinence is situational to a unique occasion.

It is hard to argue with Lancaster (1998) when he states that "the abstractor should [...] omit other information that readers would be likely to know or that may not be of direct interest to them." (p. 107)—the difficult part is to guess *which type of information* different groups of readers are likely to know. The *informedness of the intended audience* is one of the central points in user tailoring known from text generation (Spärck Jones, 1988; Paris, 1988, 1994). The summarizing industry, however, does not envisage summaries which are responsive to level of expertise of the reader. Though the concept of *subject slanting* (i.e., tailoring the summary to the anticipated interest of its users) is quite common when summaries are produced for the internal use of one organization, rather little slanting takes place in general information services (Herner, 1959).

Kircz (1991) distinguishes between uninformed, partially informed and informed readers. He argues that the level of subject knowledge influences which information readers draw from scientific articles. Uninformed readers read introductions and conclusions, and also overview figures/graphs if present, and the list of references. Partially informed readers read papers particularly for the general approaches described, the relation to other work, and the conclusions. Informed readers, in contrast, can use their scientific background knowledge in a field to find their way in the literature quickly. They typically scan articles fast; only the core of information is read, e.g. the numerical results. As traditional summaries are geared towards partially informed readers, they are therefore often too terse for uninformed readers, and too verbose for informed readers. This poses more of a problem for the uninformed than for the informed reader.

It is important to see that the inflexibility of traditional summaries is rooted in the function of summaries in the paper-based world of publications which we just described. Recently, due to the omni-presence of the world wide web and electronic journals, more and more papers are available in electronic form—it can be expected for the near future that most bibliographic document retrieval environments will provide researchers with electronic versions of the paper during search time. This development has strong influence on what the most appropriate document surrogate for the search task should look like.

Firstly, and rather obviously, the fact that the full paper is available in electronic form is a necessary precondition for realistic automatic summarization. In the early era of summarization, research was restricted by data problems, and articles had to be manually encoded and typed. Now, the manipulability of electronic text makes it possible to summarize millions of papers—different summaries of one paper can be created on the fly, and it is theoretically possible to be flexible towards length, end task and user expertise.

But electronic texts also pose new challenges, as studies of readers in electronic environments show (Dillon, 1992; Levy, 1997; Adler et al., 1998; O'Hara et al., 1998). Kircz (1998) criticizes the fact that new electronic publishing technology has mostly been used to echo the old style of paper-articles in the new medium, rather than employing new functionality. Other work concentrates on reading strategies. For example, on-line browsers like Netscape or Internet Explorer, and previewers like Ghostview or Adobe Acrobat can display the articles directly on-screen, but they cannot yet simulate the physical properties of paper. O'Hara and Sellen (1997) found that this disrupts typical reading strategies of scientists, e.g. the so-called *non-linear* reading (Samuels et al., 1987; Dillon et al., 1989). The non-linear reader jumps in a seemingly arbitrary fashion from the conclusion to the table of contents and scans the section headers and captions, in order to get an ad-hoc idea of the structure of the text. This strategy serves to efficiently build a model of the text's structure as well as to extract the main concepts of the paper, and is a typical reading behaviour for scientists (Pinelli et al., 1984; Bazerman, 1988).

But even though today's browsers might give a suboptimal representation of the article, new, intelligent display mechanisms could exploit and thus compensate for some of the functions of the material paper (O'Hara and Sellen, 1997). One way in which new functionality can help readers in an electronic environment is the support of citation indexes as an additional search strategy, which will be treated in the next section.

2.1.2. Citation Information

There are information search tasks which are specific to research, tasks concerned with *connections* between research outputs (Oddy et al., 1992). Shum (1998) stresses that researchers, a community which is constantly contesting claims, need information about scientific relationships:

> [...] *relationships* are critical for researchers, who invest a lot of energy in articulating and debating different claims about the significance of conceptual structures. (Shum 1998, p. 19, his emphasis)

Such information results in the knowledge-rich cognitive net of information which Bazerman (1985) describes for physicists and Charney (1993) for evolutionists. Experienced researchers know the important names in the field; they know institutions and their specialities and preferred methodologies; they know schools of thought and how they interrelate. These information nets are acquired over time, by reading, through research, at conferences, and by discussions with colleagues.

However, in the course of conducting a *new* piece of research a researcher is likely to come up with immediate questions for which the background knowledge provides no answers. These pressing questions often result in a document retrieval search:

Supportive Data: During the writing of a paper, the researcher might look for support in the literature for a certain claim she needs as a step in the argumentation. She might first want to check if the claim has been previously stated in print; if this is the case, it is necessary to respect that paper's prior claim of intellectual ownership by citing the given paper. Another task is to find out if the given paper is the original citation for the idea, or if that work continues somebody else's work. In interdisciplinary fields, one might need to include specific evidence coming from a particular neighbouring field, e.g. validation of the claim in the form of experimental psychological results.

Differences and contrasts: The researcher might want to check if there are published results that are contradictory to her own. She might also want to find out if there are competitors to her claim, i.e. rival approaches (approaches with the same goals, but a different methodology). Another question might emerge if she has identified a weakness of some other work—she might want to find out if that work has been criticized by somebody else before, and if so, what exactly constituted the prior criticism.

Updates of old research articles: It sometimes happens that a researcher finds an article which contains the right information (e.g. a particular scientific fact or claim needed for her current work), but which happens to have been published a long time ago. It is considered bad practice to cite the old paper without stating what happened in the meantime with respect to the scientific claim. Shum (1998) mentions the following question as pressing for scientists: *"What impact did certain evidence have?"* More recent articles need to be located which either still maintain the same claims (maybe with additional evidence), or contribute counter-evidence. If the original article is a dated *review* article, a special case of this information need applies: *each* cited article needs to be traced forward in time to some more recent research.

Information about the relatedness of scientific articles is available from citation indexes, e.g. the Institute for Scientific Information (ISI)'s multidisciplinary citation indexes (ISI, 1999). Such indexes cover only a small range of journals, which is justified by the fact that a relative small number of journals account for the bulk of significant scientific results (Garfield, 1996). Traditionally, citation indexes are used for bibliometric studies, i.e., to measure the quality and academic impact that a piece of academic work or a journal has (Garfield, 1979)—an approach which has disadvantages as well as advantages (cf. section 3.2.2). In the context of our task, and apart from impact assessment, citation links can be used in two ways:

- Citation links can provide an alternative way of accessing information in the data base.
- Similarities between articles can be determined by their citation behaviour.

Work on article clustering by citations includes bibliographic coupling (Kessler, 1963) (if two articles have similar bibliographies then they must share a topic) and co-citations Small (1973) (if two papers often occur together in other article's bibliographies then they must share a topic). There is an analogy with research on the topology of the world wide web (Kleinberg, 1998), where *authorities* (often-referred-to, seminal pages) and *hubs* (clusters of pages which list many authorities) are identified.

Citation links can also be used for information access. ISI BIDS, for example, allows users to list document surrogates of all articles citing a given one, and many on-line proceedings are internally citation-indexed (SIGMOD, 1999)—articles cited in the paper can be reached directly, but there is also a listing of all articles citing the given article *later*. Recently, tools for citation manipulation with even higher functionality have emerged. The new citation visualization tool CiteSeer, which is part of NEC's digital library ResearchIndex initiative (Giles et al., 1998) performs *Autonomous Citation Indexing*: a citation index is automatically built from all papers available to CiteSeer. References in running text are automatically determined, and the reference list is parsed. Citation forms appearing in slightly different shape in other sources are mapped onto each other. CiteSeer displays the context in which a given citation occurs in running text by showing the sentence containing the physical reference along with snippets of keywords, headlines and adjacent sentences in an extract-style. The following example citation is taken from (Giles et al., 1998, p. 94); it shows a reference to the paper "*Maximum likelihood from incomplete data via the EM algorithm*", published by Dempster et al. in 1977. The following segment has to be read in order to determine how the two papers relate to each other:

> *... other variant algorithms are also seen to be possible. Some key words: EM algorithm, incremental algorithm, free energy, mixtures Submitted to Biometrika 1 Introduction The Expectation-Maximization (EM) algorithm finds maximum likelihood parameter estimates in problems where some variables were unobserved.* **Its widespread applicability was first discussed by Dempster, Laird and Rubin (1977).** *The EM algorithm estimates the parameters iteratively, starting from some initial guesses. Each iteration consists of an Expectation (E) step, which finds the distribution for the unobserved variables, given the known values for the observed variables and the current estimate of the parameters, and a Maximization...*

Even though CiteSeer enables the visualization of the connection between related articles, it does not provide the user with automatic classification of the *type* of this connection. CiteSeer opted to be non-interpretative, objective, but unhelpful to the user; the user always has to read the citation context in order to work out the relationships.

Nanba and Okumura (1999) introduce a support tool for writing surveys which categorizes citations in text (on the basis of cue words) into "Type C" citations (contrasts), "Type B" citations (based-on relationship) and "Type O" citations (others); Type "C" links are used to display differences and similarities between documents in a *reference graph*. This is a potentially useful way to structure search results, but clusters of papers are often uninformative to users if there is no indication what is similar between papers in this cluster. Users also need to know what single papers are about in "absolute" terms, and not just in relation to other papers—which is typical summary information.

Human-written summaries, on the other hand, do not typically include information about connectedness of research—guidelines actively discourage abstractors from including information about related work. Cremmins (1996) states that it should not be included in an abstract unless the studies are replications or evaluations of earlier work (p. 15). Weil et al. (1963) tell us explicitly never to mention earlier work.

It is our idea that information about connections between papers and local information about one paper should be connected. This could result in a new type of document surrogate which would support the explorative navigation of articles. The processes of search, text skimming and relevance decision could thus be interleaved: during search, parts of a retrieved paper are highlighted; while the reader is navigating the set of returned papers, she might skim-read some of these paragraphs. These text pieces can either directly satisfy the searchers' needs, spark off a new search in a new direction, or convince her that the paper is not relevant after all.

Note how different this relevance decision in such an interactive search-and-display environment is from relevance decision in the paper-based world. There the outcome of the relevance decision was not to be seen for a long time: by the time the paper copy of a certain paper finally arrived, researchers might have half forgotten what their specific reasons for ordering it actually were. Due to this long-term character of relevance decisions, errors were difficult to amend retrospectively, and the risk of ordering the wrong paper was much higher.

Manual summaries are a construct of the paper-based world: texts were of high textual quality, but they were also long-lived and thus fixed. The type of document surrogate we propose will be more dynamic and flexible to the user and her search situation; it should allow for different abstracts to be generated dynamically when needed. Such document surrogates will have a much shorter life span than a valuable human-crafted summary. Even though they will be of lower *textual* quality when compared to such summaries, we predict many situations in which they will have an edge over traditional summaries.

The document surrogate should also include information about similarities and differences between papers; this information could be used either to provide typed links in a citation analysis tool or to enrich the generated summaries.

2.2. Automatic Abstracting

The current state of the art in automatic abstracting is characterized by a deep tension between robustness and depth of understanding. Like machine translation, summarization has been an early target for automation (Luhn, 1958), but the expectation that this is a "easily manageable task" was not fulfilled.

Since the early 90s, with computing power and storage orders of magnitude more plentiful, knowledge-poor, statistical techniques have become fashionable again. However, the view of the complexity of the task has changed within the community. Researchers today see automatic summarization as "one of the most complex tasks of all natural language processing." (Hovy and Lin, 1999, p.92).

Comprehension-based summarization, the traditional symbolic approach, is the most ambitious model for creating automatic summaries. One view is that there cannot be any summarization without a complete comprehension of the text at hand. The argumentation is simple: How should we be able to decide what is important in a text unless we have understood the text?

Figure 2.1 exemplifies the standard model for summarization by comprehension (Spärck Jones, 1994)). It comprises three steps: a) linguistic analysis of the text (syntactic, semantic, pragmatic), which results in the reconstruction of the document semantics in a representation language, b) compression of the contents, by some kind of manipulation of the representation language and finally c) generation of the summary text from the reduced representation.

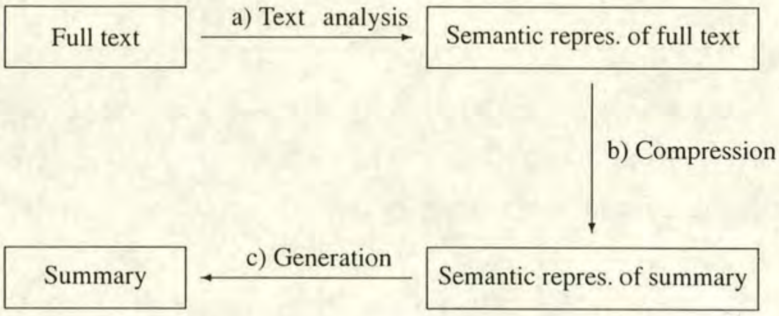


Figure 2.1: Summarization by Text Comprehension

The main problem with this approach is step a): it is not possible yet to map unrestricted text reliably and robustly into a semantic representation. Only then could one apply inference and the other operations that would take place in step b), e.g. following suggestions by Kintsch and van Dijk (1978); Alterman (1985); Brown and Day (1983) and Sherrard (1985). However, severe problems in linguistic analysis and knowledge representation (also referred to as the natural language bottleneck and the artificial intelligence bottleneck) make this model unrealistic for unrestricted text. As a result, people have been looking at alternatives for step a).

Text extraction is one of these alternatives. In this paradigm, step a) is performed in a radical way—each textual segment is condensed to a minimal representation, namely a number of features associated with the textual segment, e.g. whether or not the sentence contains the cue phrase *"to summarize"*. The determination of the features is typically performed in a shallow way, e.g. by calculating the lexical frequency of words in the textual segment, without the use of any linguistic knowledge. Step b), content selection, is performed by selecting a set of these scores, typically the *n* highest-ranking ones. Step c) is circumvented completely: the outcome of text extraction is the unchanged textual segments whose scores were chosen in step b).

The other solution is based on *fact extraction*. The representation "language" used is a set of frame-like templates (DeJong, 1982; Schank and Abelson, 1977). Step a) is performed by choosing the right template which describes the text, and by filling the slots in the template, e.g. by pattern matching operations. Step b) can be left out completely if the information contained in the templates is already little enough to make up the summary. Otherwise, condensation heuristics decide which ones of several template slots or whole templates are most relevant. Step c), the transformation of the reduced templates into natural language, can be performed either by using fixed templates or by deep generation.

We will in the following look at these two approaches in turn.

2.2.1. Text Extraction

Most of today's summarization systems use text extraction methods, including many commercially available ones, e.g. Microsoft's AutoSummarize (Microsoft, 1997), Oracle (Oracle, 1993), InXight (InXight, 1999) and ProSum (British Telecom, 1998).

The general idea of text extraction is the identification of a small number of "meaningful" sentences or larger text segments from the source text. The most common unit of text extraction is the sentence (Brandow et al., 1995; Kupiec et al., 1995), but some current systems extract paragraphs (Strzalkowski et al., 1999; Abracos and Lopes, 1997; Salton et al., 1994b).

Operational measurements of importance are based on algorithmically determinable properties of the text segment. Each text segment in the source text is scored according to this measure of importance, and subsequently the highest-rated segments are selected.

This produces *extracts* rather than abstracts: collections of the *N* most "meaningful" text units (sentences), taken verbatim from the text, and presented to the user in the order in which they appeared in the source text.

Extracts can be useful in a document retrieval environment instead of human-written indicative abstracts. A few well-chosen sentences can tell the reader about the terminology used, about the style and syntax, and about how loosely and coherently the text is written. If all the user needs is a tool for rapid relevance assessment, then such robust but uninformed methods can readily provide extracts which meet, to a reasonable degree, the information compression rates required (around 10% of the original text).

Over the years there have been many suggestions as to which low-level features can help determine the importance of a sentence in the context of a source text, such as stochastic measurements for the significance of key words in the sentence (Luhn, 1958; Baxendale, 1958), location of the sentence in the source text (Baxendale, 1958), connections with other sentences (Skorochod'ko, 1972; Salton et al., 1994a), cohesion (Morris and Hirst, 1991; Barzilay and Elhadad, 1999), co-reference information (Baldwin and Morton, 1998), sentence length (Kupiec et al., 1995), the presence of bonus/malus words (Luhn, 1958; Pollock and Zamora, 1975), title words (Edmundson, 1969), proper nouns (Kupiec et al., 1995) or indicator phrases (Paice, 1981; Johnson et al., 1993).

Single heuristics tend to work well on a certain type of document, but in that case success is concentrated on single documents that resemble each other in style and content. For the more robust creation of extracts, e.g., from texts with a high degree of variation in style, it is advantageous to combine these heuristics. The difficulty is to weigh the relative usefulness of single heuristics out of a given set. Edmundson assigns the weights manually. Kupiec et al. (1995) pioneered corpus-driven summarization research in which the combination of heuristics is learned from a training corpus and feature weights are automatically adjusted.

Kupiec et al.'s system uses supervised learning to determine the characteristic properties of those sentences which are known *a priori* to be extract-worthy (positive training examples). The features considered are: presence of particular cue phrases, location in the text, sentence length, occurrence of thematic words (document specific frequency of noun pairs) and occurrence of proper names. They redefine sentence extraction as a statistical classification task: the task is to estimate an unseen sentence's probability to occur in the summary, given its feature values and a statistical model of abstract worthiness acquired during training.

The big advantage of text extraction methods is that they are extremely robust. Due to the low level of analysis performed, it is possible to process texts of all kinds, independent of writing style, text type and subject matter. This means that unexpected turns in a news story, sudden changes in topic and other difficult phenomena can be treated in a shallow way—the extracts will, to a certain degree, reflect these particularities of the texts.

How does one measure the quality of extracts, and what lower bound (baseline) should they be compared to? Researchers have used either random choice of n sentences, or selected the n leading sentences. Which baseline makes more sense is text type dependent. Brandow et al. (1995) report that for newspaper text, a baseline defined by leading sentences can prove to be so hard to beat that more sophisticated sentence extractors perform *below* the baseline. The reason for this is that journalistic writing style already takes relevance into account by placing the most important information first. For scientific articles, a selection of leading sentences would not make an equally good baseline. Kupiec et al.'s baseline was constructed by leading sentences, and their best results achieved a 74% improvement over baseline. However, with baselines as weak as these, a look at the concrete output is needed to assess the quality of text extracts.

In order to have a concrete example of a sentence extract of a document for ongoing discussion, we used the commercial software AutoSummarize to create extracts of an example article taken from our corpus. This example article—cmp_lg:9408011—will be used throughout the thesis. It is the article most frequently cited by other articles in our collection. The full text of the article is reproduced in appendix B.2 (p. 285). We produced a 10-sentence AutoSummarize extract of the pdf version of the example article, which is given in figure 2.2.

Normally, AutoSummarize displays extracted sentences highlighted in the context where they were extracted from, but it is also possibly to list only the extracted sentences.

AutoSummarize, like many sentence extractors, extracts material other than full document sentences, e.g. titles and headlines (shown in bold face in figure 2.2).

It also selected a single line from the reference list at the end, namely item j), which is the title of a paper published by Rose et al. (1990). This paper is important for the article, but the titles of cited works are no standard summary items, especially

a) **Distributional Clustering of English Sentences**

b) **Distributional Similarity** To cluster nouns n according to their conditional verb distributions pn, we need a measure of similarity between distributions.

c) We will take (1) as our basic clustering model.

d) In particular, the model we use in our experiments has noun clusters with cluster memberships determined by p(njc) and centroid distributions determined by p(vjc).

e) Given any similarity measure d(n;c) between nouns and cluster centroids, the average cluster distortion is

f) If we maximize the cluster membership entropy

g) **Clustering Examples**

h) Figure 1 shows the five words most similar to the each [sic] cluster centroid for the four clusters resulting from the first two cluster splits.

i) **Model Evaluation**

j) 1990. Statistical mechanics and phrase transitions in clustering.

Figure 2.2: AutoSummarize Summary for Example Paper cmp_lg 9408011

if they are not signalled to the user as such. AutoSummarize did not extract sentences from the original abstract, even though the abstract was included in the full document.

In general, extracts are texts of low readability and text quality (Brandow et al., 1995). However this particular AutoSummarize extract reads surprisingly well: it contains no syntactic incoherences like dangling anaphora. None of the selected sentences is obviously displaced in the extract, and they give an idea of the general topic of the paper. We get the idea that it is about clustering, that it is a statistical, technical paper, and that it probably gives an algorithm of some kind. In a document retrieval scenario, this extract could be of use as a rough-and-ready relevance indicator.

Incorrect or confusing content characterization is a harder problem than superficial syntactic flaws, which is why Minel et al. (1997) propose independent evaluation of automatic abstracts by a) text quality and b) content characterization. Even if—like in our extract—each individual sentence is interpretable in isolation, that still does not mean that the extract as a whole will be easy to understand. Earl (1970) noted that extracts are often logically discontinuous. Problems with semantic coherence include unexpected topic shifts or repetitions, non-natural use of anaphora, and general logical incoherence.

With respect to the semantic connection between the sentences, apparent coherence of extracts can even be a *disadvantage*. Sentence d) in the extract appears 25 document sentences after sentence c)—it certainly does not elaborate on particulars related to sentence c). However, as readers are intuitively trying to coerce coherence for prose-like text, they will try to fill in the semantic gaps between potentially unconnected sentences by performing inference (Kintsch and van Dijk, 1978). Many of these inferences might introduce inappropriate semantics links and confuse the reader. In order to avoid this, many summarizers including AutoSummarize offer the possibility to show the extracted sentences highlighted in their original context; others present their extracts as a itemized list with bullet points (Kupiec et al., 1995) instead of continuous prose.

The other issue concerns the extent to which the extract characterizes the meaning of the document. The level of analysis performed seems too low to guarantee correct characterization, and Boguraev and Kennedy (1999) state:

> The cost of avoiding the requirement for a language-aware front end is the complete lack of intelligence—or even context-awareness—at the back end. The validity, and utility, of sentence-or paragraph-sized extracts as representations for the document content is still an open question [...]
>
> (Boguraev and Kennedy, 1999, p. 100)

Semantic incoherence and content selection problems become worse the longer the source document is. Typical sentence extractors compress a text down to about 15–25% of the original length—for example, they reduce a short newspaper article to a few sentences. In that case, the extract is still short enough to be read as an indicative "summary", even if the extracted sentences do not form a coherent text. However, things look different for scientific articles, which are much longer. With methods as untargetted as sentence extraction, one needs a 20% compression (or better still, 30%), in order to understand what a text is about: Morris et al.'s (1992) experiment showed that there is no difference in reading comprehension between subjects using the full text, subjects using indicative human-written summaries and subjects using extracts of 20% and 30% compression.

But this level of compression is very low. A 20-page article would have to be reduced to a 4 to 6-page collection of extracted sentences. Given that the statements in such a collection are semantically unconnected, it would be too much text to read and certainly not adequate for human consumption.

One might argue that sentence extracts are a good starting point for later automatic post-processing. However, text extraction is a completely context-insensitive method. Once the abstract-worthy sentences have been extracted, the logical and rhetorical organization of the text is lost. As a result, it becomes difficult to make sensible decisions on how to further reduce a long list of sentences without further information about the meaning of the sentences, the relationships between them or the contexts in which they occurred.

In sum, the low level of analysis performed and its context-insensitivity make text extraction a weak, albeit general and robust technique. Spärck Jones (1999) compares text extraction to looking at a text through tinted glass. All parts of the text can be "seen" by the text summarization technique, but the information we get is certainly blurred.

2.2.2. Fact Extraction

Summarization methods relying on fact extraction need a template to represent the information extracted. We will first discuss the style of these templates and then turn to the question of how to generate coherent summaries from them.

A large-scale competitive evaluation of systems for fact extraction from real-world news paper text was provided by the Message Understanding Conferences (MUC), sponsored by DARPA since the late 1980s (Grishman and Sundheim, 1995). Processing in MUC is restricted to text from a narrow domain, as figure 2.3 shows.

Competition	Domain
MUC 1 & 2	Naval sightings and engagements
MUC 3 & 4	Terrorist attacks in Central and South America
MUC 5	International joint ventures and electronic circuit fabrication
MUC 6	Changes in company management
MUC 7	Telecommunications satellite launches

Figure 2.3: Domains of Texts in Different MUC Competitions

MUC templates are shallow knowledge representation schemes without recursion, which encode information about entities and their relations. They are an instance of the frames well-known from symbolic text understanding and memory organization theories (Minsky, 1975; Schank and Abelson, 1977).

What can summarizers do with such templates? The SUMMONS system as described in Radev and McKeown (1998) and McKeown and Radev (1995) is based

Message: ID	**TST-REU-0001**
Secsource: Source	Reuters
Secsource: Date	March 3, 1996 11:30
Primsource: Source	
Incident: Date	March 3, 1996
Incident: Location	Jerusalem
Incident: Type	Bombing
Hum Tgt: Number	"killed: 18"
	"wounded: 10"
Perp: Organization ID	

Message: ID	**TST-REU-0002**
Secsource: Source	Reuters
Secsource: Date	March 4, 1996 07:20
Primsource: Source	Israel Radio
Incident: Date	March 4, 1996
Incident: Location	Tel Aviv
Incident: Type	Bombing
Hum Tgt: Number	"killed: at least 10"
	"wounded: 30"
Perp: Organization ID	

Message: ID	**TST-REU-0003**
Secsource: Source	Reuters
Secsource: Date	March 4, 1996 14:20
Primsource: Source	
Incident: Date	March 4, 1996
Incident: Location	Tel Aviv
Incident: Type	Bombing
Hum Tgt: Number	"killed: at least 13"
	"wounded: more than 100"
Perp: Organization ID	"Hamas"

Message: ID	**TST-REU-0004**
Secsource: Source	Reuters
Secsource: Date	March 4, 1996 14:30
Primsource: Source	
Incident: Date	March 4, 1996
Incident: Location	Tel Aviv
Incident: Type	Bombing
Hum Tgt: Number	"killed: at least 12"
"wounded: 105"	
Perp: Organization ID	"Hamas"

Figure 2.4: Examples of MUC-4-Style Templates

on deep generation. SUMMONS' speciality is that it compresses several descriptions about the same event from multiple news stories. It takes MUC-4 style templates as input, e.g. the templates given in figure 2.4 (taken from Radev and McKeown 1998, pp. 487-488; the corresponding original newspaper texts are reproduced in figure 2.5). The compression strategy in SUMMONS is specific both to the domain (terrorist activities) and to the text type and situation (journalistic writing, publishing at successive times):

- *Change of perspective:* If the same source reports conflicting information over time, report both pieces of information.
- *Contradiction:* If two or more sources report conflicting information, choose the one that is reported by *independent* sources.

TST-REU-0001
JERUSALEM - A Muslim suicide bomber blew apart 18 people on a Jerusalem bus and wounded 10 in a mirror-image of an attack one week ago. The carnage by Hamas could rob Israel's Prime Minister Shimon Peres of the May 29 election victory he needs to pursue Middle East peacemaking. Peres declared all-out war on Hamas but his tough talk did little to impress stunned residents of Jerusalem who said the election would turn on the issue of personal security.

TST-REU-0002
JERUSALEM - A bomb at a busy Tel Aviv shopping mall killed at least 10 people and wounded 30, Israel radio said quoting police. Army radio said the blast was apparently caused by a suicide bomber. Police said there were many wounded.

TST-REU-0003
A bomb blast ripped through the commercial heart of Tel Aviv Monday, killing at least 13 people and wounding more than 100. Israeli police say an Islamic suicide bomber blew himself up outside a crowded shopping mall. It was the fourth deadly bombing in Israel in nine days. The Islamic fundamentalist group Hamas claimed responsibility for the attacks, which have killed at least 54 people. Hamas is intent on stopping the Middle East peace process. President Clinton joined the voices of international condemnation after the latest attack. He said the "forces of terror shall not triumph" over peacemaking efforts.

TST-REU-0004
TEL AVIV (Reuters) - A Muslim suicide bomber killed at least 12 people and wounded 105, including children, outside a crowded Tel Aviv shopping mall Monday, police said. Sunday, a Hamas suicide bomber killed 18 people on a Jerusalem bus. Hamas has now killed at least 54 people in four attacks in nine days. The windows of stores lining both sides of Dizengoff Street were shattered, the charred skeletons of cars lay in the street, the sidewalks were strewn with blood. The last attack on Dizengoff was in October 1994 when a Hamas suicide bomber killed 22 people on a bus.

Figure 2.5: Articles Corresponding to Templates in Figure 2.4

- *Addition:* If additional information is reported in a *subsequent* article, include the additional information.

- *Refinement:* Prefer more specific information over more general one (name of a terrorist group rather than the fact that it is Palestinian).

- *Agreement:* Agreement between two sources is reported as it will heighten the reader's confidence in the reported fact.

- *Superset/Generalization:* If the same event is reported from different sources and all of them have incomplete information, report the combination of these pieces of information.

- *Trend:* If two or more messages reflect similar patterns over time, these can be reported in one statement (e.g. three consecutive bombings at the same location).

- *No Information:* Report the lack of information from a certain source when this would be expected.

New templates are generated by combining other templates. The most important template, as determined by heuristics, is chosen for generation.

The content planner assigns values to realization flags (McKeown et al., 1994) related to discourse features such as "similarity" and "contradiction" which guide the choice of connectives and control local choices such as tense and voice in later generation steps. These switches also govern the presence or lack of certain constituents, in order to satisfy anaphora constraints and to avoid repetition of constituents. SUMMONS uses a domain ontology for lexical choice, to enrich the input and to make generalizations. The sentence generator used is FUF (Elhadad, 1993; Robin, 1994) which employs SURGE, a large systemic grammar of English. The output of this process is the following summary:

> *Reuters reported that 18 people were killed in a Jerusalem bombing Sunday. The next day, a bomb in Tel Aviv killed at least 10 people and wounded 30 according to Israel Radio. Reuters reported that the radical Muslim group Hamas had claimed responsibility for the act.*

The fact that this summary is deep-generated is illustrated by the change of voice in the first sentence compared to its source (TXT-REU-0001), the change of tense in the third sentence from simple past to past perfect, the replacement of the phrase *"the Islamic fundamentalist group Hamas"* by *"the radical Muslim group Hamas"* (TXT-REU-0003) and the occurrence of the term "*the next day*" which did not appear in the original text, but was added by SUMMONS during the combination and surface realization phase.

A similar, but more surface-oriented approach is given in Paice and Jones (1993) for scientific papers in the field of crop husbandry. The slots in their template (cf. figure 2.6, taken from Paice and Jones 1993, p. 71) are also domain specific,

	Paper 1	Paper 2
SPECIES:	*potato*	*winter wheat*
CULTIVAR:		
HIGH LEVEL PROPERTY:	*yield*	*each field a grid*
LOW LEVEL PROPERTY:		
PEST:	Powdery mildew	*Brent Geese Branta*
AGENT:		
INFLUENCE:		
LOCATION:	*York, Lincoln and Peterbourgh, England*	*Deepsdale Marsh, Burnham, Deepdale*
TIME:		*1985, 1986*
SOIL:		
CLIMATE:		
TREATMENT:		
PROCESS:		
NUTRIENT:		

Figure 2.6: Paice and Jones' (1993) Template for Agricultural Articles

Paper 1:

Title: The assesment [sic] of the tolerance of partially resistant potato clones to damage by the potato cyst nematode Globodera pallida at different sites and in different years.

Ann. Appl. Biol., 1988, 113:79-88

This paper studies the effect the pest G. pallida has on the yield of potato. An experiment in 1985 and 1986 at York, Lincoln and Peterbourgh, England was undertaken. These results indicate clearly *that* there are consistent differences between potato cultivars in their tolerance of damage by PCN as measured by proportional yield loss.

Paper 2:

Title: The effect on winter wheat of grazing by Brent Geese *Branta Bernicla*

Journal of Applied Ecology, 1990, 27:821-833

This paper studies the effect of Brent Geese Branta on the each field a grid of winter wheat [sic]. The experiment took place at Deepdale Marsh, Burnham, Deepdale. The fact that ear density increased due to grazing in one yield indicates that there is probably little value in the farmer sowing seed at a higher density in an attempt to compensate for geese grazing.

Figure 2.7: Paice and Jones' (1993) Abstracts for the Papers in Figure 2.6

e.g. SPECIES, CULTIVAR and PEST. The concepts are identified by a heuristic pattern matching procedure, where patterns such as "*effect of* INFLUENCE *on* PROPERTY *of/in* SPECIES" are identified in text. Candidate strings for a certain slot are weighted according to their frequency and the contexts where they appeared. Oakes and Paice

(1999) introduce an automated process to generate the search patterns automatically from text.

The abstracts, cf. figure 2.7 (taken from Paice and Jones 1993, p. 74), are generated in a much simpler fashion than Radev and McKeown's. The first sentence in each abstract is generated by slotting the best candidate strings into a fixed natural language template. Note that when a wrong string has been identified, such as the string *"each field a grid of"* in the second abstract, this might lead to ungrammatical output. The second sentence in each abstract is added by traditional text extraction: if a phrase like "*results indicate that*" (underlined in figure 2.7) is encountered, the sentence is added, in the hope that this turns the abstract into an informative one.

In fact-extraction templates, domain-knowledge is hard-wired into the slot definitions, and semantic relations between the slots are known *a priori*, e.g., the knowledge that it is the PERPETRATOR of a terrorist act who causes the killing or wounding of the HUMAN TARGETS. The depth of representation and the additional knowledge about semantic relationships between slots has clear advantages: it is possible, on the basis of domain-specific templates, to generate high-quality abstracts which read well and which are logically well-structured, as exemplified by Radev and McKeown's and Paice and Jones' summaries.

One of the disadvantages of such domain-specific approaches is the huge knowledge engineering efforts required to hard-wire the knowledge into the recognizers. Worse still, the whole machinery (template filling and, as a result, summarization) is not robust enough to react to unforeseen events in the texts. Only text segments that fit the expectations expressed by the situation slots can be handled. For instance, in the SUMMONS example only those aspects which have been anticipated in the template can be treated in the summary, namely the effects of the attack in terms of physical damage. All the other information in the original text is ignored, e.g. information about Mr. Peres and his prospects in the election (an important part of Text TST-REU-0001), or the future of the peace process and the international reaction to the attack (additional information in Text TST-REU-0003). Paice and Jones can similarly only process articles from a narrow subject field.

Spärck Jones (1999) calls fact extraction methods "what you know is what you get" techniques (p. 2), as they come with "the disadvantages that the required type of information has to be explicitly (and often effortfully) specified and may not be important for the source itself" (p. 3).

In sum, we have seen that the state of the art in automatic summarization is far from creating fluent summaries of unrestricted text which characterize the text's meaning well. However, there are two practical approaches which manage to fulfill some of the requirement of this task. We will in the following suggest our own approach.

2.3. A New Approach

In our review of current abstracting techniques, we found the following requirements for a new type of automatically generated document surrogate:

- It should be more flexible towards the text than fact extract based summaries are, while retaining some of the expressiveness of these.
- It should contain more information than text extracts, while retaining some of the generality and robustness of these.
- It should be more adaptive with respect to other tasks and other users than manual summaries are, while retaining the good characterization of the article achieved by these.
- It should include types of information not typically occurring in manual summaries (e.g. related work and its relation to the current work), while integrating this information with all other aspects.

2.3.1. Design of the Approach

2.3.1.1. General Design Criteria

When designing a new document surrogate, we started from the requirement of robustness. Robustness is indeed imperative, as we are working with unrestricted, naturally occurring text; such "real-life" text is a rough species. As a direct result, we decided to take orthographic sentences as unit of annotation, in analogy to most text extraction methods. Sentences can be identified robustly; smaller units seem fraught with problems. The concept of a clause, for example, has had linguists arguing for a long time.

Of course, a document surrogate based on textually extracted sentences presupposes that sentences which can act as parts of summaries are indeed found in the document, as Radev and McKeown (1998) point out. If this is not the case, nothing but deep-generation will help. However, we assume that explicit material for summaries will be available, due to the authors' motivation to formulate their important claims clearly.

One of our central observations is that the importance of a sentence within the whole text is crucially influenced by its rhetorical status: depending on whether the sentence describes the purpose of the research, the conclusion, or the author's criticism of other research, the content of a given sentence might be more or less useful for a given information need. For example, sentences which describe weaknesses of previous research can provide a good characterization of the scientific articles in which they occur, since they are likely to also be a description of the problem the paper is intending to solve. Take a sentence like *"Unfortunately, this work does not solve problem X"*: if X is a shortcoming in somebody else's work, the sentence might be a very good candidate for extraction. However, a very similar-looking sentence can play a completely different rhetorical role: if X refers to limitations of the approach presented in the paper, the sentence is *not* a good characterization of the article at all.

Our novel contribution is that we attach *additional rhetorical information* to the extracted sentences, in the form of fixed labels. The purpose of the labels is to capture the global context in which the sentence occurred in respect to the overall argumentation in the document. In contrast to fact extraction methods, the semantics of these labels is not defined by domain-specific knowledge, as this was the reason for the inflexibility which plagues fact extraction methods. This is in the line of Kircz (1991) and Sillince (1992) who have argued that *rhetorical* (or argumentative) indexing will provide more domain-independence in document retrieval applications than semantic indexing does. The exact definition of the labels will be given in section 2.3.2 and justified in chapter 3. As a result of how the labels are defined, they should apply equally well to articles coming from different disciplines; the approach is thus *domain independent but text type dependent*.

Some of these labels we define will encode different types of *connections* between articles: contrastive vs continuative mentions of other work, as motivated in section 2.1.2. The advantages of such a typing of links become apparent for large volume search, where a pre-sorting by type of link will save the user valuable time. However, the typing is subjective in nature (cf. section 3.2.2). Humans might disagree about certain cases, and a system performing the differentiation will sometimes make errors. We are aware of this risk, but think that the advantages outweigh the risks. Additionally, we invest some effort to measure the subjectivity of such decisions.

It is the working hypothesis of this thesis that shallow *argumentative* analysis is a promising approach for document characterization in a document retrieval environment. We take the deliberate decision not to model the *scientific content* of the article—in contrast to other approaches, which shallowly model content by term frequency methods (Salton et al., 1994b), lexical chaining methods (Baldwin et al., 1998; Barzilay and Elhadad, 1999), TextTiling (Hearst, 1997) or lexical similarity (Kozima, 1993). One of the reasons for our decision is the observation that even in human summarization it is not always the case that knowledge-intensive methods are the method of choice. Cremmins (1996) states that professional abstractors do not attempt to fully "understand" the text, but use surface-level features such as headings, key phrases and position in paragraphs. They also use *discourse* features such as overall text structure to organize abstracts and extract information. Endres-Niggemeyer et al. (1995) found that they

- prefer top-level segments of documents,
- build topic sentences,
- consider beginnings and ends of units as relevant,
- examine passages and paragraphs before individual sentences,
- exploit document outlines,
- pay attention to document formatting,
- determine the role of each section in the problem-solving process by reading the first and last sentence of each section or each paragraph and
- paraphrase relations between theme and in-text summaries.

However, our emphasis on the rhetorical side of the analysis does not mean that we believe that domain knowledge should never be included in a summarizer for scientific articles. On the contrary, scientific knowledge about the contents of the articles is undoubtedly going to improve the overall summarization process. Our long-term vision is that a better system would incorporate both *form* and *content* approaches, as we expect them to complement each other perfectly by recovering different aspects of meaning in the article. However, given the state of the art, we feel it is currently most promising to use shallow approaches of *form* rather than *content*.

The fundamental question, of course, is the question of depth of analysis, to which we will return in detail in chapter 5. Our approach will opt for robust, low-level techniques, because we believe that many of the problems encountered can be successfully addressed with fairly shallow techniques. Our approach is corpus-based: we will observe or learn features from a large amount of naturally occurring text. In sum, our approach

- uses shallow analysis;
- relies on sentences as units of extraction and analysis;
- does not model scientific content;
- attaches rhetorical information to sentences, e.g. the type of relation to other work.

The document surrogate we sketched so far bears comparison to structured abstracts, as sentences are classified into different types of information. Therefore, we will now review the literature on structured abstracts.

2.3.1.2. Structured abstracts

The literature on abstracting has identified the following four *content units* for informative summaries of articles in the experimental sciences (ANSI, 1979; ISO, 1976; Rowley, 1982; Cremmins, 1996):

- PURPOSE/PROBLEM
- SCOPE/METHODOLOGY
- RESULTS
- CONCLUSIONS/RECOMMENDATIONS

There is more disagreement about "peripheral" content units, such as RELATED WORK, BACKGROUND, INCIDENTAL FINDINGS and FUTURE WORK. According to Alley (1996), BACKGROUND is a useful content unit in an abstract if it is restricted to being the first sentence of the abstract (p. 22). Other authors (Rowley, 1982; Cremmins, 1996) recommend not to include any background information at all. Similar disagreement concerns the content unit RELATED WORK, as already discussed.

Buxton and Meadows (1978) provide a comparative survey of the contents units in summaries in the physics domain. They studied which rhetorical section in the source text (*Introduction–Method–Result–Discussion*) corresponds to the information in the summaries and found, for example, that summaries tend not to report material from the *Method* section. Milas-Bracovic (1987) performed a similar experiment on sociological and humanities summaries. Tibbo (1992) compares science (chemistry), social science (psychology) and humanities (history) with respect to the following content categories: BACKGROUND, PURPOSE/SCOPE, HYPOTHESES, METHODOLOGY, RESULTS, and CONCLUSIONS. Although the ANSI standard claims applicability of the above-mentioned four information units for abstracting in the social sciences and humanities as well, she found that fewer than 40% of the sentences in the history summaries fell into one of the ANSI categories.

Some innovative approaches suggest completely new information units and new structures. Trawinski (1989) introduces *problem structured abstracts*, with the main categories DOCUMENT PROBLEM, PROBLEM SOLUTION and TESTING METHOD, RELATED PROBLEMS, and 63 more fine-grained content elements such as SPECIFICATION OF OBJECTS USED IN TESTING and POSSIBLE USAGE AREAS IN SCIENCE. Broer (1971) uses graphic block-like units in his two-dimensional summaries, with the following units: WHAT? TITLE, WHAT/WHY? – INSTRUMENT, WHAT/WHY? – PRELIMS, WHAT? – CONSTRUCTION, HOW? – BASIC, HOW? – AID and WHY? – PERFORMANCE. His approach sounds promising but has not been used in practice.

Liddy (1991) showed experimentally that professional abstractors use an internalized building-plan when they write summaries. Her description of the components of summaries of empirical articles is based on professional abstractors' intuitions and a corpus of summaries.

Figure 2.8 gives an overview of the components (taken from Liddy 1991, p. 71). The seven most important components ("*prototypical components*") are displayed in capitals and bold face. The next level of importance ("*typical components*") is shown in capitals. The components found by Liddy cover short text spans (parts of sentences rather than sentences) and they can be embedded recursively into each other. Liddy concludes that abstractors, even if they might not choose the same sentences, still choose the same *type* of contents when they fill the fixed building-plans.

In the medical field, structured abstracts (Adhoc, 1987; Rennie and Glass, 1991) have long replaced free text summaries. Abstract information is given using prescribed headings which are dependent on the type of research being reported. Rather elaborate rules for their preparation have been established (cf. for example,

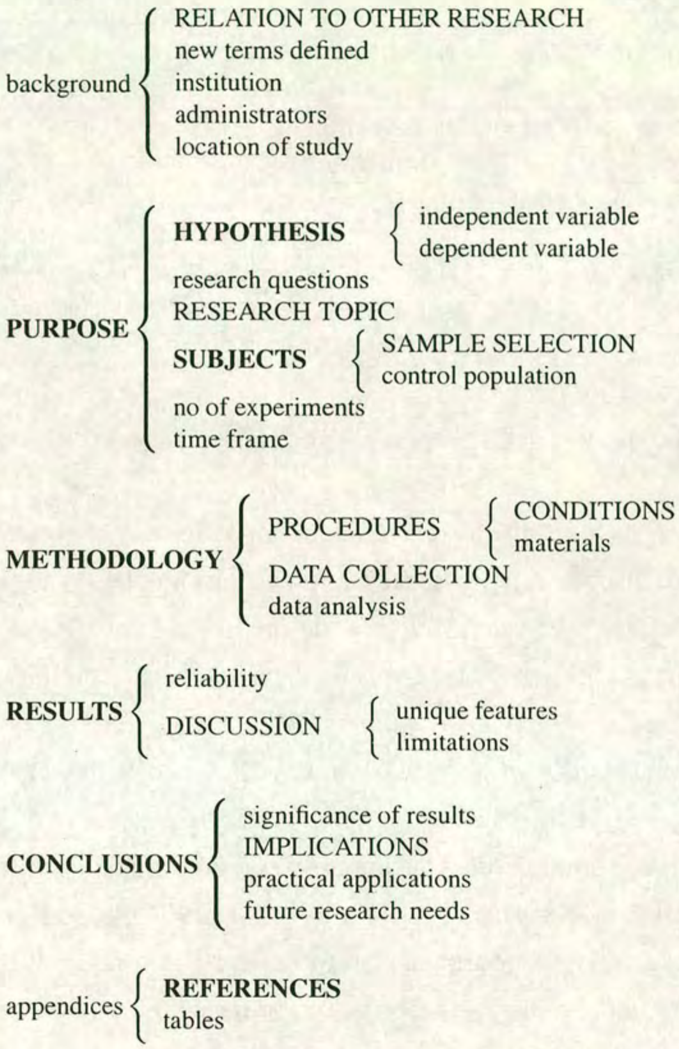


Figure 2.8: Liddy's (1991) Empirical Summary Components

Haynes (1990)). The following headings are used for descriptions of clinical trial reports in the *Annals of Internal Medicine*: BACKGROUND, OBJECTIVE, DESIGN, SETTING, PATIENTS, INTERVENTIONS, MEASUREMENTS, RESULTS and CONCLUSIONS. For reviews, headings include OBJECTIVE DATA SOURCES and STUDY SELECTION. Summaries in the *Archives of Dermatology* (Arndt, 1992) are structured into: BACKGROUND/DESIGN, RESULTS, CONCLUSIONS (CLINICAL), BACKGROUND/OBSERVATIONS and CONCLUSIONS (OBSERVATIONAL).

Several researchers found problems with the application of structured abstracts. Salager-Meyer (1992) researches empirically the linguistic and discoursal quality of

Task	Information required
Browsing the Literature	OBJECTIVES and CONCLUSIONS of a clinical study
Evaluating Clinical Studies	EXPERIMENTAL DESIGN plus CONCLUSIONS of the research (STUDY TYPE, STATISTICS, LIMITATIONS)
Matching Patients with Clinical Studies	ELIGIBILITY AND EXCLUSION CRITERIA, EXPERIMENTAL SETTING
Treating/Counseling Patients	INTERVENTIONS, RISK FACTORS, DIAGNOSTIC TESTS, ADVERSE EFFECTS and CONCLUSIONS
Planning Clinical Research	OBJECTIVE, CONCLUSIONS, DISCUSSION of UNANSWERED QUESTIONS and FUTURE WORK, LIST OF REFERENCES

Figure 2.9: ACP's Annals Extracts: Tasks and Components

medical summaries, in connection to content units. She found almost half to be "poorly structured", i.e. discoursally flawed. Froom and Froom (1993) showed that structured abstracts in *Annals of Internal Medicine* do not always contain all of the information requested in the guidelines for authors, even when the information needed was present in the article itself.

However, Hartley et al. (1996) and Hartley and Sydes (1997) present experiments which give evidence that structured abstracts are easier to read and overall more efficient than prose summaries. Hartley (1997) argues that structured abstracts should also be applied to social sciences. Taddio et al. (1994), based on a larger study of 300 summaries from three journals, also found that the structured abstracts were more likely to contain more complete information of research importance than unstructured abstracts were.

A new summarization/extraction application in the medical domain tests the plausible assumption that task flexibility can be realized based on such content units: the American College of Physicians (ACP) has recently started providing *task-specific* summaries for the papers in *Annals of Internal Medicine* (ACP online, 1997; Wellons and Purcell, 1999). There is a choice of five different types of (manually created) extracts for each paper; each of the five types is geared towards a different medical tasks. These tasks have been identified as frequently recurring in the different types of professional work of the readership of the *Annals*. Each of these tasks requires a different type of information from the medical articles, cf. figure 2.9.

And finally, Buckingham Shum and colleagues propose a specific meta data scheme for expressing relationships between articles (Shum, 1998; Sumner and Shum, 1998; Shum et al., 1999). It is a meta-data scheme for a Scientific Knowledge Web (SKW) of scientific papers in the field of HCI (Human–Computer Interaction) which concentrates on scholarly discourse, and the expression of relations between papers. The status of the units of this document surrogate is not anchored in any scientific domain knowledge, but rather in higher-level aspects which connect the instances of research, e.g. similarities and differences between scientific approaches. We will take the same approach in the design of our document surrogate. Their suggestion is unusual in its emphasis on *relations* between pieces of research, another aspect which has inspired the design of our document surrogate. An example for a representation of a paper according to this meta-description can be seen in figure 2.10 (taken from Shum 1998, p. 19).

There are 10 relations which describe how scientific works might be related to each other: ANALYSES, SOLVES, DESCRIBES-NEW, USES/APPLIES, MODIFIES/EXTENDS, CHARACTERIZES /RECASTS, EVALUATES (SUPPORTS or PROBLEMATISES or CHALLENGES).

The suggested concepts are entities which are important in the domain (HCI), namely the following 9 categories: APPLIED-PROBLEM, THEORETICAL-PROBLEM, METHOD, LANGUAGE, SOFTWARE, EVIDENCE, THEORY/FRAMEWORK, TREND,

REF: Smith, J. (1997) ATC Overload, Journal of ATC, 3 (4), 100-150

ANALYSES	APPLIED-PROBLEM	*Air traffic controller cognitive overload*
USES/APPLIES	THEORY/FRAMEWORK	*use of video, undergraduate university physics, student ability*
PROBLEMATISES	SOFTWARE	*GOMS cognitive modelling tools*
MODIFIES/EXTENDS	LANGUAGE	*Knowledge Interchange Format (KIF)*
CHARACTERIZES/RECASTS	TREND	*Electronic trading over the internet*
CHALLENGES	SCHOOL-OF-THOUGHT	*Postmodernism*
SUPPORTS	EVIDENCE	*multimedia, school chemistry teaching*

Figure 2.10: Shum's (1998) Design for Document Representations in a Scientific Knowledge Web (SKW)

SCHOOL-OF-THOUGHT. Each of the concepts can be further refined by keywords or names and connected to a reference or a URL.

The design of the SKW slots has not been verified by cognitive experiments with users, but is currently in a beta-testing phase, where researchers in the HCI field can contribute example encodings of their own papers, suggestions and comments. In the setup that Shum (1998) has in mind, a human expert would select one of these possible slots and fill them manually with domain-specific material, sometimes requiring background knowledge and inference. This is typical for meta-data approaches, which assume in general that humans (authors or indexers) provide mark-up. Shum (1998) argues pessimistically about the task of filling the slots in his scheme by an automatic process:

> It is possible that useful information may be extracted through intelligent analyses of text, but often this information is not explicit in documents, but implicit in the minds of domain experts. (Shum, 1998, p. 16)

On the one hand, we welcome the meta-data approach because meta-indexing provided by authors can be expected to be of high quality. On the other hand, it might take some time before such meta-data approaches will have an impact on writer's behaviour when papers are written and submitted.

The main difference between our design and this scheme is the fact that our analysis is aiming to provide filling material *automatically*. As a result, the fillers which our planned document representation provides have to be of a much simpler kind: mere surface strings.

Another difference is that in Shum's approach nodes themselves are "neutral" (i.e., not associated with local semantic information); the only semantics that a node has comes from the links and its position in a research web. In our approach, the characterization of the paper on its own is also important. This has the advantage that papers can be summarized and characterized as single items without looking at their connections (which the system does not necessarily have knowledge of).

2.3.2. Rhetorical Document Profiles (RDPs)

The outcome of these design decisions is a new document surrogate. We call this document surrogate a *Rhetorical Document Profile* (RDP) because it consists of rhetorical units (slots) and because it profiles different kinds of information about the document. RDPs were designed to encode typical information needs of new readers in a systematic and structured way. Figure 2.11 shows an empty RDP.

1. Solution Identifier	—			
2. Specific Aim/Scope	—			
3. Background	Aim	Problem/Phenomenon		
	—	—		
4. Solution/Inventive Step	—			
5. Claim/Conclusion	—			
Rel. to Other Work	6. Rival/ Contrast	Reference	Solution ID	Type of Contrast
		[...]	—	—
		[...]	—	—
	7. Basis/ Continuation	Reference	Solution ID	Type of Continuation
		[...]	—	—
		[...]	—	—
External Structure	Headlines		8. Textual Structure	
	—		—	
	—		—	

Figure 2.11: An Empty Rhetorical Document Profile (RDP)

On the following pages, we will walk the reader through a *filled* RDP (namely the one for example article cmp_lg/9408011) slot by slot. This RDP was manually filled by us with textual material taken verbatim from the source article (excluding the human-written summary). These surface strings are often whole sentences, and sometimes segments of sentences. Slot fillers are identified by sentence numbers, which act as pointers into the original text where the textual material was extracted from (cf. sentence numbers in XML representation of the article, appendix B.1).

The exact filling criteria will be elaborated later. The solution displayed is *one* possible solution; as the filling criteria rely on human intuition, other solutions would have been possible too. We claim, however, that other humans would have filled the slots sufficiently *similarly*; chapter 4 will provide experimental evidence for this claim.

1. Solution Identifier —

Solution Identifier: Sometimes a paper introduces a new approach and gives it a name. Later papers might refer to it using that term. In our domain, these are often artefacts: names of programs, methods, algorithms or theories. Information about well-known methods in the field is extremely important to uninformed and partially informed readers (cf. section 1.1). Examples for what we will consider as identifiers for solutions are the following: *"the SPLATTER parser", "Maximum Entropy classifier", "Minimum Description Length (MDL)", "Data Oriented Parsing (DOP)", "the Centering algorithm*" and *"Rhetorical Structure Theory (RST)*". A solution identifier does not always have to be a proper name, but can be any other description, e.g. "*Hobbs' anaphora resolution algorithm*" or "*simulated annealing*".

Our example article does introduce a named solution: a new method which later articles refer to as "*soft word clustering*". But unfortunately, there is no explicit mention of this particular term in the example article itself. A similar expression ("*hierarchical "soft" clustering*") does appear in the author-written summary, but we decided not to use information from the summary. As it is, the slot remains empty.

2. Specific Aim/Scope

164 to group words according to their participation in particular grammatical relations with other words

10 how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves

44 how to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams

11 how to derive the classes directly from distributional data

46 learning a joint distribution of pairs from a large sample of pairs

22 we will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs

45 we will only address the more specific problem in which the objects are nouns and the contexts are verbs that take the nouns as direct objects

The slot SPECIFIC AIM/SCOPE contains descriptions of the research goal specific to the article. We believe that fillers of this slot can be the single most characteristic information about a scientific paper (particularly if they occur in a sentence together with the methodology used).

Our example article happens to contain unusually many explicit mentions of the specific research goal. The slot-fillers differ in the level of abstraction at which they describe the research goal, and in their focus on a particular aspect of the problem. Some of them are paraphrases of each other, or contribute more detailed information. This leads to a certain degree of redundancy. Note that slot fillers **11** and **46** do not just talk about the research goals, but additionally give some information about the solution, i.e., *how* the task is solved. In general, it can be difficult to keep goals and solutions apart. Slot fillers **22** and **45** stand in the context of a contrastive *scope* delimitation: the authors stress that they do *not* classify verbs, just nouns.

3. BACKGROUND

AIM	PROBLEM/PHENOMENON
1 automatically classifying words	**4** The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.

BACKGROUND information divides into two kinds: BACKGROUND (AIM) can be considered as the paper's topic, a high level characterization of the task, e.g. "*machine translation*". In our example, the high level goal is the automatic classification of words. BACKGROUND (PROBLEM/PHENOMENON) gives high level problems in the field (in this case: data sparseness). If the paper aims at an explanatory account, then BACKGROUND (PROBLEM/PHENOMENON) can contain sentences describing phenomena to be explained.

4. SOLUTION/INVENTIVE STEP

164 a general divisive clustering procedure for probability distributions can be used [. . .]
12 we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN> for each word w.

The nature of the SOLUTION/INVENTIVE STEP depends on the type of discipline we are considering. In some empirical disciplines, a new empirical claim or a new hypothesis is the main innovation of the paper; the research goal, namely to verify or disprove the hypothesis, is left implicit. In those disciplines, the methodology is often standardized. In disciplines like computational linguistics, the main idea is often the technical solution (methodology) — exactly because there are few fixed rules as to which methodologies can be used.

In our case, there are some high-level descriptions of the innovative step: the authors apply a well-known general divisive clustering procedure, and part of their solution is to model word senses as clusters.

5. CLAIM/CONCLUSION

165 The resulting clusters are intuitively informative, and can be used to construct class-based word coocurrence models with substantial predictive power.

The CLAIM/CONCLUSION slot concerns explicit claims. Explicit claims, hypotheses and predictions are typically found in experimental papers. Even though this particular paper is a technical paper (something is engineered), we still encounter a claim. This claim, however, is not a claim about the scientific domain, but rather a meta-claim: it is a statement that the problem has been solved, and that the result makes sense. Such sentences, if correctly identified, can give valuable information about the paper's problem-solving process.

Two slots describe the relation of the current work to other work. The two categories are 6. CONTRASTIVE relations and 7. CONTINUATION of research relations. The slot RIVAL/CONTRAST approaches is filled with information on other work which is in a contrastive or comparative relationship to the given work, or information about a specific weakness of the other work. The other work can be identified either by a formal explicit reference or by a solution *identifier*, in analogy to the SOLUTION IDENTIFIER slot discussed on p. 58.

6. RIVAL/CONTRAST

REFERENCE	SOLUTION ID	TYPE OF CONTRAST
• [Hindle 1990] – **5**		**9** it is not clear how it can be used directly to construct word classes and corresponding models of association
• [Brown et al. 1992] – **13**	**13** other class-based modeling techniques	**13** Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information
• [Resnik 1992] – **11**		**11** preexisting sense classes (Resnik) vs. we derive the classes directly from distributional data
•	**43** agglomerative clustering techniques	**43** need to compare individual objects being considered for grouping (advantage of authors' method)
• [Church and Gale 1991] – **40**	**40** smoothing zero frequencies appropriately	**41** However, this is not very satisfactory as our goal is to avoid the problems of data sparseness by clustering words together

With respect to contrastive approaches, the authors seem to have identified certain weaknesses with Hindle's (1991) and Brown et al.'s (1993) work. There is also a contrast in task with Resnik (1992), and an advantage over both agglomerative clustering techniques and Church and Gale's (1991) approach.

7. BASIS/CONTINUATION

REFERENCE	SOLUTION ID	TYPE OF CONTINUATION
• [Rose et al. 1990] – **113**	**113** deterministic annealing	**113** The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering [Rose et al. 1990] . . .
• [Dagan et al. 1993] – **155**		**155** based on a suggestion by
•	**29** Kullback-Leibler (KL) distance	**29** used
• [Hindle 1993] – **19**		**19** automatically parsed by Hindle's parser
• [Church 1988] – **20**		**20** with the help of a statistical part-of-speech tagger
• [Yarowsky 1992] – **20**		**20** [with the help of] tools for regular expression pattern matching on tagged corpora

The BASIS/CONTINUATION describes work which provides a starting point for the current work, or which provides data, theoretical apparatus or methodology that the current work uses. It might also support the claims of the given paper, or fit in with the paper's claims without contradiction. Information about intellectual ancestry, i.e., the knowledge of who builds their work on who else's work, is of great importance to users trying to orient themselves in a new area (cf. section 1.1). Note that contrasted and continued research are not necessarily mutually exclusive classes. Researchers might use a certain work as starting point but identify problems with it which they then try to rectify.

In the example paper, the single most important continuation is the fact that the authors use Rose et al.'s annealing procedure. They also use Hindle's (1993) parser, Church's (1988) POS tagger, Yarowsky's (1992) regular expression tools and a commonly agreed upon statistical measure (KL). Also, they use a suggestion in a paper by Dagan et al. (1993).

EXTERNAL STRUCTURE

HEADLINES

1.	Introduction
1.1	Problem Setting
1.2	Distributional Similarity
2.	Theoretical Basis
2.1	Distributional Clustering
2.1.1.	Maximum Likelihood Cluster Centroids
2.1.2.	Maximum Entropy Cluster Membership
2.1.3.	Minimizing the Average KL Distortion
2.1.4.	The Free Energy Function
2.2.	Hierarchical Clustering
3.	Clustering Examples
4.	Model Evaluation
4.1.	Relative Entropy
4.2.	Decision Task
5.	Conclusions

8. TEXTUAL STRUCTURE

127 All our experiments involve the asymmetric model described in the previous section.

EXTERNAL STRUCTURE is a concerned with explicit representations of structure in the article: a simple listing of all headlines found in the text (sub-slot HEADLINES) or explicit textual information about the section structure (sub-slot TEXTUAL STRUCTURE). In this paper, only one explicit statement about textual structure was found (and even this one is not a clear case). It is a reference back to the previous section, and can give some indication of the contents of that section.

The full RDP is given in appendix B.3; appendix B.4 lists the sentences from the original text corresponding to the textual material in the RDP.

We have by now redefined the goal of the thesis: to verify if it is possible to automatically identify these types of information in real world texts. The output of this thesis, namely relevant textual material for the RDP slots, could be regarded as a final result. We believe that lists of RDP slot fillers are already better textual extracts than those provided by today's sentence extraction methods. Additionally, we predict that RDP slot fillers would provide useful information for human abstractors, shortening the time it takes them to construct a full textual abstract. Conceptually however, the extraction step described in this thesis was designed in such a way that its output would be of greatest possibly usability to the follow-on processing steps.

We will now discuss the use of RDP type information in a document retrieval environment.

2.3.3. RDPs for Tailored Summaries

If RDPs could be automatically compiled in an off-line fashion for each document in a large collection of papers, this would have definite advantages for document retrieval. RDPs in themselves provide a detailed, tabularized summary of the article. Users could get an overview of the contents of the paper by directly scanning them. However, RDPs are big document surrogates containing a lot of redundancy. Users might not want to invest the time to directly read them.

Users who prefer more traditional summaries could be provided with user, length and task tailored summaries generated from RDPs. Imagine two kinds of users (informed vs. uninformed readers), three kinds of "tasks" (general purpose, contrastive use of summaries, determining intellectual ancestry between papers) and two lengths of summaries (longer vs. shorter). In figure 2.12, simple recipes (or building-plans) for summaries are given for combinations of expertise, length and task. The building-plans vary in the number and type of individual slot fillers which are included in the summary. Following from our considerations in section 2.1.1, the building-plan mirror the following intuitions about differences in expertise:

- More background material (e.g. in the introduction) is needed for uninformed readers, whereas informed readers do not require any background information. For uninformed readers, the approaches of other researchers are *described*; for informed readers, they are only *identified* (by direct citation or by solution identifier).

- This should make summaries for uninformed reader in general longer than summaries for informed readers.

- Sentences with more general terms are preferred for uninformed readers, and sentences with more technical terms for informed readers. Sentence **44** in figure 2.13, which contains for example the specific term *"ngram", "linguistic objects"*, was chosen as expression of the SPECIFIC AIM for informed readers, whereas sentence **164** in figure 2.17 was chosen for uninformed readers, as it contains more general terms (*"group", "words", "grammatical relations"*).

The second factor we considered was *task-tailoring*:

- General purpose summaries consist of as few SPECIFIC AIM sentences as possible, in order to avoid redundancy.

- Longer general purpose summaries should include some SOLUTION/INVENTIVE STEP material, in order to simulate informative summaries.

- For comparative or contrastive summaries, the "most important" rival approaches should be presented to the reader. One simple way to determine importance of an approach is by measuring how much space the description of the approach is given in the paper (see also a later discussion of this point in section 3.4).

- In analogy, the most important based-upon other work needs to be identified for intellectual-ancestry summaries.

We manually generated summaries to illustrate the building-plans. Many ways

	Informed reader	Uninformed reader
General purpose, short	*Summary 1:* 2 SPECIFIC AIM	*Summary 5:* 1 BACKGROUND (AIM) + 1 BACKGROUND (PROBLEM) + 2 SPECIFIC AIM
General purpose, longer	*Summary 2:* 2–3 SPECIFIC AIM + 1 INVENTIVE STEP	*Summary 6:* 1 BACKGROUND (AIM) + 1 BACKGROUND (PROBLEM) + 2–3 SPECIFIC AIM + 1 INVENTIVE STEP
Contrastive	*Summary 3:* 2 SPECIFIC AIM + 1–2 (SOLUTION ID + TYPE OF CONTRAST)	*Summary 7:* 1 BACKGROUND (AIM) + 1 BACKGROUND (PROBLEM) + 2 SPECIFIC AIM + 1–2 (DESCR. OF OTHER WORK + TYPE OF CONTRAST)
Ancestry	*Summary 4:* 2 SPECIFIC AIM + 1–2 (SOLUTION ID + TYPE OF CONTINUATION)	*Summary 8:* 1 BACKGROUND (AIM) + 1 BACKGROUND (PROBLEM) + 2 SPECIFIC AIM + 1–2 (DESCR. OF OTHER WORK + TYPE OF CONTINUATION)

Figure 2.12: Building-Plans for Task and Expertise Tailored Summaries

of arriving at the actual summary text are imaginable for this illustration, resulting in summaries of a different quality. We decided to select good candidates amongst the RDP slot fillers and to change them as little as possible. The output is enriched with templates, and some minimal surface repair is performed in order to make the result easier to read.

We simulated a selection process amongst RDP slot fillers for each slot given in the building-plan. The rules for choosing a given sentence for a slot over its competitors are that it has to be a) minimally similar to any other chosen sentence for that slot, in order to reduce redundancy and b) maximally similar to as many other candidates for that slot as possible—which are, as a consequence of a), *not* chosen. The argumentation for this is due to Edmundson (1969) who voiced the intuition that more important material appears redundantly in text. The occurrence of similar slot fillers thus raises our confidence that the given slot fillers are good characterizations for the semantics of its slot.

Surface repair can be imagined as follows: for a summary sentence about research goal, strings are taken from the corresponding RDP slot, the semantic verb is identified and transformed into the syntactic form fitting to the template context (*"This paper's goal is to"*). Template material is shown underlined in the following summaries.

As there is more space for the discussion of other approaches in summaries for uninformed readers, it is not always necessary to process the sentences further. In contrast, generating concise sentences for informed readers is a more complex task, as the material needs to be found from different sources and assembled correctly. Consider, for example, the sentence constructed from sentences **5** and **9** in figure 2.15, where sentence **5** supplies the solution identifier and sentence **9** supplies the criticism/contrast. In order to correctly handle comparison and negation in sentences **5/9** and **14**, some more complex templates or deeper generation mechanisms would have to be used here.

44 *This paper's goal is to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.* **22** *More specifically: the goal is to classify nouns according to their distribution as direct objects of verbs.*

Figure 2.13: Summary 1: Informed Reader, General Purpose, Short

44 *This paper's goal is to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.* **22** *More specifically: the goal is to classify nouns according to their distribution as direct objects of verbs.* **11** *The goal is to derive the classes directly from distributional data.* **164** *A general decisive clustering procedure for probability distributions is used.*

Figure 2.14: Summary 2: Informed Reader, General Purpose, Longer

44 *This paper's goal is to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.* **22** *More specifically: the goal is to classify nouns according to their distribution as direct objects of verbs.* **5** *Unlike, [Hindle 1990],* **9** *this approach constructs word classes and corresponding models of association directly.* **14** *In comparison to [Brown et al. 92], the method is combinatorially less demanding and does not depend on frequency counts for joint events involving particular words, a potentially unreliable source of information.*

Figure 2.15: Summary 3: Informed Reader, Contrastive

44 *This paper's goal is to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.* **22** *More specifically: the goal is to classify nouns according to their distribution as direct objects of verbs.* **113** *It uses the deterministic annealing procedure introduced by [Rose et al 1990].*

Figure 2.16: Summary 4: Informed Reader, Intellectual Ancestry

1 *This paper's topic is to automatically classify words.* **4** *The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.* **164** *This paper's specific goal is to group words according to their participation in particular grammatical relations with other words,* **22** *more specifically to classify nouns according to their distribution as direct objects of verbs.*

Figure 2.17: Summary 5: Uninformed Reader, General Purpose, Short

1 *This paper's topic is to automatically classify words.* **4** *The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.* **164** *This paper's specific goal is to group words according to their participation in particular grammatical relations with other words,* **22** *more specifically to classify nouns according to their distribution as direct objects of verbs.* **11** *Another goal is to derive the classes directly from distributional data.* **12** *The authors model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN> for each word w.*

Figure 2.18: Summary 6: Uninformed Reader, General Purpose, Longer

1 *This paper's topic is to automatically classify words.* **4** *The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.* **164** *This paper's specific goal is to group words according to their participation in particular grammatical relations with other words,* **22** *more specifically to classify nouns according to their distribution as direct objects of verbs.*
5 *[Hindle 1990] proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen.* **8** *In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events.* **9** *It is not clear how his notion of similarity can be used directly to construct word classes and corresponding models of association.*
13 *Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes [Brown et al. 1990].* **14** *Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information.*

Figure 2.19: Summary 7: Uninformed Reader, Contrastive

1 *This paper's topic is to automatically classify words.* **4** *The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.* **164** *This paper's specific goal is to group words according to their participation in particular grammatical relations with other words,* **22** *more specifically to classify nouns according to their distribution as direct objects of verbs.*
113 *The authors use a deterministic annealing procedure for clustering [Rose et al. 1990], in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter <EQN/> following an annealing schedule.*

Figure 2.20: Summary 8: Uninformed Reader, Intellectual Ancestry

The summaries read fluently and convey different kinds of information for different readers and different tasks. Manipulation of length and of syntactic constructions in the sentences is possible due to the rhetorical information coming from the RDP slots. This information is not domain-specific, in contrast to similar fact-extraction templates.

Multi-document summarization could also profit from RDPs for scientific articles: articles mentioning similar concepts in the same RDP slots might be candidates for collective characterization in one summary for all these articles. Documents returned by a users' query for the term "*Decision Tree Learning*" might be described ("summarized") as follows:

> *In your query results, there are 13 papers that have the term* `PP attachment` *in their* SPECIFIC AIM *slot. There are 33 papers with* `cross-validation` *in the* SOLUTION *slot.*

2.3.4. RDPs for Citation Maps

The information contained in RDPs can help users understand the relationship of one particular paper to other papers: either to papers contained in a set of search results, or to papers already known to the user.

We suggest generating a new construct called *local citation maps* on the fly for papers of interest. Figure 2.21 shows such a (manually created) citation map, including all those papers from our document collection which cite our example paper, Pereira et al. (1993). Each article of this starting set is displayed in a rectangle and identified by name of authors and year of publication. The map also shows articles referenced by these papers (i.e. those not contained in our document collection) which are displayed without rectangles. (The difference in status between articles within and outwith our collection is of course that we cannot trace the citations contained in the latter.)

The information contained in RDPs allows to display *typed* links, where the green links corresponds to CONTRAST ("contrasting the work to other work") and purple links to BASIS/CONTINUATION ("building the work onto previous solutions"). If no particular stance could be determined, a "neutral" citation link is displayed in black.

We claim that citation maps could help users picture document similarities and differences in an immediate and natural way. Especially for uninformed searchers, such a representation of links would be extremely useful for a local exploration of a wide range of questions.

Certain kinds of similarities and differences between papers can be seen at first glance. Figure 2.21 shows that Nitta and Niwa (1994) and Resnik (1995) cite Pereira et al. (1993) and the other four papers in our collection only contrastively, and they both cite some other papers, and in a contrastive way (e.g. Schütze (1993) and Hirst (1991)). Two of the other three papers, on the other hand, also form a natural sub-cluster: Dagan

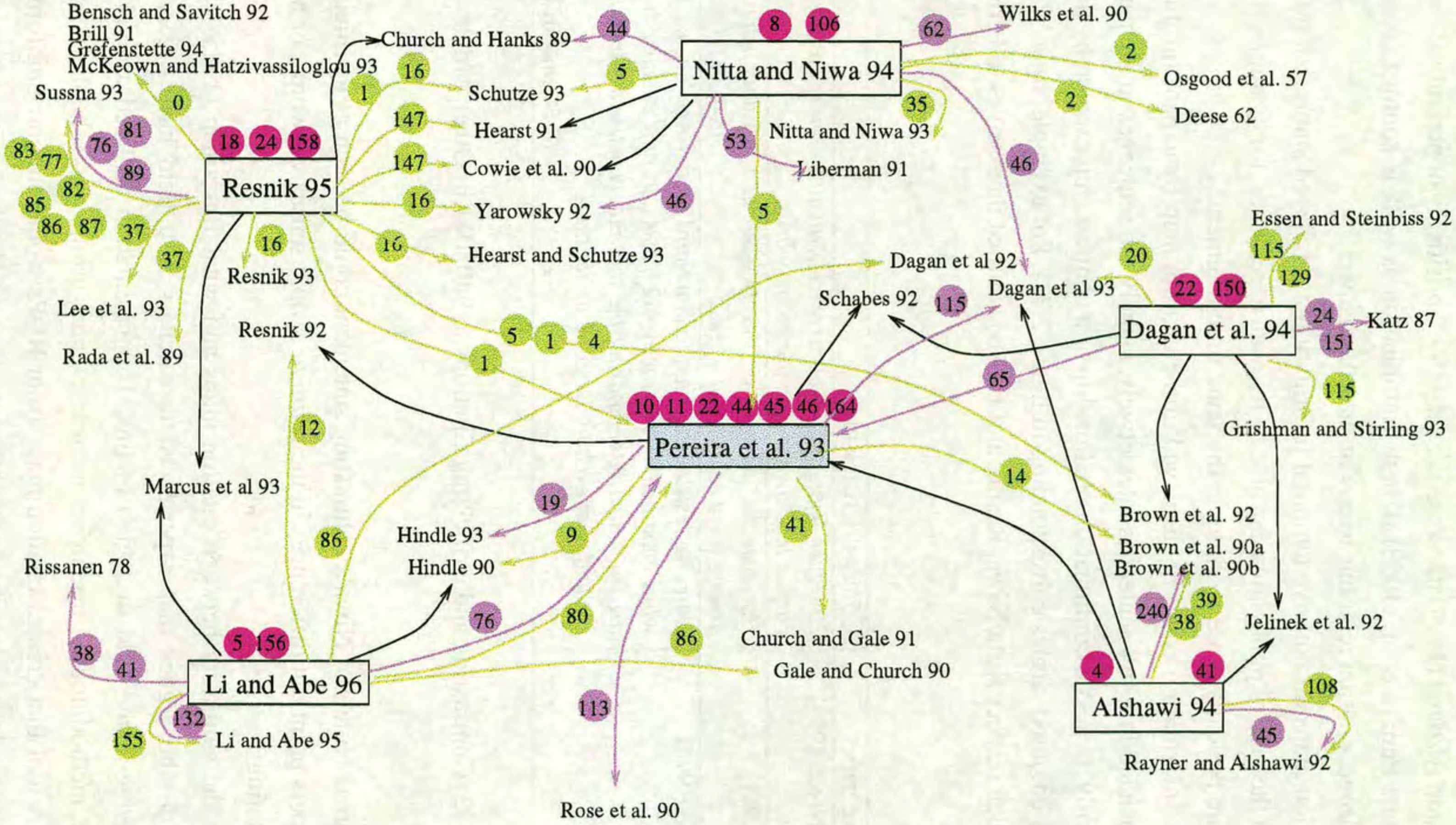


Figure 2.21: Citation Map for Document 9408011

et al. (1994) and Alshawi (1994) cite Pereira et al. (1993) positively or neutrally. Li and Abe (1996) cite Pereira et al. (1993) in both continuation as well as contrast context and have no direct citation relations to any of the other papers.

Citation maps do not give temporal information a privileged status, but information about the time of publication can also be relevant to searches: for example, rival approaches are typically those working in the same time fragment.

More information could be displayed in the citation map by expansion: links could be expanded into full sentences interactively, namely the sentences in the paper which explicitly express a continuation relationship or a contrast (represented by their numbers and coloured circles corresponding in figure 2.21). For example, figure 2.22 shows in which respect Nitta/Niwa, Resnik and Li/Abe contrast themselves to Pereira et al. (1993).

Contrasting paper	Contrast/Criticism
[Nitta and Niwa, 1994]	*However, using the co-occurrence statistics requires a huge corpus that covers even most rare words.* (S-5, 9503025)
[Resnik, 1995]	*However, for many tasks, one is interested in relationships among word senses, not words.* (S-1, 9511006)
[Li and Abe, 1996]	*Here, we restrict our attention on 'hard clustering' (i.e., each word must belong to exactly one class), in part because we are interested in comparing the thesauri constructed by our method with existing hand-made thesauri.* (S-80, 9605014)

Figure 2.22: Contrasting and Criticizing Citations to 9408011 in Other Articles

Whereas Nitta and Niwa's contrasting statement could be seen as a criticism, the other papers point out differences in their *aim or scope*: senses vs. words, or hard vs. soft clustering.

Note the similarity between citation maps and what Bazerman (1985) calls *research maps*: he argues that experienced researchers in a field have organized their knowledge in the field in a kind of linked representation centered around research goals, methodologies, researcher names, research groups and schools (cf. section 2.1.2). A tool that creates citation maps from RDPs would support uninformed users in acquiring their own mental research map more efficiently. Local and content-enriched citation maps present information in an immediate, powerful and natural way.

Uninformed users could start using citation maps without any knowledge of the terminology in the field. They get an overview of relations amongst papers and incidentally come across relevant terms in sentences which are displayed. This boot-strap knowledge will make subsequent keyword searches more efficient.

2.4. Conclusion

In this chapter we have looked at state-of-the-art summarization techniques. An overview of the paper-based world of hand-written summaries has shown that such summaries are of high quality but inflexible. They also do not provide much-needed information about contrastive and ancestral relations between similar articles. With respect to automatic summarization, we found that fact extraction methods, while providing informative output, are too domain-dependent and not robust enough towards unexpected turns in unrestricted texts—whereas text extraction methods, which are robust to the extreme, do not provide enough information about the extracted material. We have argued that what is missing is some form of context with respect to the overall document content. As a possible way out of this predicament, this chapter has introduced RDPs (Rhetorical Document Profiles).

- Similar to *text-extraction* methods, RDPs will use sentences as extraction units. In contrast to text-extraction output, RDPs contain information attached to each sentence, namely the information about the rhetorical status of a sentence with respect to the whole paper. This makes different kinds of postprocessing possible.

- Similar to *fact-extraction* approaches, summaries can be (re)generated, due to the information connected with the textual material. In contrast to fact-extraction templates, RDP slot semantics are not domain dependent: RDP slots do not encode anything about the subject matter of science. However, RDP slots are text type dependent.

- Similar to *human-written abstracts*, information about functional units in the document will help construct and structure the abstract in an RDP-based approach. In contrast to human-written summaries, RDPs provide information about connections between articles; they can be tailored to user expertise and task requirements.

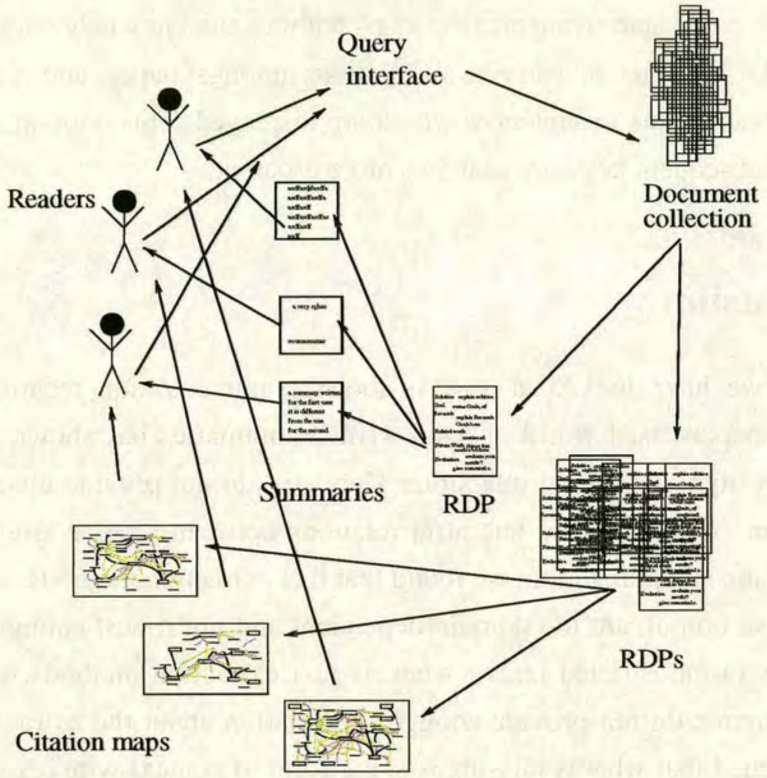


Figure 2.23: The Role of RDPs in a Document Retrieval Environment

- Similar to *citation-indexing tools*, RDPs provide information about relatedness of articles. In contrast to them, RDPs distinguish the *type* of links between documents and also provide *static, semantic* information about the document.

As figure 2.23 shows, RDPs could support scientists' information foraging activities in an actual document retrieval environment by providing the information needed for automatically generated, expertise and task tailored summaries and for citation maps.

This thesis will not go all the way in producing RDPs automatically—RDPs are highly informative document surrogates, the automatic generation of which is too ambitious a task for the scope of this thesis. Instead, this thesis will constitute the first step in the production of RDPs, namely the production of a list of sentences which are good slot fillers for RDPs.

In this context, the next chapter will place the concept of an RDP (which is a reader-centered construct) with the concept of argumentative zones in text (which is a writer-centered construct). It will pave the way for an automatic procedure for filling RDP slots, by looking at strategies for finding good slot fillers in running text.

Chapter 3

Argumentative Zoning

In the previous chapter, we motivated a new document surrogate, the RDP or rhetorical document profile. We showed that the RDP is a desirable construct in a document retrieval environment, as it provides the right kind of information for the flexible generation of summaries.

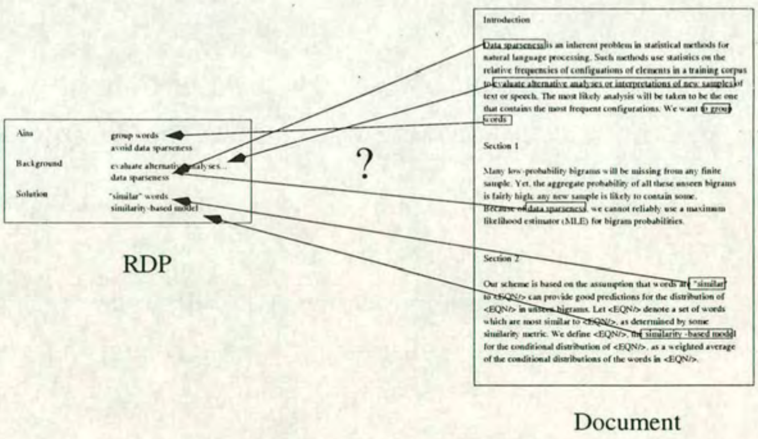

Aim
group words
avoid data sparseness
Background
evaluate alternative analyses...
data sparseness
Solution
"similar" words
similarity-based model
RDP
?
Introduction
Data sparseness is an inherent problem in statistical methods for natural language processing. Such methods use statistics on the relative frequencies of configuations of elements in a training corpus to evaluate alternative analyses or interpretations of new samples of text or speech. The most likely analysis will be taken to be the one that contains the most frequent configurations. We want to group words
Section 1
Many low-probability bigrams will be missing from any finite sample. Yet, the aggregate probability of all these unseen bigrams is fairly high; any new sample is likely to contain some. Because of data sparseness, we cannot reliably use a maximum likelihood estimator (MLE) for bigram probabilities.
Section 2
Our scheme is based on the assumption that words are "similar" to <EQN/> can provide good predictions for the distribution of <EQN/> in unseen bigrams. Let <EQN/> denote a set of words which are most similar to <EQN/>, as determined by some similarity metric. We define <EQN/>, the similarity-based model for the conditional distribution of <EQN/>, as a weighted average of the conditional distributions of the words in <EQN/>.
Document

Figure 3.1: From Documents to RDPs

In this chapter we discuss how to get from text to RDPs. Some constraints of the task were already discussed at the end of the previous chapter: our analysis will be shallow and robust, using full sentences as filling material, and it will aim at attaching rhetorical information to the extracted sentences (cf. figure 3.1).

In the previous chapter, the semantics of RDP slots was justified by the document retrieval task: the slots are defined by the kinds of information that *readers* want *out* of the text. In this chapter, we will define the slot semantics by looking at what the *writer* put *into* the text, in particular how she organized and structured her text. This has a parallel to the situation in summarization in general, about which Paris writes:

> Summarising depends on the recognition of both the intention of the writer in writing the original text (with respect to what he or she was trying to convey) as well as the goals and knowledge of the reader (why do they want a summary and how much do they know about the domain). (Paris, 1993, p. 1)

However, it is not obvious *what* kind of rhetorical information should define the slot semantics. We will see in section 3.1 that fixed section structure cannot offer much help. We base our structural analysis instead on a new model of prototypical scientific argumentation. The theory behind the model, described in section 3.2, is based on authors' communicative acts—-these communicative acts are predictable from text type-specific expectations. The model draws from different strands of research:

- *Argumentative moves:* Swales (1990) claims that there is a restricted range of prototypical argumentative goals that a writer of a scientific article has to fulfill, e.g., to convince her readers that the problem she addresses has some interest to the field (cf. section 3.2.1).

- *Authors' stance towards other work:* The field of Content Citation Analysis categorizes semantic relations between citing and cited work (cf. section 3.2.2).

- *Intellectual ownership:* Authorship in scientific discourse is typically explicitly given: either the statements are presented as own work, as well-known facts in the field, or as other authors' claims. We will argue in section 3.2.3 that a segmentation based on this distinction is an essential step for our task. To our knowledge, this aspect of scientific text has not received any attention in computational approaches yet.

- *Problem-solving statements:* Scientific research papers can be seen as biased reports of a problem-solving activity: they contain many statements about problem-solving activities: own as well as other researchers' (cf. section 3.2.4). Some of these problem-solving activities are portrayed as successful, others as flawed.

Our model of scientific argumentation is operationalized in section 3.3, where we introduce our practical annotation scheme and the task of *Argumentative Zoning*, i.e. the task of applying the scheme to text. Section 3.4 makes the connection back to RDPs and shows how Argumentative Zoning serves the construction of RDPs.

The task introduced in this chapter, Argumentative Zoning, is new, but fits in with the recent surge of interest in document profiling, argumentation and discourse analysis. We will contrast Argumentative Zoning with related work in section 3.5.

3.1. Fixed Section Structure

RDP slots are in many cases identical with the common section headings in scientific articles. The task of filling the slots would be simplified a great deal if we knew from which section in the paper to extract the corresponding material.

The single most prominent property which is the same across many scientific articles is their common external global structure in *rhetorical sections* (or *rhetorical divisions*) and corresponding section headers (van Dijk, 1980). This highly structured building plan for research articles is particularly well-established in the life and experimental sciences, e.g. experimental physics, biology and psychology. The most famous structure is four-pronged and contains the sections *Introduction, Method, Results, Discussion*. In some disciplines, there is a fifth typical section, namely *Conclusions*. Rhetorical sections often contain other rhetorical sections, e.g., a *Method* section in a psychology article is often divided into *Subjects, Materials* and *Procedure*. Rigid section structures enhance efficiency of understanding and information searching: researchers in psycholinguistics, for example, know with great accuracy where to find the number of experimental subjects in any given article.

It has been argued that this structure has evolved and become petrified because texts which serve a common purpose among a community of users eventually take on a predictable structure of presentation (Mullins et al., 1988; Hyland, 1998).

Knowing how to write in this style is important for the career of scientists, but they are rarely trained in it during their undergraduate degrees. Part of the training of young researchers consists in experienced researchers showing them "how to write papers such that they get accepted". Rules on how to fit material into sections do exist (e.g., "*report only numerical results in the* RESULTS *section; if there's interpretation involved, put it into the* DISCUSSION *section*", "*description of machinery belongs into the methodology except if...* "). Prescriptive style manuals and writer aids abound (Mathes and Stevenson, 1976; Blicq, 1983; Alley, 1996; Conway, 1987; Day, 1995;

Farr, 1985; Houp and Pearsall, 1988; Michaelson, 1980; Mitchell, 1968; van Emden and Easteal, 1996; Lannon, 1993). Writing style manuals urge writers to explicitly mark explicit structure, e.g.:

- by clear physical format/layout: orthographically recognizable indications of text structure;
- by mapping of conceptual paragraphs to physical paragraphs;
- by use of informative sub-headings as very short summaries;
- by adherence to conventionalised text structure;
- by explicit signalling of text macrostructure (*"in section 2, we will ...*");
- by clear discourse/rhetorical relations;
- by clear and logical elaboration of the subject matter (topicality and nuclearity).

There have been more or less formal attempts by discourse analysts to model this section structure. Van Dijk (1980) presented conventionalized schematic forms for several text types (apart from experimental research reports, also for narratives, arguments, newspaper articles).

Figure 3.2 shows Kircz' (1991) taxonomy of argumentative entities (taken from Kircz 1991, p. 368), which is more fine-grained than van Dijk's, and specifically designed for physics articles. It also includes dependencies between these entities in the form of see-also links and in the form of logical implications (i.e., there cannot be any experimental constraints if there is no experimental setup), which we have not reproduced here. This structure, though it covers the whole article, is similar to Liddy's structured abstract and other abstract templates. Kando (1997) presents a similar structure which she uses to make queries in a DR environment more distinctive, cf. figure 3.3, taken from (Kando, 1997, p. 70).

Models such as Kando's and Kircz' describe papers from the experimental sciences well. However, our corpus covers an interdisciplinary science. In cognitive science and computational linguistics, where the focus is the investigation and simulation of intelligent action and language processing, a wide range of scientific areas is covered: experimental sciences (psychology, neuroscience), engineering (computer

1. Definition of the research subject in broad terms
    - (a) Redefinition of the problem in the actual research context
2. Experimental setup
    - (a) Experimental constraints
    - (b) Experimental assumptions
    - (c) Experimental ambiguities
    - (d) Relation of experimental setup with other experiments
3. Data collection
    - (a) Data handling methods
    - (b) Data handling criteria
    - (c) Error analysis
4. Presentation of raw experimental data
    - (a) Presentation of smoothed experimental data
    - (b) Pointers to pictorial or tabular presentation
    - (c) Comparison of own data with other results
5. Theoretical model
    - (a) Theoretical constraints
    - (b) Theoretical assumptions
    - (c) Theoretical ambiguities
    - (d) Relation of theoretical elaboration with other works
6. Theoretical/mathematical elaboration
7. Presentation of theoretical results/predictions
    - (a) Comparison with other theoretical results
    - (b) Pointers to pictorial or tabular presentation
8. Comparison of experimental results with own theoretical results
    - (a) Comparison of experimental results with other theoretical results
    - (b) Pointers to pictorial or tabular presentation
9. Conclusions
    - (a) Experimental conclusions
    - (b) Theoretical conclusions
10. Reference to own previous published work
    - (a) Reference to own work in progress
11. Reference to other people's published work
    - (a) Reference to other people's work in progress

Figure 3.2: Kircz' (1991) Argumentative Taxonomy

- A. PROBLEMS
  - A.1 BACKGROUND
    - A.1.1 stating background WITHOUT REFERENCES
    - A.1.2 REVIEW or relevant previous research
  - A.2 RATIONALE
    - A.2.1 GAP of knowledge
    - A.2.2 IMPORTANCE
    - A.2.3 INDUCEMENTS to start the study
    - A.2.4 INTERESTS OF THE AUTHOR(S)
  - A.3 RESEARCH TOPIC
    - A.3.1 RESEARCH QUESTIONS
      - A.3.1.1 HYPOTHESIS
      - A.3.1.2 PURPOSES
    - A.3.2 SCOPE of the study
      - A.3.2.1 OUTLINE of methods
      - A.3.2.2 OUTLINE of discussion
      - A.3.2.3 principle RESULT or conclusion
      - A.3.3.4 ORGANIZATION of the paper
  - A.4 TERM DEFINITION
- B. VALIDITY of the evidence or METHODS
  - B.1. FRAMEWORK of the study
    - B.1.1. RESEARCH DESIGN
    - B.1.2. ENVIRONMENT
    - B.1.3. MODELS/ASSUMPTIONS used in the study
    - B.1.4. REASONS for selecting the framework
  - B.2. SUBJECTS
    - B.2.1. ATTRIBUTES of the subjects
    - B.2.2. SELECTION CRITERIA of the subjects
    - B.2.3. NUMBERS of the subjects
    - B.2.4. REASONS for selecting the subjects
    - B.2.5. ETHICAL CONTROLS for the subjects
  - B.3. OPERATIONS/inventions
    - B.3.1. PROCEDURES of the operation
    - B.3.2. TOOLS used in the operation
    - B.3.3. MATERIALS used in the operation
    - B.3.4. CONDITIONS of the operation
    - B.3.5. REASONS for selecting the operation
  - B.4. DATA COLLECTION
    - B.4.1 PROCEDURES and ITEMS of the data collection
    - B.4.2. TOOLS used in the data collection
    - B.4.3. MATERIALS used in the data collection
    - B.4.4. CONDITIONS of the data collection
    - B.4.5. MEASUREMENT CRITERIA
    - B.4.6 REASONS for selecting the data collection
  - B.5. DATA ANALYSIS
    - B.5.1 PROCEDURES and TECHNIQUES of analysis
    - B.5.2. TOOLS and S/W used in the data analysis
    - B.5.3. REASONS for selecting the analysis
  - B.6. LOGICAL EXPANSION
- C. EXAMINATION of the EVIDENCE
  - C.1. PRESENTATION OF EVIDENCE
  - C.2. ORIGINAL EVIDENCE, mentioned again
  - C.3. ORIGINAL EVIDENCE + opinion
  - C.4. ORIGINAL EVIDENCE + SECONDARY EVIDENCE
  - C.5. ORIGINAL EVIDENCE + SECONDARY EVIDENCE + OPINION
  - C.6. SECONDARY EVIDENCE
  - C.7. SECONDARY EVIDENCE + OPINION
  - C.8. OPINION
- E. ANSWERS
  - E.1. SUMMARY of the study
  - E.2. CONCLUSIONS
  - E.3. FUTURE RESEARCH
  - E.4. APPLICATIONS
  - E.5. SIGNIFICANCE

Figure 3.3: Kando's (1997) Categories

science, language engineering, artificial intelligence), humanities (philosophy), sociology (sociology of science), applied sciences (discourse analysis, English for a Specific Purpose), medicine and theoretical sciences (linguistics and mathematics). The scientific traditions of the authors represented in our corpus vary according to many dimensions:

- *Structure:* In contrast to experimental scientists, humanists comply much less to the classic model for scientific writing (Tibbo, 1992). Tibbo states that the contents of humanistic writing frequently appear as seemingly unstructured text lacking standardized section headings. Historical discourse, for example, consists mainly of interpretative arguments and narrative supporting those arguments.

- *Research style:* In young disciplines new methods evolve fast, as researchers use and combine relatively new techniques with old and new tasks. Additionally, new disciplines often have not agreed on what a good evaluation strategy is. An example for this is the current state of the field of automatic summarization.

- *Cultural differences:* Different language traditions prefer different argumentative structure, as has been shown in the case of English–German (Clyne, 1987) and Polish–English (Duszak, 1994). The main difference seems to be that in the German-Polish tradition the results are kept "hidden" as long as possible, in order to retain the readers' curiosity, whereas the English texts preview the structure of the entire article and give results away early.

- *Conference and Presentation style:* The presentation of a paper can be influenced by how conferences are organized. In philosophy, speakers read their talks from paper, whereas in linguistics free talks prevail, supported by handouts. In computational linguistics, computer science and psychology, where talks are also free, there are printed proceedings and no handouts. In neuroscience, however, talks are often accompanied by a slide show.

- *Peer reviewing:* Researchers in interdisciplinary fields often have to review papers with material coming from a discipline adjacent to their own. They typically do not feel that they should criticize the presentation of that material. As a result, there is a general leniency towards writing style; papers with diverging structure *are* accepted at conferences and in journals.

As predicted, the structure of the papers in our corpus is indeed heterogeneous. Even though most of our articles have introduction and conclusions sections (sometimes occurring under headers with different names), the presentation of the problem and the methodology/solution are idiosyncratic to the domain and personal writing style. In some cases, prototypical headers are used, in others, headers contain subject-matter terms.

Figure 3.4 compares the most frequent headlines in our corpus (left hand side) with those in a comparison corpus of cardiology papers. 74% of all 823 headers in our data are not prototypical. 32% of all papers contain no explicitly marked *Conclusion* section. In the entire CL corpus, there were only two sections titled *Method* or *Methods*.

Computational Linguistics (80 papers)			Cardiology (103 papers)		
Headline	Frequency		Headline	Frequency	
Introduction	63	79%	Introduction	103	100%
Conclusion	34	43%	Results	97	94%
Discussion	13	16%	Discussion	97	94%
Conclusions	13	16%	Methods	95	92%
Acknowledgments	12	15%	Tables	81	79%
Results	8	10%	Statistics	41	40%
Experimental Results	8	10%	Patients	30	29%
Evaluation	7	9%	Limitations	29	28%
Background	7	9%	Conclusions	26	25%
Implementation	6	8%	Statistical Analysis	23	22%
Example	6	8%	Conclusion	18	17%
Acknowledgements	6	8%	Patient Characteristics	9	9%

Figure 3.4: Frequencies of Headlines in CL and Cardiology Corpus

In contrast to the computational linguistics corpus, where the external structure of the paper if obviously a matter of personal style, the section structure in the medical corpus is very homogeneous: each headline out of the typical *Introduction, Method, Result, Discussion* structure is present in almost each paper. The least frequent component, *Methods*, is still present in 92% of all papers. Some papers (25%) contain a *Conclusion* section as a fifth section structure. The only headings that were not prototypical occurred at a deeper level of embedding (e.g. names of specific medical procedures or methodologies such as "*Measurement of lipid hydroperoxides*").

Of course, rhetorical sections in our data might still be present logically even if they are not explicitly marked. In the absence of an *Introduction* section, the same function is sometimes fulfilled by sections titled *Motivation* or *Background*, or by the first paragraphs of the first section. However, in this case it is much harder to find the corresponding types of information.

Overall, if section structure is not the dominant structure in our data, we will have to consider other possible commonalities between the papers. The variation in our data forces us to steer clear of distinctions that are too domain specific. We will have to go "deeper" into the structure of the papers—we believe that more interesting theoretical questions will emerge this way. However, due to the robustness requirements of our approach, we cannot go indefinitely deep: the commonalities we are looking for must still be traceable on the surface.

3.2. A Model of Prototypical Scientific Argumentation

3.2.1. Argumentative Moves: Swales (1990)

We have so far presented scientific articles as purpose-free, objective descriptions of research. The rigid section structure reinforces the impression that the research presented was performed following a strictly logical procedure. However, the process by which a scientific paper is created is very complex—there are many levels of actions that interact, presentational as well as scientific (Latour and Woolgar, 1986). The presentation of research in scientific papers does not normally follow the chronological course of the research. Ziman (1969) states that the authors do not inform of false starts, mistakes, unnecessary complications, difficulties and hesitations. On the contrary, the procedure is shown as simple, precise, profitable and the conclusions derived as inevitable. If we accept a definition of argument as "any proof, demonstration, or reason that is useful for persuading the audience of the validity of a statement" (Myers, 1992), then arguing is an important part of presenting science, even in disciplines where overt argumentation is not part of the presentational tradition.

Swales (1990) assumes that the main communicative goal authors of scientific papers is to convince readers of the validity and importance of their work, as this is the only way to have the paper reviewed positively, and published as a result. Authors need to show that the presented research is justified (i.e., that it addresses an interesting problem), that it is a contribution to science, that the solution presented is a good solution, and that the evaluation is sound.

His CARS model ("Creating a Research Space") describes the structure of introductions to scientific articles according to prototypical rhetorical building plans. The unit of analysis is the argumentative *move* ("a semantic unit related to the writer's purpose"), typically one clause or sentence long. There is a finite number of such moves, and they are subdivided into "steps". The model, a successor of his earlier model (Swales, 1981), is schematically depicted in figure 3.5. It is based on empirical studies on two data collections: firstly, a collection of several hundred research articles in the physical sciences and secondly, a mixed collection of research articles from several science and engineering fields.

One such rhetorical move is to motivate the need for the research presented (Move 2), which can be done in different ways, e.g. by pointing out a weakness of a previous approach (Move 2A/B) or by explicitly stating the research question (Move 2C). Note that context plays an important role for the classification of a sentence in Swales' model: the example sentence for Move 2C (which characterizes the question actually addressed in the article) would constitute a different move if it had appeared towards the end of the article, e.g. under the heading *Future Work*.

Swales' model has been used extensively by discourse analysts and researchers in the field of English for Specific Purposes, and for tasks as varied as teaching English as a foreign language, human translation and citation analysis (Myers, 1992; Thompson and Yiyun, 1991; Duszak, 1994). Salager-Meyer (1990, 1991, 1992) establishes similar moves for medical abstracts. Busch-Lauer (1995) did not find these moves in all abstracts of her German medical corpus; she concludes that presentation and arrangement of moves are related to the author's intentions and summarizing skills.

An inspection of introduction sections in our corpus showed that Swales' definition of argumentative moves seem to generalize well to the domain of computational linguistics and cognitive science. (Crookes (1986), however, reports that is not the case for the social science literature.) As a result of the shortness of our texts, however, the optional move 3.3 (INDICATE ARTICLE STRUCTURE) was rare. The right hand side of figure 3.5 shows real examples coming from our corpus.

Even though Swales' model is non-computational, i.e. not aimed at automatic recognition of the moves, one important assumption in Swales' work is that the argumentative status of a certain move is visible on the surface by linguistic cues. This is important for our task.

We will use a description based on argumentative moves to describe structural similarities between papers in our corpus, but we feel that we cannot use Swales' model

MOVE 1: ESTABLISHING A TERRITORY

1.1	CLAIMING CENTRALITY	• *Recently, there has been a lot of interest in Earley deduction* [...] (S-0, 9502004)
1.2	MAKING TOPIC GENERALIZATIONS (BACKGROUND KNOWLEDGE) OR	• *The traditional approach has been to plot isoglosses, delineating regions where the same word is used for the same concept.* (S-3, 9503002)
	(DESCRIPTION OF PHENOMENA)	• *In the Japanese language, the causative and the change of voice are realized by agglutinations of those auxiliary verbs at the tail of current verbs.* (S-56, 9411021)
1.3	REVIEWING PREVIOUS RESEARCH	• *Brown et al. (1992) suggest a class-based n-gram model in which words with similar cooccurrence distributions are clustered in word classes.* (S-12, 9405001)

MOVE 2: ESTABLISHING A NICHE

2A	COUNTER-CLAIMING	• *I argue that Hidden Markov Models are unsuited to the task* [...] (S-9, 9410022)
or 2B	INDICATING A GAP	• [...] *and to my knowledge, no previous work has proposed any principles for when to include optional information* [...] (S-9, 9503018)
or 2C	QUESTION-RAISING	• *How do children combine the information they perceive from different sources?* (S-15, 9412005)
or 2D	CONTINUING A TRADITION	• *Within a current project on adapting bilingual dictionaries* [...] *the need arose for a POS-disambiguator to facilitate a context sensitive dictionary look-up system.* (S-4, 9502038)

MOVE 3: OCCUPYING A NICHE

3.1A	OUTLINING PURPOSE	• *The aim of this paper is to examine the role that training plays in the tagging process* [...] (S-32, 9410012)
or 3.1B	ANNOUNCING PRESENT RESEARCH	• *In this paper, we discuss the interaction of temporal anaphora and quantification over eventualities.* (S-2, 9502023)
3.2	ANNOUNCING PRINCIPLE FINDINGS	• *In our corpus study, we found that three types of utterances (prompts, repetitions and summaries) were consistently used to signal control shifts* [...] (S-139, 9504006)
3.3	INDICATING ARTICLE STRUCTURE	• *This paper is organized as follows: We first review a general algorithm for least-errors recognition* [...] (S-27, 9502024)

Figure 3.5: Swales' (1990) CARS Model; Examples from our Corpus

without adjustment. Firstly, whereas Swales' scheme covers only the introduction we need a model that describes the whole article; some moves might have to be added. Also, many of Swales' definitions are vague. For example, the difference between the two moves 2D (Continuing a tradition) and 2C (Indicating a gap) is that for move 2D "there is a weaker challenge to the previous research" (Swales, 1990, p. 156). Our feeling is that the scheme would need to be operationalized before it could be applied by groups of annotators.

Swales's (1990) model is more flexible than models of fixed section structure like van Dijk's. However, it still assumes an argumentative structure which is rather close to the textual form, with a fixed order of moves. We empirically found that the order he suggests is typically indeed the most frequent, but we also found many cases in our heterogeneous corpus where the argumentative moves were ordered in unexpected ways. For example, six of our texts started with a specific goal statement, and 14 introductions do not contain any explicit goal statement at all. Duszak (1994) reports similar problems with Swales' assumption of a fixed move order.

	Swales' move name	Our move name
1.1	Claiming Centrality	Describe: General Goal Show: Own Goal/Problem is Important/Interesting Show: Solution to Own Problem is Desirable Show: Own Goal/Problem is Hard
1.2	Making Topic Generalizations	Describe: General Problem Describe: General Conclusion/Claim
1.3	Reviewing Previous Research	Describe: Other Conclusion/Claim
3.1A	Outlining Purpose	Describe: Own Goal/Problem
3.1B	Announcing Present Research	Describe: Own Goal/Problem
3.2	Announcing Principle Findings	Describe: Own Conclusion/Claim
3.3	Indicating Article Structure	Describe: Article Structure Preview: Section Contents Summarize: Section Contents

Figure 3.6: Move Names in Swales' and in our Model

We borrow Swales' moves given in figure 3.6 and expand them to the moves in figure 3.7. These 12 moves are a useful description of a large part of the material occurring in the introduction sections and some other material too.

The moves for textual presentation (Swales' "Indicate Article Struc-

1. DESCRIBE: GENERAL GOAL
   Abstract generation is, like Machine Translation, one of the ultimate goal [sic] of Natural Language Processing. (S-0, 9411023)

2. SHOW: OWN GOAL/PROBLEM IS IMPORTANT/INTERESTING
   Both principle-based parsing and probabilistic methods for the analysis of natural language have become popular in the last decade. (S-0, 9408004)

3. SHOW: SOLUTION TO OWN PROBLEM IS DESIRABLE
   The knowledge of such dependencies is useful in various tasks in natural language processing, especially in analysis of sentences involving multiple prepositional phrases, such as: [...] (S-10, 9605013)

4. SHOW: OWN GOAL/PROBLEM IS HARD
   Correctly determining number is a difficult problem when translating from Japanese to English. (S-0, 9511001)

5. DESCRIBE: GENERAL PROBLEM
   The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities. (S-4, 9408011)

6. DESCRIBE: GENERAL CONCLUSION/CLAIM
   It has often been stated that discourse is an inherently collaborative process [...] (S-171, 9504007)

7. DESCRIBE: OTHER CONCLUSION/CLAIM
   Nonetheless there is psychological evidence that language has an unplanned, spontaneous aspect as well (Ochs 1979). (S-9, 9410032)

8. DESCRIBE: OWN GOAL/PROBLEM
   The aim of this paper is to examine the role that training plays in the tagging process [...] (S-32, 9410012)

9. DESCRIBE: OWN CONCLUSION/CLAIM
   [...] *we found that three types of utterances (prompts, repetitions and summaries) were consistently used to signal control shifts.* (S-139, 9504006)

10. DESCRIBE: ARTICLE STRUCTURE
    This paper is organized as follows: We first review a general algorithm for least-errors recognition [...] (S-27, 9502024)

11. PREVIEW: SECTION CONTENTS
    In this section, we are going to motivate the reasons which lead us to choose grammatical words as discriminant. (S-21, 9502039)

12. SUMMARIZE: SECTION CONTENTS
    The previous section provided illustrative examples, demonstrating the performance of the algorithm on some interesting cases. (S-125, 9511006)

Figure 3.7: Moves Based on Swales' CARS Model

ture, our moves 10, 11, 12) are important, even though they have no direct connection to the argumentation. When reporting their research, the authors have to solve the problem of how to linearize their statements in such a way that a reader will be able to understand the main points. In disciplines where fixed section structure is not typical, authors often inform the reader explicitly of which content to expect in each section.

Swales' moves 2A through 2D, which have to do with how other work is introduced and cited, are not included in these 12 moves. In order to operationalize these moves, we should take a closer look at how authors express a stance towards other work, and how this information could be encoded.

3.2.2. Citations and Author Stance

This section will look at results from Content Citation Analysis, one strand of research within library science and the sociology of science, in order to define the concept of authors' stance towards other work. Researchers in content citation analysis have determined and classified semantic relationships between citing and cited works. As we will see it is a highly political matter whether a researcher cites another or not, and what they write about the other's work.

Whereas in industry, the patent system registers intellectual property and thus encourages researchers to produce and contribute new ideas and results, the reward system in science is based on publication and citation (Luukkonen, 1992). To publish an idea means staking a claim of intellectual ownership for that idea (Myers, 1992). The assumption is that other researchers who use the idea must acknowledge them as the authors' intellectual ownership; this is done by formal citation.

Research institutions are rewarded by exercises like the British RAE (Research Assessment Exercise), which measures intellectual output by number of publications in quality journals; individual researchers are affected because publishing is one of the main criteria used in promotion and tenure decisions—this is captured in the well-known motto of "publish or perish".

Other bibliometric measures assesses the quality of a researcher's output, also in a purely quantitative manner, by counting how many papers *cite* a given paper. Content citation analysis is critical of the application of pure citation counting as a measurement of quality and impact of scientific work. Bonzi (1982), for example, points out that *negational* citations, while pointing to the fact that a given work has been *noticed* in a field, does not mean that that work is *received well*, and Ziman (1968), following a slightly different argumentation, states that many citations are done out of "politeness" (towards powerful rival approaches), "policy" (by name-dropping and argument by authority) or "piety" (towards one's friends, collaborators and superiors). Researchers also often follow the custom of citing some particular early, basic paper, which gives the foundation of their current subject ("paying homage to pioneers").

Researchers in content citation analysis believe that the classification of motivations is a central element in understanding the relevance of the paper in the field. Many classification schemes for properties of citations have been invented to this end (Weinstock, 1971; Swales, 1990; Oppenheim and Renn, 1978; Frost, 1979; Chubin and Moitra, 1975). Based on such annotation schemes and hand-analyzed data, different influences on citation behaviour can be determined. As one of the earliest such studies, Moravcsik and Murugesan (1975) divide citations in running text into four dimensions: conceptual or operational use (i.e., use of theory vs. use of technical method); evolutionary or juxtapositional (i.e., own work is based on the cited work vs. own work is an alternative to it); organic or perfunctory (i.e., work is crucially needed for understanding of citing article or just a general acknowledgement); and finally confirmative vs. negational (i.e., is the correctness of the findings disputed?). They found, for example, that 40% of the citations were perfunctory, which casts further doubt on the mere citation-counting approach.

As another example of a finer-grained scheme, we reproduce Spiegel-Rüsing's (1977) scheme (taken from p. 105) in figure 3.8. Spiegel-Rüsing's results are that of 2309 citations examined, 80% substantiated statements (category 8), 6% discussed history or state of the art of the research area (category 1) and 5% cited comparative data (category 5).

Annotation schemes such as the ones discussed above are subjective, the suggested classifications are difficult to operationalize and annotation is usually not confirmed by reliability studies. Swales (1986), for example, calls researchers in Content Citation Analysis "zealously interpretative" (p. 44).

We are interested in the role that authors' stance plays in the overall argumentation of the paper, as this stance can provide the information of *relatedness* (e.g. rivalry and ancestry) between papers. It is natural to expect that authors should express a stance towards work they introduce: real estate in the paper is sparse, so authors will tend to try and put it to good use for strengthening the argument. If the other work is used as part of her solution, we expect the author to express a positive stance; if she compares her own work with it or if she has identified a problem with it, we expect a

1. Cited source is mentioned in the introduction or discussion as part of the history and state of the art of the research question under investigation.
2. Cited source is the specific point of departure for the research question investigated.
3. Cited source contains the concepts, definitions, interpretations used (and pertaining to the discipline of the citing article).
4. Cited source contains the data (pertaining to the discipline of the citing article) which are used sporadically in the article.
5. Cited source contains the data (pertaining to the discipline of the citing particle) which are used for comparative purposes, in tables and statistics.
6. Cited source contains data and material (from other disciplines than citing article) which is used sporadically in the citing text, in tables or statistics.
7. Cited source contains the method used.
8. Cited source substantiated a statement or assumption, or points to further information.
9. Cited source is positively evaluated.
10. Cited source is negatively evaluated.
11. Results of citing article prove, verify, substantiate the data or interpretation of cited source.
12. Results of citing article disprove, put into question the data as interpretation of cited source.
13. Results of citing article furnish a new interpretation/explanation to the data of the cited source.

Figure 3.8: Spiegel-Rüsing's (1977) Categories for Citation Motivations

contrastive stance. We also expect other work which is more relevant to receive more space in the paper. While we do not deny that there are many other motivations for citing apart (e.g. citations for general reference, background material, homage to pioneers (Ziman, 1968)), we still assume here that citations which are afforded some space in the paper will be used to support the overall scientific argumentation.

In this context it is interesting to consider negational citations. Both Moravcsik and Murugesan and Spiegel-Rüsing found that negational citations are rare.

MacRoberts and MacRoberts (1984) argue that the reason why pure negational citations are rare is that they are potentially politically dangerous, and that they must therefore be made more acceptable. They claim that authors dissemble in order to diffuse the impact of negative references, hiding a negative point behind insincere praise, or diffusing the thrust of criticism with perfunctory remarks ("damning them with faint praise"). Brooks's (1986) interviews of scholars and classification of 437 references confirms this hypothesis. In our data we found ample evidence of this effect, cf. the following examples:

> *This account makes reasonably good empirical predictions, though it does fail for the following examples:* (S-75, 9503014)

> *Hidden Markov Models (HMMs) (Huang et al. 1990) offer a powerful statistical approach to this problem, though it is unclear how they could be used to recognise the units of interest to phonologists.* (S-24, 9410022

> *Even though these approaches often accomplish considerable improvements with respect to efficiency or termination behavior, it remains unclear how these optimizations relate to each other and what comprises the logic behind these specialized forms of filtering.* (S-21, 9604019)

When there was apparent simultaneous positive and negative evaluation of a citation in one paper, the positive negation always precedes the negative one, suggesting that the real intention was to criticize.

The moves given in figure 3.9 are based on author stance. The first of these moves describes a weakness of previous research (cf. Spiegel-Rüsing's 10, 12, possibly 13; Moravcsik/Murugesan's "negational/juxtapositional"). The next three describe comparisons between own and other work (cf. Spiegel-Rüsing's category 5; no Moravcsik/Murugesan category). The move expressing the fact that other work is advantageous is best expressed with Spiegel-Rüsing's category 9, and Moravcsik/Murugesan's "confirmative". The final move, a statement of intellectual ancestry, is expressed in many of Spiegel-Rüsing's categories (2, 3, 4, 5, 6, 7, possibly 9), and in Moravcsik/Murugesan's "evolutionary" category.

Note that our main distinction into positive/continuing and negative/contrastive stances can be expected to be intuitive: all annotation schemes enumerated here make this distinction, including Shum's (1998) meta-data scheme. Spiegel-Rüsing's and many other schemes, however, typically make finer distinctions.

13. Show: Other Solution is Flawed
Goal-freezing [...] is equally unappealing: goal-freezing is computationally expensive, it demands the procedural annotation of an otherwise declarative grammar specification, and it presupposes that a grammar writer possesses substantial computational processing expertise.
(S-59, 9502005)

14. Show: Own Solution is Different from Other Solution
The use of the chart to store known results and failures allows the user to develop hybrid parsing techniques, rather than relying on the default depth-first top-down strategy given by analysing with respect to the top-most category. (S-146, 9408006)

15. Show: Own Goal/Problem is Different from Other Goal/Problem
Unlike most research in pragmatics that focuses on certain types of presuppositions or implicatures, we provide a global framework in which one can express all these types of pragmatic inferences. (S-124, 9504017)

16. Show: Own Claim is Different from Other Claim
Despite the hypothesis that the free word order of German leads to poor performance of low order HMM taggers when compared with a language like English, we have shown that the overall results for German are very much along the lines of comparable implementations for English, if not better. (S-117, 9502038)

17. Show: Other Solution is Advantageous
CUG (Categorial Unification Grammar; Uszkoreit (1986)) is advantageous, compared to other phrase structure grammars, for parallel architecture, because we can regard categories as functional types and we can represent grammar rules locally. (S-10, 9411021)

18. State: Other Solution Provides Basis for Own Solution
We present a different method that takes as starting point the back-off scheme of Katz (1987).
(S-24, 9405001)

Figure 3.9: Moves Based on Author Stance

Our move 18 State: Other Solution Provides Basis for Own Solution might well be split into a) theoretical basis b) use of data or c) definition of used methodology—however, what interests us here is the *positive* tenet and the idea of intellectual ancestry more than the exact aspect of agreement with the prior work.

Content citation analysis experiments seem to point to the fact that humans are in principle capable of determining author stance in running text—we will, in section 4.3, employ human judgement for a similar task. However, as already mentioned in section 2.1.2, we are concerned about the potentially high level of subjectivity, a general problem with many studies in the field of content citation analysis.

We try to increase the objectivity of the task by giving exact guidelines and instructing our annotators to only mark citation stance when the authors have explicitly stated it. Also, the most subjective categories are not part of our scheme ("paying homage to pioneers"), which should put us on fairly objective ground. Nevertheless, in order to make sure that these decisions can indeed be made reliably, we also measure reproducibility and stability between several annotators formally.

Other content citation analysis research which is important for us concentrates on relating textual spans to authors' descriptions of other work. For example, in O'Connor's (1982) experiment, *citing statements* (one or more sentences referring to other researchers' work) were manually identified. The main problem encountered in that work is the fact that many instances of citation context are linguistically unmarked. Our data confirms this: articles often contain large segments, particularly in the central parts, which describe research in a fairly neutral way. In order to capture the role of these long neutral segments for the overall argumentation, we needed to define different types of moves. The basis of this definition will be the attribution of intellectual ownership, as motivated in the next section.

3.2.3. Attribution of Intellectual Ownership

We have discussed in the previous section how knowledge claims of *other* authors are acknowledged in the reward system of science. Of course, it is equally essential that the knowledge claims of the current paper itself are registered properly (Myers, 1992), as the intellectual rights to the solution or claim associated with the research are not owned by the authors until they have been accepted by the community via peer review (Zuckerman and Merton, 1973).

Whereas it is arguably in the interest of every researcher to publish as many articles as possible, new research results are a scarce and valuable substance. Research might be presented and possibly perceived as coming naturally in different "sizes"—journal-article-length, conference-length or workshop-length packets of scientific knowledge—but it is clear that this is not how research is done. It is more typically a continuous activity carried out over decades by an individual and her co-workers, such that it is not obvious how much of it should be reported in one paper. Instead, the amount of new research going into a paper is a strategic decision for every researcher.

One strategy for publishing more is to present as many aspects of one piece of research in as many publications as will get accepted, with as few changes as possible. This results in authors breaking research down into "smallest publishable units". This phenomenon is illustrated by clusters of papers with titles which are close variations of one theme—it can be assumed that the scientific innovations presented in these papers will show a high level of overlap. However, there is a tension between the interest of the individual to publish and the interest of the field not to be swamped by near-identical papers. The main quality control mechanism in science is the peer-reviewing process, which guarantees a minimum size of the smallest publishable unit, by making sure that in principle each published paper contains at least *something* new ("original" and "previously unpublished").

A scientific paper contains many ideas and statements which are not the authors' own ideas and beliefs, but which are needed to guide the reader towards accepting their own ideas and beliefs. Other ideas, methods or results are associated with other researchers, namely those which own the intellectual rights for them. Of course, the author does not claim intellectual ownership of those statements; instead, she should recognize the other authors' knowledge claims for them.

We think of documents as divided into segments of different intellectual ownership, where each segment plays a certain role in the overall scientific argumentation:

- General statements about the field's problems and methodologies; statements are portrayed as generally accepted in the field (BACKGROUND).
- More specific descriptions of other researchers' work, e.g. rival approaches (OTHER).
- As the real interest of an author is to stake a new knowledge claim, she needs to make clear what exactly her new contribution is (OWN).

The logical tri-section into types of intellectual ownership is related to the semantics of all moves introduced so far, and it also defines the three new moves shown in figure 3.10. These moves constitute larger textual units than the moves introduced so far which are typically associated with single sentences. For a coverage of the entire paper, the longer moves are indispensable.

We believe that clear attribution of intellectual ownership is one aspect of overall writing quality of a paper: readers often have difficulty recognizing attribution of intellectual ownership in unclearly written papers. Section 4.3.2 will address this question by first experimentally testing if humans can in principle attribute ownership reli-

19. DESCRIBE: GENERAL SOLUTION
The traditional approach has been to plot isoglosses, delineating regions where the same word is used for the same concept. (S-3, 9503002)

20. DESCRIBE: OTHER SOLUTION
Instead, Katz's back-off scheme redistributes the free probability mass non-uniformly in proportion to the frequency of <EQN/>, by setting <EQN/> (S-56, 9405001)

21. DESCRIBE: OWN SOLUTION
The basic idea [...] *is to move from dealing with a single model to dealing with a collection of models linked by an accessibility relation.* (S-196, 9503005)

Figure 3.10: Moves Based on Intellectual Ownership

ably; it will then argue that those texts where they disagree much more than expected must be less clearly written.

How do humans understand who a certain statement in a scientific article is attributed to?

- *Top-down information:* Readers anticipate certain argumentative moves; when interpreting the text they infer the probable communicative intentions of the author.

- *World-Knowledge:* Experts use world knowledge to infer intellectual ownership. They know which statements in a text are established fact and which are intellectually owned by other researchers, and assume that everything else must be the authors' conjecture or knowledge claim.

- *Agent markers:* Agents (other researchers or the authors) typically appear in ritualized roles—they are often portrayed as rival researchers ("*Chomsky argues that*", *"workers in AI"*), as contributors of supportive research (*"several discourse linguists"*) and as representatives of the general opinion in the field (*"It is a well-known fact that*").

- *Segmentation and boundaries:* However, not every sentence contains agent markers. On the contrary, even in clear and well-written papers, most sentences are unmarked propositions which state facts about the object world. Their status can be inferred from surrounding attribution boundaries. Readers assume that unmarked statements are attributed to the previously explicitly mentioned agents, until a new explicit attribution redefines the status of the next segment, or until an obvious conflict catches the reader's eye.

- *Linguistic Cues:* Readers use linguistic cues like tense and voice and non-linguistic cues like location to check that they are still in the type of segment they expect to be in.

Of course, there are papers which show a less pronounced tri-section of intellectual ownership. Work which is "close" to the authors—particularly previous own or co-authored work, but also work of friends or colleagues of the same institution—is usually treated in the text similarly to how the own work is treated, e.g. it is evaluated more positively than other work cited. In some cases, the authors continue a tradition, i.e., add a small amount of research to own previous work described elsewhere. Often the largest part of such papers describes the *previous* own work in a tenet that might make the reader mistake it for the actual *new* contribution of the given paper, if she does not know the prior paper ("smallest publishable unit"). Attribution might then be ambiguous for large portions of the text, an unclarity which might actually even be in the interest of the author.

However, we consider close work as distinct from the current work: As motivated in chapter 1, our task is to determine each paper's contribution with respect to other papers in order to support searchers in a document retrieval environment. Their choice is bound to be particularly difficult if the papers are by the same authors in a similar time frame. The idea is that it is the knowledge claim of each paper which should provide the selection criterion.

In review or position papers, all intellectual work is at a meta-level (reasoning about research work)—no own "technical" object-level work is performed. Thus, the distinction of own and other work does not really apply. A similar case of meta-level research are evaluation papers, i.e. papers in which one approach (typically, one's own) is formally evaluated on a given task, or several approaches are formally compared (one's own approach typically being one of these).

For now, there is one last piece missing in the argumentational mosaic before we can move on to the overall model. This piece has to do with statements describing research as a sequence of (successful or unsuccessful) problem-solving activities.

3.2.4. Statements about Problem-Solving Processes

There are different descriptions of the internal logic of the scientific research process; some of these are oriented in the hypothesis testing framework (Suppe, 1998). An alternative is to regard scientific papers as reports of a problem-solving activity (Hoey, 1979; Solov'ev, 1981; Jordan, 1984; Zappen, 1983; Trawinski, 1989).

In theoretical sciences, the problem is to find an adequate and explanatory *model* that accounts for the evidence obtained from observing the real world, whereas in experimental sciences, the problem is to find *evidence* for some theory about how the world works. In engineering, *artefacts* are designed which fulfill a certain predefined function. Accordingly, what counts as an acceptable solution is discipline specific.

We describe now a simple view of academic research acts. In this model, one atomic research act is associated with exactly one paper. A situation Sit_0 is perceived as unsatisfactory because problem $Prob_0$ is associated with it. The first step in the research process is the formulation of a research goal $Goal_0$. Problem $Prob_0$ is solved (or at least "*addressed*") by applying a solution $Solu_0$ (a new methodology, or an experiment), which leads to a situation Sit_1. Whereas the problem $Prob_0$ might or might not be already known in the field, the solution $Solu_0$ is always assumed to be new (at the least, the application of the solution in the given problem situation is new). Evaluation measures how well the goal was achieved, i.e., how much the overall situation has improved, by implicitly or explicitly comparing situations Sit_0 and Sit_1. There might be remaining problems $Prob_1$ associated with Sit_1 which are not addressed in the current paper; they are the *limitations of the approach.* They are typically portrayed as less severe than the problems which motivated the research ($Prob_0$).

For the argumentation in the paper, Situation Sit_0 needs to be portrayed as undesirable; to improve Sit_0 is the central motivation of the paper. Alternatively, one could show that Sit_1 is desirable; at the very least, situation Sit_1 should be more desirable than situation Sit_0, even if only because in Sit_1 more knowledge is available.

With respect to knowledge claims, the solution is the single entity which is most proprietary about one problem-solving process; the authors want to be attributed with it. To a lesser degree, the research goal can also be considered as the authors' contribution. In some fields, e.g. in complexity theory, the invention of new problems is itself a research goal which would justify the publication of a paper. Such meta-problems do not fit well with our simple problem-solving model.

Not only can the *own* problem-solving process be described by such atomic research acts. The argumentation in a paper also involves descriptions of other people's problem-solving activities. The background of a problem can be introduced as (possibly successive) problem-solving actions, including general problems in the field, general solutions, research goals and evaluation methodologies. The problem addressed in the paper (*Prob*$_0$) could be a specific weakness of prior solutions which have led to the situation *Sit*$_0$, or it could be a general, long-standing problem in the field.

The own solution can be portrayed as building on some other problem-solving process: some other methodology or idea is taken as the basis for the reported research and applied either with or without changes.

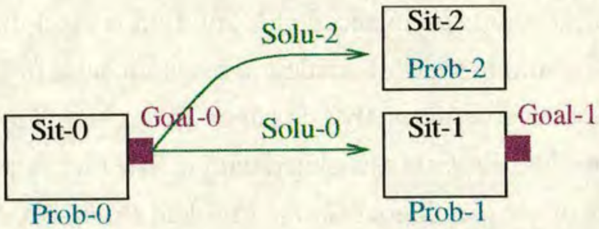


Figure 3.11: Rival Problem-Solving Processes

Figure 3.11 shows a situation where the own paper solution *Solu*$_0$ solves a *known* problem *Prob*$_0$, i.e. a problem to which some other researchers have already presented a solution *Solu*$_2$. The problem solving process presented by the other researchers leads to a different situation *Sit*$_2$. *Sit*$_2$ is similar to *Sit*$_1$, the one favoured by the authors, in that both *Sit*$_1$ and *Sit*$_2$ are not associated with the original problem *Prob*$_0$ anymore, but they differ in some other respect. It is the task of the authors to motivate that the own solution is better than the rival solution. For example, there might be (new) problems associated with *Solu*$_2$, or *Solu*$_2$ might be inferior according to some default criteria—solutions are supposed to be explanatory, elegant, simple, and efficient.

Statements about own and other problem solving processes abound in our data. Figure 3.12 summarizes our moves based on author stance and problem-solving statements. Note that moves describing somebody else's unsuccessful problem solving activity also express contrastive stance and could have been classified as belonging to the moves in figure 3.9.

As the reader has now seen almost all moves we propose and should have an idea of the constructions this thesis is interested in, we will turn to the important aspect of *how* such statements are typically expressed in scientific articles.

22. SHOW: OWN SOLUTION SOLVES OWN PROBLEM
This account also explains similar differences in felicity for other coordinating conjunctions as discussed in Kehler (1994a) [. . .] (S-100, 9405010)

23. SHOW: OWN SOLUTION IS NECESSARY TO ACHIEVE OWN GOAL
We have argued that obligations play an important role in accounting for the interactions in dialog. (S-217, 9407011)

24. SHOW: OWN SOLUTION AVOIDS PROBLEM
This paper presents a treatment of ellipsis which avoids these difficulties, while having essentially the same coverage as Dalrymple et al. (S-9, 9502014)

25. SHOW: OTHER SOLUTION DOES NOT SOLVE PROBLEM
Computational approaches fail to account for the cancellation of pragmatic inferences: once presuppositions or implicatures are generated, they can never be cancelled. (S-20, 9504017)

26. SHOW: OTHER SOLUTION SOLVES PROBLEM
The Direct Inversion Approach (DIA) of Minnen et al. (1995) overcomes these problems by making the reordering process more goal-directed and developing a reformulation technique that allows the successful treatment of rules which exhibit head-recursion. (S-15, 9502005)

27. SHOW: OTHER SOLUTION INTRODUCES NEW PROBLEM
Specifically, if a treatment such as Hinrichs's is used to explain the forward progression of time in example <CREF/>, then it must be explained why sentence <CREF/> is as felicitous as sentence <CREF/>. (S-12, 9405002)

28. SHOW: OWN SOLUTION IS BETTER THAN OTHER SOLUTION
We found that the MDL-based method performs better than the MLE-based method. (S-11, 9605014)

29. SHOW: OWN GOAL/PROBLEM IS HARDER THAN OTHER GOAL/PROBLEM
[. . .] *disambiguating word senses to the level of fine-grainedness found in WordNet is quite a bit more difficult than disambiguation to the level of homographs (Hearst 1991; Cowie et al. 1992).* (S-147, 9511006)

Figure 3.12: Moves Based on Problem-Solving Statements

3.2.5. Scientific Meta-Discourse

In section 3.2.3 we hypothesized that there are superficially recognizable correlations of boundaries of zones of intellectual attribution, e.g. expressions like *"Chomsky claims that"*. We believe that meta-discourse is one of the most universally applicable structure markers in scientific text.

Meta-discourse, commonly defined as *discourse about discourse*, is a name for

Category	Function	Examples
Textual meta-discourse		
Logical connectives	express semantic relation between main clauses	*in addition; but; therefore; thus*
Frame markers	refer to discourse acts or text stages	*to repeat;our aim here; finally*
Endophoric markers	refer to information in other parts of the text	*noted above; see Fig 1; below*
Evidentials	refer to source of information from other texts	*according to X; Y (1990)*
Code glosses	help readers grasp meanings of ideational material	*namely; eg; in other words*
Interpersonal meta-discourse		
Hedges	withhold author's full commitment to statements	*might; perhaps; it is possible*
Emphatics	emphasize force or author's certainty in message	*in fact; definitely; it is clear; obvious*
Attitude markers	express author's attitude to propositional content	*surprisingly; I agree; X claims*
Relational markers	explicitly refer to or build relationship with reader	*frankly; note that; you can see*
Person markers	explicit reference to author(s)	*I; we; my; mine; our*

Figure 3.13: Hyland's (1998) Categories of Meta-Discourse

all those statements which fulfill other functions but to convey pure propositional contents (the "science" in the paper). Meta-discourse is a pragmatic construct by which writers signal their communicative intentions (Hyland, 1998; Swales, 1990). It is ubiquitous in scientific writing: Hyland (1998) found a meta-discourse phrase on average after every 15 words in running text, hedges being the most frequent type of meta-discourse in his texts. His classification of meta-discourse is given in figure 3.13.

Some of Hyland's categories (Attitude markers, Person markers, Evidentials, Endophorics and Frame Markers) seem immediately relevant to the effects discussed in this chapter. Another set of meta-discourse which we are particularly interested in are meta-statements about the own research. Much of that type of scientific meta-discourse is conventionalized, particularly in experimental sciences, and particularly in the methodology or result section; linguistically, there is not much variation (e.g. "*we present original work...*", or "*An ANOVA analysis revealed a marginal interac-*

tion/a main effect of..."). Such formulaic expressions occur less often in the discussion section and the introduction where there is more room for personal style. Swales (1990) lists many such fixed phrases as co-occurring with the moves of his CARS model (p.144;pp.154–158;pp.160–161). Another type of meta-discourse points to the current research process ("*in this paper*", "*here*"), expresses affect ("*unfortunately*") or knowledge states ("*to the best of our knowledge*"; "*it has long been known*").

It is well-known that different disciplines use different meta-discourse. Hyland (1998) argues that meta-discourse variation between scientific communities can be attributed to the fact that meta-discourse has to follow the norms and expectations of particular cultural and professional communities—scientific communities impose linguistic standardization pressures. He found significant differences in meta-discourse use across disciplines (Microbiology, Marketing, Astrophysics and Applied Linguistics), though the articles displayed a remarkable similarity in the *density* of meta-discourse. Marketing and Applied Linguistics papers used far more interpersonal meta-discourse than those in Biology and Astrophysics, which, on the other hand, use far more textual meta-discourse. Due to the particularities of our data we expect meta-discourse in our corpus to be varied.

And even within one discipline, there is a large class of expressions which express similar, prototypical moves, even though the resulting sentences do not look similar on the surface. This is particularly the case for statements referring to aspects of the problem-solving process or to the author's stance towards other work: expressions of contrast to other researchers and for statements of research continuation. Figure 3.14 shows that there are many ways to express the fact that one piece of work is based on some previous other work.

The surface forms of these sentences are very different despite the similar semantics they express: in some sentences the syntactic subject is a method, in others it is the authors, and in others the originators of the based-upon idea. Also, the verbs used are very different. This wide range of linguistic expression presents a real challenge—later parts of this thesis will be concerned with finding a method for recognizing a large subset of such variable meta-discourse (cf. section 5.2.2).

After this brief look at the syntactic variability of the moves, we now return to our model of overall strategy of argumentation.

- *Thus, we base our model on the work of Clark and Wilkes-Gibbs (1986), and Heeman and Hirst (1992) who both modeled (the first psychologically, and the second computationally) how people collaborate on reference to objects for which they have mutual knowledge.* (S-15, 9405013)

- *The starting point for this work was Scha and Polanyi's discourse grammar (Scha and Polanyi 1988; Pruest et al. 1994).* (S-4, 9502018)

- *We use the framework for the allocation and transfer of control of Whittaker and Stenton (1988).* (S-36, 9504007)

- *Following Laur (1993), we consider simple prepositions (like "in") as well as prepositional phrases (like "in front of").* (S-48, 9503007)

- *Our lexicon is based on a finite-state transducer lexicon (Karttunen et al. 1992).* (S-2, 9503004)

- *Instead of feature based syntax trees and first-order logical forms we will adopt a simpler, monostratal representation that is more closely related to those found in dependency grammars (e.g. Hudson (1984)).* (S-116, 9408014)

- *The centering algorithm as defined by Brennan et al. (BNF algorithm), is derived from a set of rules and constraints put forth by Grosz et al. (Grosz et al. 1983; Grosz et al. 1986).* (S-56, 9410006)

- *We employ Suzuki's algorithm to learn case frame patterns as dendroid distributions.* (S-23, 9605013)

- *Our method combines similarity-based estimates with Katz's back-off scheme, which is widely used for language modeling in speech recognition.* (S-151, 9405001)

Figure 3.14: Variability of Statements Expressing Research Continuation

3.2.6. Strategies of Scientific Argumentation

Scientific articles are *biased* reports; the argumentation follows the interest of the author. Indeed, we see the whole paper as one rhetorical act, as Myers (1992) does. The high level communicative goal in a paper, apart from conveying a message, is to persuade the scientific community of the relevance, reliability, quality and importance of the work (Swales, 1990; Kircz, 1998). There are parallels to politeness theory (Brown and C., 1987), where the commodity that is traded is "face"; in the case of scientific writing, the commodity is "credibility".

There are some "high level" moves which are essential for the overall argumentation: One needs to show that the research process is successful, i.e. that the total knowledge available to the community must have increased. The most important ones

Show: Own Research is Valid Contribution to Science

Show: Research is Justified

Show: Authors are Knowledgeable

Show: Other Researchers have Tried to Solve the Problem

Show: Own Solution Process is New

Show: Nobody has Used Same Solution for Same Problem Before

30. Show: Own Goal/Problem is New
    [...] *and to my knowledge, no previous work has proposed any principles for when to include optional information* [...] (S-9, 9503018)

31. Show: Own Solution is Advantageous
    The substitutional treatment of ellipsis presented here [...] *has the computational advantages of* [...] (S-210, 9502014)

Figure 3.15: Moves Based on Higher-Level Intentions

of these moves are given in figure 3.15.

The first six moves in figure 3.15 are not numbered and contain no corpus example. The reason for this is that these moves are not typically made explicit; instead, the reader is left to induce them. The last two high-level moves, however, do occur explicitly, making our set of 31 argumentative moves complete (summarized in figure 3.16).

Relations between the moves are shown in figure 3.17. The tree relation means "Is A Sub-Move Of". An argumentation strategy might be as follows: One might say that the own problem is hard, then introduce the own solution, argue that it solves the problem, argue that this solution is better than somebody else's solution or state the fact that the problem has never been addressed before.

Not all of these moves have to occur in a scientific article for the argumentation to be successful or complete. For example, the problem addressed ($Prob_0$) can be new to the field; this can be stated explicitly (30). Additionally, one can shown that similar problems addressed before are different from the given one. This would additionally fulfill the function of showing that the authors are knowledgeable in their field. But problems need not be new; they might have been addressed by others before (cf. the

I. Moves borrowed from Swales

1. Describe: General Goal
2. Show: Own Goal/Problem is Important/Interesting
3. Show: Solution to Own Problem is Desirable
4. Show: Own Goal/Problem is Hard
5. Describe: General Problem
6. Describe: General Conclusion/Claim
7. Describe: Other Conclusion/Claim
8. Describe: Own Goal/Problem
9. Describe: Own Conclusion/Claim
10. Describe: Article Structure
11. Preview: Section Contents
12. Summarize: Section Contents

II. Moves defined by author stance

13. Show: Other Solution is Flawed
14. Show: Own Solution is Different from Other Solution
15. Show: Own Goal/Problem is Different from Other Goal/Problem
16. Show: Own Claim is Different from Other Claim
17. Show: Other Solution is Advantageous
18. State: Other Solution Provides Basis for Own Solution

III. Moves defined by attribution of ownership

19. Describe: General Solution
20. Describe: Other Solution
21. Describe: Own Solution

IV. Moves defined by problem solving statements

22. Show: Own Solution Solves Own Problem
23. Show: Own Solution is Necessary to Achieve Own Goal
24. Show: Own Solution Avoids Problems
25. Show: Other Solution does not Solve Problem/does not Achieve Goal
26. Show: Other Solution Solves Problem
27. Show: Other Solution Introduces New Problem
28. Show: Own Solution is Better than Other Solution
29. Show: Own Goal/Problem is Harder than Other Goal/Problem

V. High level moves

30. Show: Own Goal/Problem is New
31. Show: Own Solution is Advantageous

Figure 3.16: List of Argumentative Moves

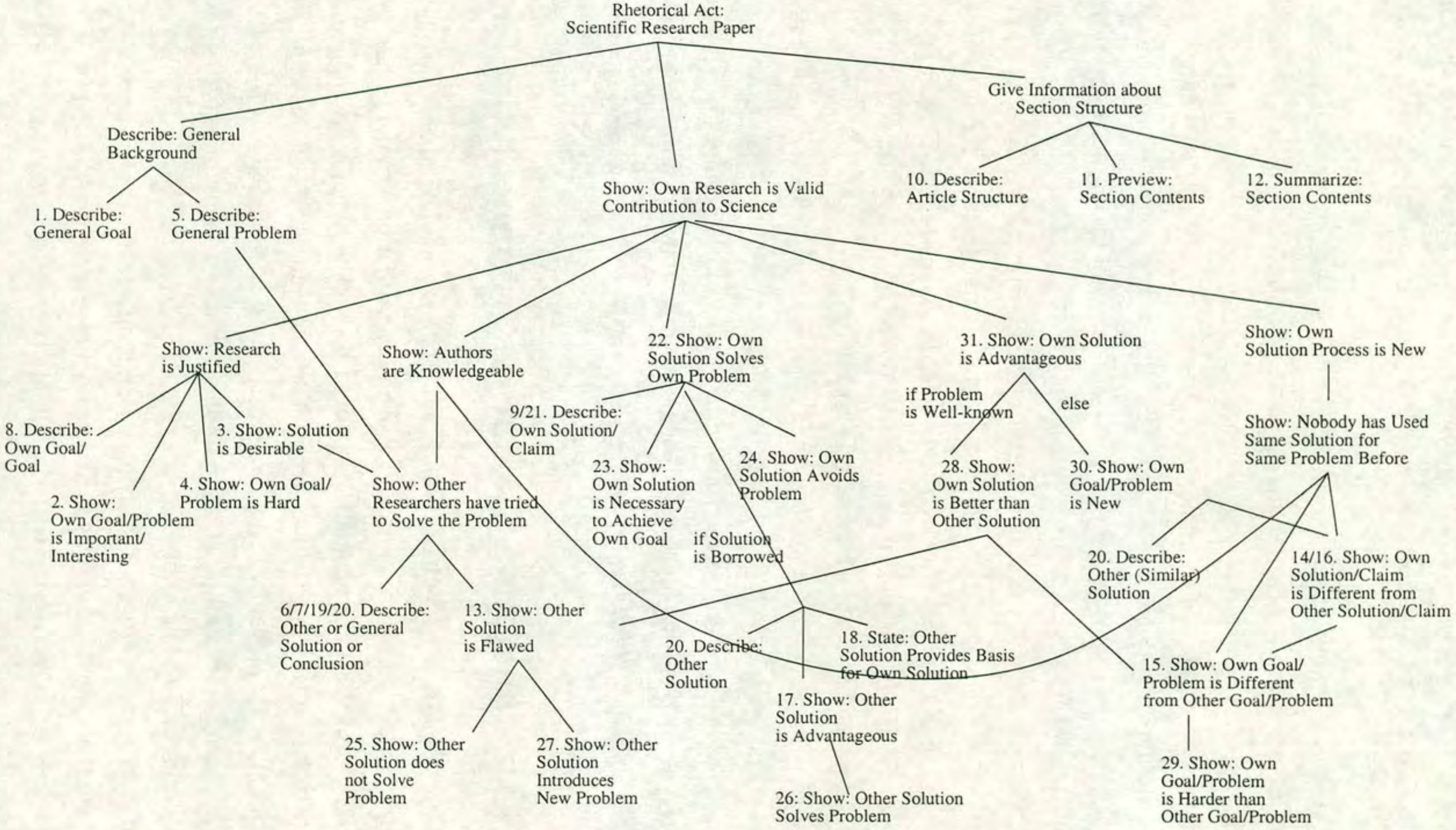

Rhetorical Act:
Scientific Research Paper
Describe: General Background
Show: Own Research is Valid Contribution to Science
Give Information about Section Structure
10. Describe: Article Structure
11. Preview: Section Contents
12. Summarize: Section Contents
1. Describe: General Goal
5. Describe: General Problem
Show: Research is Justified
Show: Authors are Knowledgeable
22. Show: Own Solution Solves Own Problem
31. Show: Own Solution is Advantageous
Show: Own Solution Process is New
8. Describe: Own Goal/ Goal
3. Show: Solution is Desirable
2. Show: Own Goal/Problem is Important/ Interesting
4. Show: Own Goal/ Problem is Hard
Show: Other Researchers have tried to Solve the Problem
9/21. Describe: Own Solution/ Claim
23. Show: Own Solution is Necessary to Achieve Own Goal
24. Show: Own Solution Avoids Problem
if Solution is Borrowed
if Problem is Well-known
else
28. Show: Own Solution is Better than Other Solution
30. Show: Own Goal/Problem is New
Show: Nobody has Used Same Solution for Same Problem Before
20. Describe: Other (Similar) Solution
14/16. Show: Own Solution/Claim is Different from Other Solution/Claim
6/7/19/20. Describe: Other or General Solution or Conclusion
13. Show: Other Solution is Flawed
20. Describe: Other Solution
18. State: Other Solution Provides Basis for Own Solution
17. Show: Other Solution is Advantageous
15. Show: Own Goal/ Problem is Different from Other Goal/Problem
25. Show: Other Solution does not Solve Problem
27. Show: Other Solution Introduces New Problem
26: Show: Other Solution Solves Problem
29. Show: Own Goal/Problem is Harder than Other Goal/Problem

Figure 3.17: Relations between Argumentative Moves

situation in figure 3.11, where a rival solution was suggested). In that case, one needs to show that the own solution is better (28) or that the other solution is flawed (25 or 27).

All of the moves cover a textual span at least as long as a sentence, and in some cases they cover much larger textual spans. Some moves—particularly the moves of type SHOW—can be explicitly stated in one single sentence, but many moves typically span longer segments, for example the moves of type DESCRIBE, which detail problems, solutions and goals in a neutral way and whose purpose is informative rather than rhetorical. We consider the whole move as one unit for our purposes, disregarding possible internal move structure.

Some moves in the diagram tend to occur with other moves, e.g., moves describing other work (6, 7, 19 or 20) co-occur with statements about the role of this other work for the current work (critical stance in moves 13, 25, 27; contrastive stance in moves 14, 15, 16, 29; positive stance in moves 17, 18, 26). Relations of such kinds between moves are not shown in the diagram.

Moves sometimes serve more than one communicative and argumentative purpose at once. The move OTHER RESEARCHERS HAVE TRIED TO SOLVE THE PROBLEM describes the history of the problem, provides background knowledge, proves that the authors know the literature in the field, and it shows that the problem is indeed justified and that a solution is desirable.

3.3. An Annotation Scheme for Argumentative Zones

In the previous section, we have introduced a rather complex model of discourse and argumentative effects in scientific text. We believe that our implicit claim—that the model explains our data adequately—should be substantiated by demonstrating that other humans can apply the account consistently to actual texts. In this section, we will operationalize our model by defining a practical annotation based on it.

In general, designing an annotation scheme has many pitfalls. One wants the annotation scheme to be a) predictive and informative, so that it will prove useful for an end task and b) intuitive, or at least learnable, such that it can be applied consistently by different annotators and over time. If an annotation scheme is simple and intuitive and the task well-described, it will result in high consistency, but there is a danger that the information contained in it might not be informative enough for the given task. On the other hand, if the categories are informative their definition is necessarily vague, leaving a lot of leeway for subjective interpretation. In this case, it is likely that different annotators will disagree in their judgements. The process of finding a workable annotation scheme is thus a tight rope act between the conflicting requirements of informativeness and consistency. This section reports on our quest for a good annotation scheme, and shows why two predecessors of the final annotation scheme fall short of the requirements.

The first annotation scheme (Teufel, 1998) contains 23 categories defined directly by argumentative moves, similar to those in figure 3.16. Such a scheme based on moves is very informative and encodes valuable information for subsequent fact extraction from the sentences. For example, a sentence of type "SHOW: OWN SOLUTION IS ADVANTAGEOUS" contains both a mention of the own solution and a statement of the advantage of the own solution, a fact which could be exploited for information extraction from such a sentence.

We used two unrelated annotators in the definition phase. As is typical for high-level, information-rich classification tasks, the annotation scheme had to be changed repeatedly during this time. Settling on an exhaustive list of moves which annotators agreed on proved very difficult. We were constantly tempted to add more moves for situations where a given sentences does not quite fall into the semantics already defined. Once the scheme mentioned above (23 categories) had emerged, we wrote guidelines detailing criteria for each move.

After the definition phase, we ran a pilot study with our two, by now, task-trained annotators. This experiment revealed that the scheme was not reliable. Even repeated changes to the annotation scheme at this late stage did not improve agreement significantly. Within the mind of one annotator, private understandings of these categories may well be rather consistent—we annotated 10 randomly sampled, previously annotated papers again after 4 weeks and achieved reasonable agreement with the previous annotation (the concept of stability will be introduced in section 4.2). However, if these understandings cannot be communicated to others, something is wrong with the scheme. Low agreement between different annotators (*reproducibility*; detailed in section 4.2) finally convinced us that a fixed, exhaustive list of such high-level categories at this pragmatic level is not universal enough to train annotators.

In order to make the next scheme easier and more objective, we reduced the number of categories and simplified their definitions, while trying to retain as much of the information as possible for our task. Our second attempt at an annotation scheme

B	BACKGROUND
T	TOPIC
W	RELATED WORK
P	PURPOSE/PROBLEM
S	SOLUTION/METHOD
R	RESULT
C	CONCLUSION/CLAIM

Figure 3.18: Annotation Scheme Based on Functional Abstract Units

(figure 3.18) consisted of just seven categories (Teufel and Moens, 1998, 1999a), which are similar to the functional units well-known from summarizing guidelines (cf. section 2.3.1.2).

Again, we achieved respectable stability when re-annotating parts of the corpus. This is a good sign, but we nevertheless noticed fundamental problems with the type of annotation. It proved extremely difficult to associate textual units as big as sentences (i.e. propositional contents) with categories which describe high-level concepts (i.e. nominal phrases). An additional, orthogonal problem was the fact that some high level entities such as PURPOSE/PROBLEM and SOLUTION can be difficult to distinguish in real-world text. To give an example, we were not sure about the right annotation for the following sentence:

> *We then show how different classes of pragmatic inferences can be captured using this formalism, and how our algorithm computes the expected results for a representative class of pragmatic inferences.* (S-29, 9504017)

Is the sentence to be counted as TOPIC, because "*pragmatic inferences*" are the TOPIC of the paper? Or is it rather the case that *"capturing different classes of pragmatic inferences"* is the PROBLEM/PURPOSE? Or should this sentence be classified as SOLUTION, as the phrase *"our algorithm computes the expected results*" could be interpreted as a high level description of the approach used?

Allowing for multiple annotation seemed to ameliorate the problems, but it lead to so many multiply annotated sentences that we started doubting the informativeness contained in this annotation. We redesigned the scheme radically, resulting in the third and final annotation scheme (figure 3.19).

A simpler version of the scheme (the "basic scheme") encodes only intellectual ownership (figure 3.20). Pilot studies with our annotators with both schemes showed that they were much more comfortable and accurate when applying these schemes to real texts. These are the schemes we will use for the extensive human annotation experiments reported in chapter 4 (Teufel et al., 1999), and for the prototypical implementation reported in chapter 5 (Teufel and Moens, 1999b).

BACKGROUND	Generally accepted background knowledge
OTHER	Specific other work
OWN	Own work: method, results, future work...
AIM	Specific research goal
TEXTUAL	Textual section structure
CONTRAST	Contrast, comparison, weakness of other solution
BASIS	Other work provides basis for own work

Figure 3.19: Final Annotation scheme—Full Version

BACKGROUND	Generally accepted background knowledge
OTHER	Specific other work
OWN	Own work: method, results, future work...

Figure 3.20: Final Annotation Scheme—Basic Version

As with the other annotation schemes, the categories are to be read as mutually exclusive labels, one of which is attributed to each sentence. Each category is associated with a colour to make human annotation more mnemonic.

We call the categories which occur only in the full scheme but not in the basic scheme *non-basic* categories (i.e. AIM, CONTRAST, TEXTUAL and BASIS). The seven

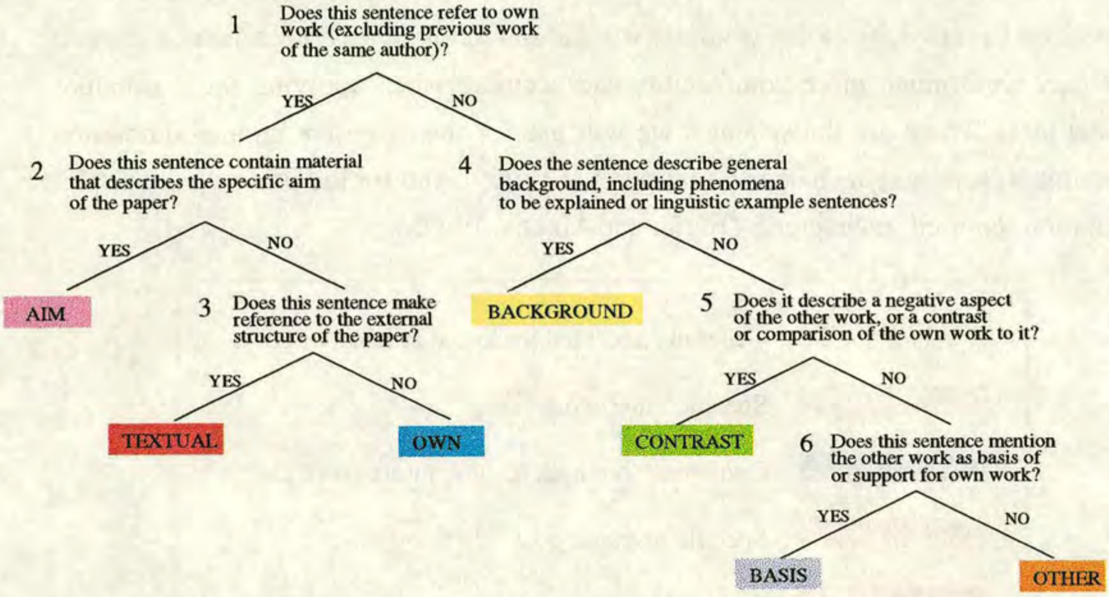

1 Does this sentence refer to own work (excluding previous work of the same author)?
YES
NO
2 Does this sentence contain material that describes the specific aim of the paper?
4 Does the sentence describe general background, including phenomena to be explained or linguistic example sentences?
YES
NO
AIM
3 Does this sentence make reference to the external structure of the paper?
YES
NO
BACKGROUND
5 Does it describe a negative aspect of the other work, or a contrast or comparison of the own work to it?
YES
NO
TEXTUAL
OWN
YES
NO
CONTRAST
6 Does this sentence mention the other work as basis of or support for own work?
YES
NO
BASIS
OTHER

Figure 3.21: Decision Tree for Full Annotation Scheme

categories of the full annotation scheme are closely related to the different aspects of our model (Swales' categories, author stance, intellectual ownership, and problem-solving statements). The semantics of our scheme is best explained with the decision tree in figure 3.21, based on six yes/no questions.

Question 1 focuses on attribution of ownership, distinguishing between statements which describe the authors' own *new* contributions and those which describe research outside the given paper, including the authors' own previous work, generally accepted statements and statements which are attributed to other, specific researchers.

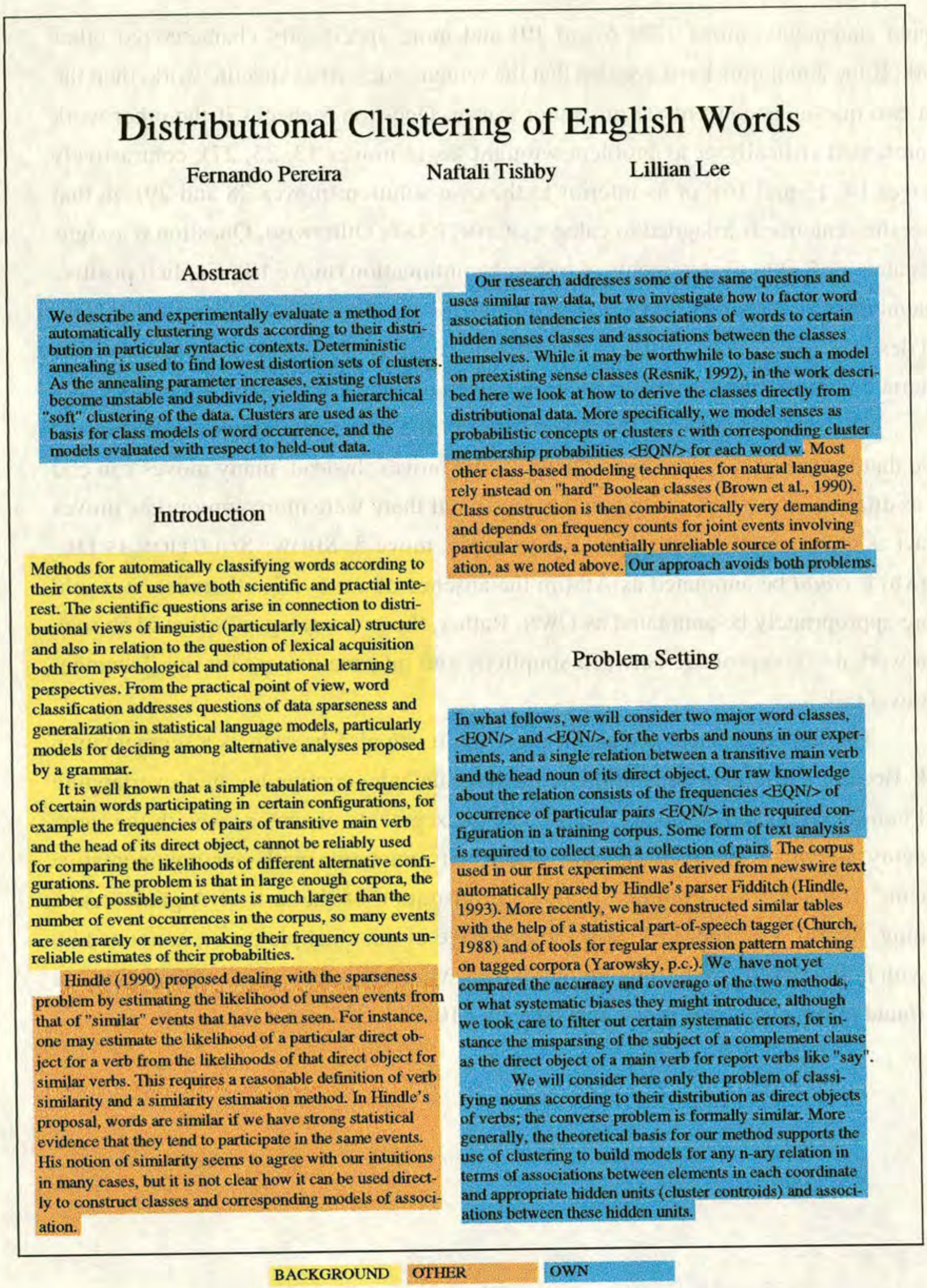
Once annotators decide that the statement describes own work, Question 2 determines AIM sentences. Such sentences describe the research goal addressed in the paper. The most explicit type of AIM sentences is provided by move 8 (DESCRIBE OWN GOAL/PROBLEM in figure 3.16). But dependent on the annotators' intuitions, other moves can in principle be AIM sentences too, e.g. moves 2, 3, 4, 22, 23, 24, 30 and 31.

Question 3 singles out TEXTUAL sentences, i.e. those giving explicit information about section structure. This corresponds to moves 10, 11 and 12. All other statements about own work, in particular move 21, but also all moves not deemed AIM sentences, receive the label OWN.

Question 4 distinguishes between BACKGROUND material (i.e. generally accepted statements; move 1, 5, 6 and 19) and more specifically characterized other work. If the annotators have decided that the sentence describes specific work, then the last two questions concentrate on author stance. Question 5 checks if the other work is presented critically or as problem-wrought (as in moves 13, 25, 27), contrastively (moves 14, 15 and 16), or as inferior to the own solution (moves 28 and 29); in that case, the sentence is assigned to category CONTRAST. Otherwise, Question 6 assigns the category BASIS to statements of research continuation (move 18). Explicit positive statements about other work (i.e. moves 17 and 26) can also be assigned to BASIS. Neutral descriptions of other work get assigned the category OTHER. Details and decision criteria on how to answer the questions are given in the guidelines (cf. appendix C.2).

The relation between the categories and the moves is complex: it is not the case that the categories are super-classes of the moves. Instead, many moves can end up as different zones, depending on the question if there were more appropriate moves to act as argumentative categories. For example, move 3. SHOW: SOLUTION IS DESIRABLE *could* be annotated as AIM in the absence of a move 7; otherwise, it would more appropriately be annotated as OWN. Rather, the seven categories should be seen as a workable compromise between simplicity and informativeness for our document retrieval task.

The task is defined as classification, but it can also be seen as a segmentation task. Because the kind of annotation we envisage includes contiguous, non-overlapping and non-hierarchical sequences, we refer to the segments of sentences with the same category as *zones*. We then call the process of annotation with our argumentative scheme *Argumentative Zoning*. To give an illustration of the task of Argumentative Zoning, figures 3.22 and 3.23 show the first page of our example paper, annotated by us with both versions of the annotation scheme. More human example annotations can be found in the guidelines in the appendix (p. 310, 311, 327 and 328).



Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

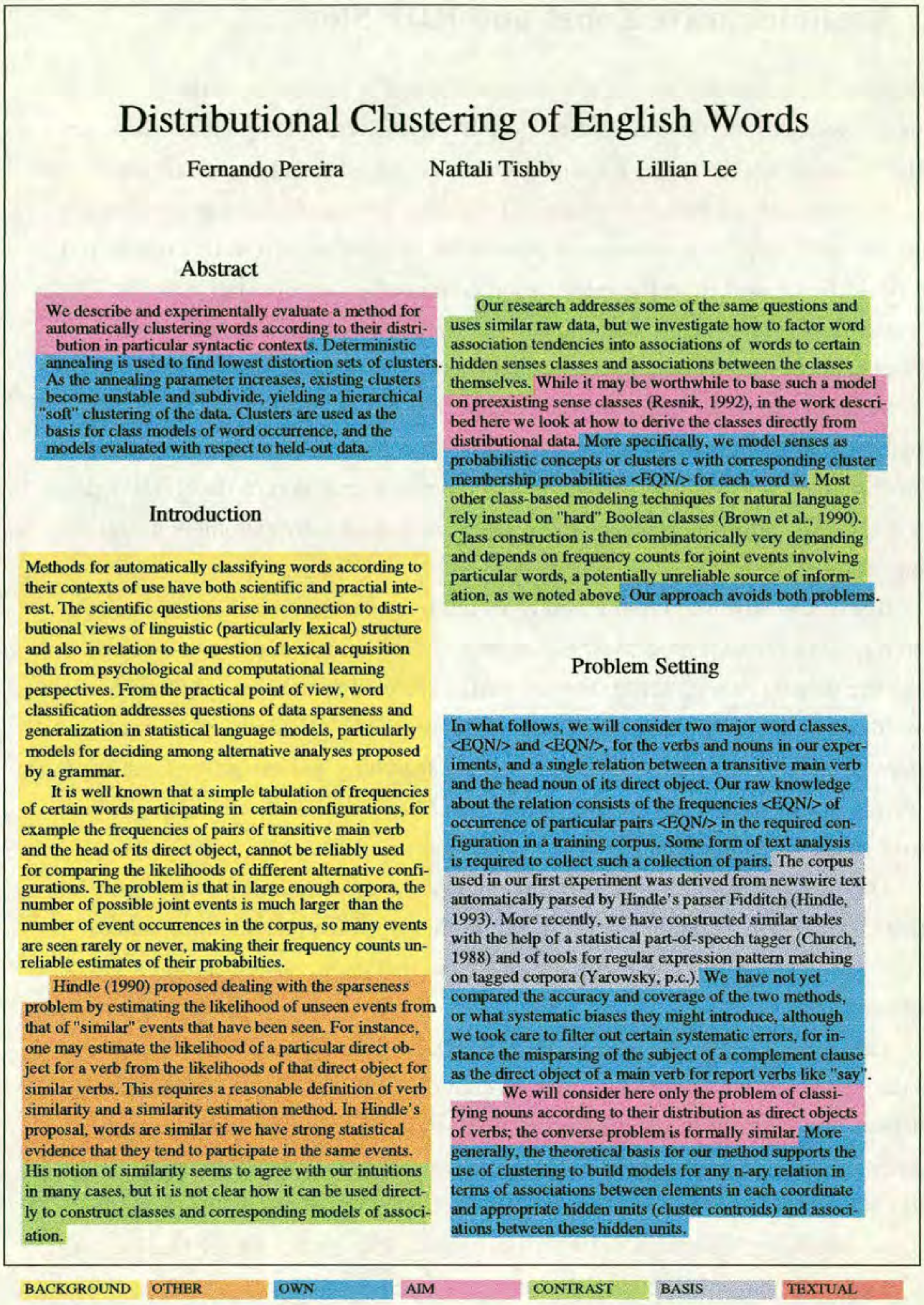
Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster centroids) and associations between these hidden units.

BACKGROUND OTHER OWN

Figure 3.22: First Page of Example Paper, Annotated with Basic Annotation Scheme



Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster centroids) and associations between these hidden units.

BACKGROUND OTHER OWN AIM CONTRAST BASIS TEXTUAL

Figure 3.23: First Page of Example Paper, Annotated with Full Annotation Scheme

3.4. Argumentative Zones and RDP Slots

In this thesis, we originally set out to generate RDPs. The semantics of the individual argumentative zones are obviously very close to RDP slots, but argumentative zones and RDP slots are not the same. We will now discuss the relation between the two.

Argumentative zones can be seen as providing the material of text which *might* go into the RDP slots. In a subsequent processing step not treated in this thesis, full RDPs could be created from the information contained in argumentative zones. The RDP presented in section 2.3.2 was manually created based on an annotated version of the example paper, obtained in the annotation exercise to be described in chapter 4.

Some of the zones, the non-basic categories, are short and contain important information; they can therefore act as *direct slot fillers* without requiring much further work. AIM zones, for example, constitute a good characterization of the entire paper, which is typically only one sentence long. They are thus already extremely useful for the generation of abstracts.

But BACKGROUND, OTHER and OWN are longer zones, which should be seen as search ground for later processes. For example, as simple sentence extraction does not take the context of a sentence into account, a selected sentence might turn out to be describing *other* people's work. This is a grave error, particularly if the sentence expresses a statement which the authors reject. By searching and extracting from argumentatively zoned articles, where zones such as OWN and OTHER are distinguished, this error should be eliminated.

There is another task which argumentative zones as search ground is useful for. This task is the association of identifiers of other work (formal citations, names of researchers, names of solutions) with the statement that expresses the author's stance towards the work.

This task is needed in order to generate RDPs from argumentative zones. Our approach has a more concise definition of citation context (cf. O'Connor's (1982) work) than previous approaches. Citation maps display only *one* sentence, namely the sentence which expresses the evaluative statement. In contrast, Lawrence et al.'s (1999) CiteSeer (which displays contexts in a text extract fashion, cf. the example on p. 34), and Nanba and Okumura's (1999) tool operate with a much *larger* citation context. Consider Nanba and Okumura's example of a contrastive citation context (taken from p. 927):

> **1** In addition, when Japanese is translated into English, the selection of appropriate determiners is problematic.
> **2** Various solutions to the problems of generating articles and possessive pronouns and determining countability and number have been proposed **[Murata and Nagao, 1993]**.
> **3** The difference between the way numerical expressions are realized in Japanese and English has been less studied.
> **4** In this paper we propose an analysis of classifiers based on properties in both Japanese and English.
> **5** Our category of classifier includes both Japanese josushi 'numerical classifiers' and English partitive nouns.

Nanba and Okumura's tool displays sentences **2–4** (the *reference area*). In our approach, only sentence **3** would be displayed, which implies that one must additionally determine which other work the current context refers to. In this case, the formal citation in sentence **2** must be extracted. As an additional difficulty, the authors might have used different kinds of identification of the other work, e.g. author name or solution identifier. We aim to treat these types of identification alike, instead of recognizing only formal citations (like Nanba and Okumura do).

Nanba and Okumura's approach relies on the simplifying assumption that identification and citation of an approach occur in the same sentence, or at least very close together. However, this does not have to be the case. In our example paper, the description of the work of Hindle (1993) and its weaknesses extends from sentences 5 to 9. Textual separation is an issue that needs to be addressed, as it is even *more* likely for important references, where the authors will take some time and space describing the other work (we also noticed that textual separation is more likely for CONTRAST zones than for BASIS zones, as these are often longer).

Argumentative zones can help us associate textual spans belonging to authors' descriptions of other work because of regularities between zones which we call rhetorical patterns. For example, neutral descriptions of other researchers' work often occur in combinations with statements expressing a stance towards that work. We believe that those kinds of dependencies can be helpful for automatic Argumentative Zoning: in section 5.3.4.2, we will use an ngram model operating over sentences to model these regularities. From informal inspections of our corpus, however, we suspect that in our corpus the dependencies are not as strong as Swales' claims about fixed order would imply—possibly due to the interdisciplinarity of our corpus.

Figure 3.24 illustrates typical argumentative patterns. The identifiers (i.e. re-

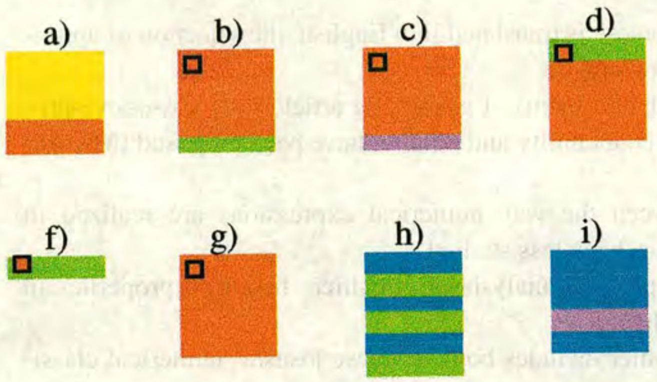


Figure 3.24: Typical Rhetorical Patterns

searchers' names, formal citations or solution names) are signified by small squares.

a) General statements typically precede more specific ones; e.g., general background material is followed by descriptions of specific other work.

b) A prototypical pattern for CONTRAST: The other solution is identified, described and criticized.

c) A prototypical pattern for BASIS: The other solution is identified and described, then a statement of intellectual ancestry follows.

d) The other work is identified and criticized before it is described. This pattern is rather common, though it does not occur as frequently as pattern b).

e) The other work is identified after it has been described and criticized. This pattern reads somewhat awkwardly, but it does occur several times in our corpus.

f) A less important contrastive approach which does not get much "real estate" in the paper.

g) Other work is introduced and identified, but no stance is expressed. In section 3.2.2 we argued that such patterns contribute nothing to the argumentation and that the authors waste space in the paper which such moves. Nevertheless, we found many such patterns in our corpus. One of the possible reason why they were were used nevertheless is that they serve the move SHOW: AUTHORS ARE KNOWLEDGEABLE. As predicted, most of patterns g) found in our corpus are short, i.e. the work is presumably not crucial to the argumentation.

h) After the own solution has been introduced, advantages of it can be presented by comparisons to other work, often in parallel steps. This is a prototypical pattern for comparisons with other work, particularly in conclusions and discussion sections.

i) A statement of intellectual ancestry occurs in the middle of a description of the own solution. We found that if some other work is cited in an OWN segment, it is generally more likely to be a BASIS zone than a CONTRAST zone. BASIS zones are also overall shorter than CONTRAST zones; many of these statements just state the fact that work is based on other work, or acknowledge methods or data used.

In an approach based on Argumentative Zoning, adjacency of argumentative zones and assumptions about their connection to a given zone can be used to find the most likely citation association. For example, if a zone expressing author stance has been identified which does not contain an identifier, adjacent zones of other researchers' work can be searched for identifiers most likely to be associated with the zone.

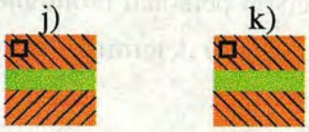


Figure 3.25: Likely and Unlikely Rhetorical Patterns

An aide in this could be provided by the following observation which is illustrated in figure 3.25: we found that if two zones of neutral description occur around a criticism zone, it is very unlikely that the neutral zones refer to the same work (as in j); it is far more likely that they refer to different work (as in k).

Additionally, argumentative zones could be used in Content Citation analysis to provide a simple and automatic means of estimating the importance of a cited work for the citing work, as more relevant OTHER work will probably receive more space in the article.

3.5. Related Work

Argumentative Zoning is a new task, but there is much work in computational and theoretical linguistics and in language engineering which is closely related. Firstly, there are other types of zoning of text, i.e. methods which break documents into segments; it is the definition of the zones which is new in our approach. While most other approaches try to segment papers into topic-related zones (Morris and Hirst, 1991; Hearst, 1997), our approach is more similar in nature to Wiebe's (1994) work. Her approach also attempts to determine a rhetorical feature, namely *evidentiality* or point of view in narrative. The task is to determine the source of information in text which might be either subjective or objective. In news reporting and narrative, this distinction is important as coherent segments presenting opinions and verbal reactions are mixed with segments presenting objective fact. Her four categories are given in figure 3.26 (examples taken from Wiebe et al. 1999, p. 247).

Subjectivity is a property which is related to the attribution of authorship as well as to author stance, but there are obvious differences between Wiebe's and our distinction, which are rooted in differences between the text types covered. As will be discussed in chapter 5, some of the sentential features we use are comparable to hers (e.g. occurrence of first or third person personal pronouns). However, her processing does not go as "deep" as ours in trying to determine the agent/action structure of the text.

Another kind of discourse segment altogether is defined by topic segments (Morris and Hirst, 1991; Kozima, 1993; Hearst, 1997; Kan et al., 1998; Raynar, 1999). The general notion behind work like this is that there is a connection between the discovery of aboutness or discourse topics and textual organization.

Practical work in topic segment determination goes back to Skorochod'ko

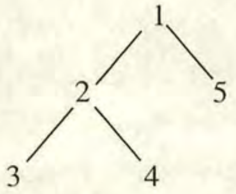
Subjective	*At several different levels, it's a fascinating tale.*
Objective	*Bell Industries Inc. increased its quarterly to 10 cents from seven cents a share.*
Subjective Speech Act	*The South African Broadcasting Corp. said the song "Freedom now" was "undesirable for broadcasting".*
Objective Speech Act	*Northwest Airlines settled the remaining lawsuits filed on behalf of 156 people killed in a 1987 crash, but claims against the jetliner's maker are being pursued, a federal judge said.*

Figure 3.26: Wiebe's (1994) Subjectivity Categories

(1972) who makes the connection between topical segmentation and relatedness of terms: whenever the value of "semantic relatedness" of a sentence with respect to the preceding chunk of sentences falls below a threshold, he proclaims a new topical text segment to begin. This idea is taken up in approaches to topic segmentation such as Hearst's (1997) TextTiling. The assumption is that words which are related to a certain topic will be repeated whenever that topic is mentioned, and that the choice of vocabulary will change when a new topic emerges. Hearst determines boundaries of *topic segments* by calculating vocabulary similarity between two adjacent windows of text. Similarity is defined using the frequency of non-stop word terms in each segment, without taking their inverse document frequency into account. Variations of her approach are discussed in Richmond et al. (1997) where the concepts of *global frequency* and *local burstiness* (proximity of all or some occurrences of multiply occurring content words in a text) are used to refine the definition of segment similarity. Raynar's (1999) system works by similar principles, but includes a range of other heuristics, similar to the ones used in text extraction methods (cf. section 2.2.1).

Our work is different in its interest in rhetorically, rather than topically, coherent segments. The argumentative zone a sentence belongs to is a distinction which often cuts across subtopic zones. One subtopic might be mentioned in several adjacent argumentative zones. For example, the name of a problem might be repeated in the introduction, in the description of other researchers' work, the statement which describes weaknesses of that work, in the goal statement and in the description of the own solution. On the other hand, some of our larger zones, particularly the OWN zone, will contain many subtopics. Thus, the apparent similarities between topic segmentation methods and Argumentative Zoning are superficial.

There is a second group of work, providing models of argumentation which have a more general aspiration, analyzing argumentative scientific discourse from a theoretical and logic point of view (Toulmin, 1972; Perelman and Olbrechts-Tyteca, 1969; Horsella and Sindermann, 1992; Sillince, 1992). Argumentation in these approaches is concerned with arbitrary facts about the world and their relation. For a computational treatment to cover this, full text comprehension would be required. Cohen's (1987) work is more computationally minded. It is a general framework of argumentation for all text types, based on the construction of claim-evidence trees from argumentative text (cf. figure 3.27, taken from Cohen 1987, p. 15):



1 The city is a disaster area
2 The parks are a mess
3 The park benches are broken
4 The grassy areas are parched
5 Returning to city problems, the highways are bad too

Figure 3.27: Cohen's (1987) Evidence-Claim Trees

Argumentative structure in her approach is related to linear order and surface meta-discourse ("clues") like the phrase "*returning to city problems*". Processing is incremental; rules express where in the tree incoming propositions can be attached. This is similar to Polanyi's (1988) discourse grammars where the rightmost node at each level of the tree is always open and all other nodes closed for attachment. Cohen suggests the implementation of a separate clue module within her framework and considers clue interpretation as "not only quite useful but feasible" (p. 18).

Cohen's approach is not implemented. The reason for this is that it presumes a "evidence oracle" which can determine if a certain incoming proposition is evidence for another statement already in the discourse tree. This is a hard task, requiring general inference on the object level which we are trying to avoid at all cost.

An approach for the *generation* of natural language arguments is given by Reed and Long (1998) and Reed (1999). The approach is based on argumentation theory (cf. van Eemeren et al. (1996) for an overview). Their RHETORICA system uses planning to generate persuasive texts by modelling users' goals and beliefs. Apart from the fact that this approach is not concerned with the *analysis* of arguments, the biggest difference between this work and ours is that instead of formally manipulating relations between facts in the world we model *prototypical (fixed) scientific argumentation* in a far more shallow way.

The third group of work related to Argumentative Zoning are discourse theories for rhetorical structure. Discourse structure is concerned with two aspects of the organization of sentences: a) the fact that the sentences in one topical or rhetorical segment of the text are in relation to each other and b) that different segments also have an inter-segmental ordering of intentional relations. This is often referred to as *micro vs. macro-structure* (van Dijk, 1980). Other names for macro-structure are discourse-level structure, or large scale text structure. In a well-written text, the function of micro segments with respect to the macro segment, as well as the function of a macro segment with respect to the text as a whole, is signalled by surface cues. Cues at micro-level are for example connectives between clauses (*"but, thus"*) or enumeration markers (*"first, second, last... "*). Cues at macro level are phrases of the kind *"next we will show that... "*.

We consider here general theories of text structure which are based on intentional or communicative acts of the writer. Examples of rhetorical functions are "to convince a reader", "to provide an example" or "to recapitulate". The common assumption is that in trying to communicate a (set of) messages, e.g., in an argumentative text, humans employ a hierarchical intentional structure.

A bottom-up approach to rhetorical relations, based on a model of human memory organization, is described in the seminal paper by Kintsch and van Dijk (1978). Their main claims about discourse organization are that text content is hierarchical and that relevance is an aspect of discourse organization. Their model starts from a manually-created, logical, but surface-oriented representation for propositions. Connectedness is calculated using the overlap of grammatical arguments in this representation. Even though their theory of text comprehension is plausible, we do not consider it here, as their approach bypasses the essential text analysis phase—this means that it cannot be used for practical summarization of unrestricted text (section 2.2). Instead, we turn to theories which work by considering more superficial cues.

Grosz and Sidner (1986) present a hierarchical discourse structure based on three types of structure: linguistic, intentional and attentional. Intentional structure in their model is defined by those intentions that the writer or speaker intended the hearer to recognize (in contrast to private intentions like to impress somebody). Intentional structure is associated with linguistic units, discourse segments. Two structural relations (dominance and satisfaction-precedence) hold between the segments. In contrast to Swales' model, and similar to Cohen's, an infinite number of different intentions is possible.

Grosz and Sidner state that three kinds of information play a role in the determination of the discourse segments: specific linguistic markers, utterance-level intentions and general knowledge about actions and objects in the domain of discourse. One of their main claims is that the use of certain linguistic expressions like referring expressions is constrained by the attentional structure. The attentional structure contains information about the different possible foci of attention in the conversation: salient objects, properties and relations.

The need to recognize the intentions and their relation to previous intentions is aided in Grosz and Sidner's example, as a strongly hierarchical task-structure underlies their example dialogue. This task-structure provides common knowledge about the task and also acts as a special case of the intentional structure posited.

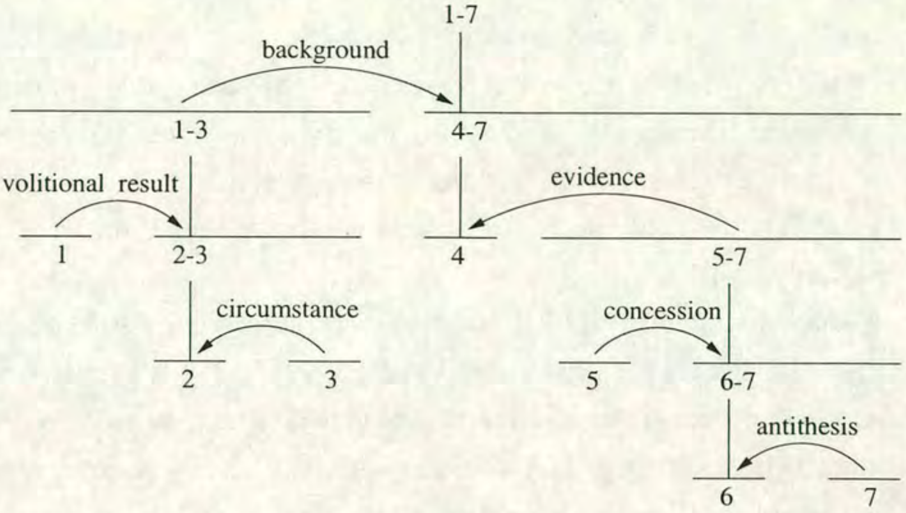
Rhetorical Structure Theory (RST; Mann and Thompson 1987, 1988) is also based on the notion that text structure serves a communicative role. In contrast to Grosz and Sidner, the document structure is based on a *fixed* set of rhetorical relations holding between any two adjacent clauses or larger text segments. Their main claims are that discourse is characterized by strong hierarchical relations and by the predominance of structural patterns of nucleus/satellite type. The relations are typically asymmetric and include CIRCUMSTANCE, SOLUTION-HOOD, ELABORATION, BACKGROUND, ENABLEMENT, MOTIVATION, EVIDENCE, JUSTIFICATION, CAUSE (VOLITIONAL AND NON-VOLITIONAL), RESULT (VOLITIONAL AND NON-VOLITIONAL), PURPOSE, ANTITHESIS, CONCESSION, CONDITION, INTERPRETATION, EVALUATION, RESTATEMENT, SUMMARY, SEQUENCE and CONTRAST. The definitions of the rhetorical relations are kept general on purpose, as illustrated by the one for JUSTIFY:

> JUSTIFY: a JUSTIFY satellite is intended to increase the reader's readiness to accept the writer's right to present the nuclear material.
>
> (Mann and Thompson, 1987, p. 9)

During the analysis, the analyst effectively provides a plausible reason the writer might have had for including each part of the whole text, cf. figure 3.28, taken from (Mann and Thompson, 1987, p. 13–14).

Ambiguity of relations and structure are considered normal in RST (Mann and Thompson, 1987, p. 28). This vagueness poses a problem for computational applications as it leads to multiple RST analyses for a given piece of text. Another dilemma is that researchers building their work on RST have often invented their own, similar relations, such that there was a proliferation of private RST-like schemes; Maier and Hovy (1993) list more than 400 RST-type relations used in the field. This dilemma could be mitigated by a corpus-based approach like Knott's (1996).

Another difficulty is the unit of annotation. It has long been debated, and is still entirely unclear, what the formal linguistic criteria defining such units might be. Consider, for example, unit 7 in figure 3.28 ("*not laziness*"). This unit has been determined



1 Farmington police had to help control traffic today
2 when hundreds of people lined up to be among the first applying for jobs at the yet-to-open Marriott Hotel.
3 The hotel's help-wanted announcement—for 300 openings—was a rare opportunity for many unemployed.
4 The people waiting in line carried a message, a refutation, of claims that the jobless could be employed if only they showed enough moxie.
5 Every rule has its exceptions,
6 but the tragic and too-common tableaux of hundreds of even thousands of people snake-lining up for any task with a paycheck illustrates a lack of jobs,
7 not laziness.

Figure 3.28: Sample RST Analysis

as "clause-like" as it obviously carries a lot of information in this particular argument. However, syntactically, this unit is only a single NP in a VP ellipsis construction—one is now in need of a general syntactic criterion which defines this phrase as a clause, but excludes similar other NPs.

RST has been extensively and successfully used for text generation, e.g. of tutor responses (Moore and Paris, 1993), and of texts describing ship movements and air traffic control procedures (Hovy, 1993). For this purpose Moser and Moore (1996) suggest a synthesis of RST and Grosz and Sidner's theory. On the analysis side, a problem of recognizing RST relations is that most rhetorical relationships are not explicitly marked by connectives, or that it is not clear at which level in the tree a given unit should connect.

Marcu uses heuristics based on punctuation and cue phrases to recognize fully hierarchical RST structure in popular science text (Marcu, 1997a, 1999a,b). One of the applications of the generated structure is summarization. The texts Marcu uses are heavily edited, unlike ours; this makes parsing easier as punctuation can be expected to be standardised. The texts are also well-written: whereas in our texts experts communicate with experts, these texts are aimed at making a possibly non-expert audience understand difficult scientific facts. To do so, causality and other rhetorical relations are often overtly signalled.

Another system that uses RST relations for summarization (of Japanese texts) is BREVIDOC (Miike et al., 1994; Sumita et al., 1992; Ono et al., 1994). Connective expressions in sentences are identified and used to build a representation of the rhetorical relations between sentences. A cumulative penalty scoring technique is used to select the most plausible binary tree. Abstracts of variable length are produced interactively from this structure.

At first glance RST-type rhetorical relations might look a bit like RDP slots, but they have a different status: whereas RST models micro-structure, i.e. relations holding between clauses, RDP slots denote macro structure, i.e. global relations between the given statement and the rhetorical act of the whole article.

While we agree with RST that micro-level structure is likely to be hierarchical and can be well described by RST relations, we choose not to model these relations. For example our move DESCRIBE: OWN SOLUTION, which is particularly long, includes a description of the methodology, evaluation strategy etc. The internal hierarchical structure of this move does not receive any attention in our approach, because we believe that many of the local rhetorical relations between sentences and clauses are irrelevant for our task.

We believe that it is macro-structure and not micro-structure which is useful for summarization and document representation. We also believe that RST is not ideally suited to model macro-structure and that macro-structure is more usefully described by an annotation scheme like ours. When humans are asked to assign RST relations between between paragraphs and larger segments, they often have to resort to the trivial RST relation JOINT. There seem to be fewer constraints on relations between such segments, and we doubt that this structure is hierarchical in the same way that micro-level relations are.

A related fact showing that it is indeed *micro*-level relations that are modelled by RST is the fact that the cue phrases used in RST approaches tend to be connectives, which operate between clauses (Knott, 1996; Marcu, 1997b).

Moreover, even though Mann and Thompson (1987) claim that RST is "unaffected by text size and has been usefully applied to a wide range of text size" (p. 46), RST analysts typically use short texts. Marcu (1997b), for example, uses text with an average of 14.5 sentences, and Mann and Thompson describe a text of 15 utterances as a "larger text" (p. 22)—whereas we wanted to reliably annotate articles several pages long.

To summarize our observations from looking at intention-based accounts, hierarchical intentional relations at micro-level might not be necessary for our task; we believe that global text structure is far more important. Secondly, rhetorical relations between two segments can be recognized by overt clues if they are present. If they are not, there is a problem. The remaining possibilities are the following, all of which are not very appealing:

- One could use simple, short, well-edited texts with standardized punctuation (Marcu, 1997a).

- One could use task-structured texts (Grosz and Sidner, 1986).

- One could posit an "evidence oracle", i.e., put the task outside one's remit (Cohen, 1987).

- One could perform "deep" intention modelling and recognition (Pollack, 1986).

In contrast, the task of Argumentative Zoning relies on more superficial expressions of scientific argumentation.

3.6. Conclusion

We have introduced a model of scientific argumentation which describes the argumentative structure of the articles in our corpus. This model incorporates ideas from Swales' CARS theory of argumentative moves, a certain view on the problem-structure of scientific research and authors' statements about problem-solving processes, a distinction of contrastive vs. continuative author stance, and our own observations about the attribution of ownership in scientific articles. We have operationalized this model as a 7-pronged annotation scheme. We call the process of applying it to text, i.e. of determining the rhetorical status of each sentence, *Argumentative Zoning*.

We conclude that texts and discourses can have multiple structures at the same time, which are not necessarily isomorphic. Certain structures are particularly dominant in some *text types*, and certain structures are particularly useful for some *tasks*. It seems that for scientific texts our model—relying on fixed, text-type specific argumentative moves—describes one such structure for which both is true at the same time.

The novel aspects of our scheme are that it applies to different kinds of scientific research articles, because it relies on the *form and meaning of argumentative aspects* found in the text type rather than on contents or physical format. It should thus be independent of article length and article discipline.

Other structural descriptions, though useful in their own right, do not fit as nicely to both task and text type: the fixed rhetorical structure of scientific articles, described by models like van Dijk's, Kando's and Kircz', relies on expectations specific to certain domains and therefore cannot describe our data well. General frameworks such as the ones discussed in the previous section, however, do not exploit text type-specific expectations and therefore cannot offer much help for automatic structure recognition.

Figure 3.29 shows the role of RDPs and Argumentative Zoning as intermediaries between reader and writer: whereas RDPs are a representation of what the *reader* wants out of a text (cf. chapter 2), argumentative zones are a representation of what the *author* put into the text.

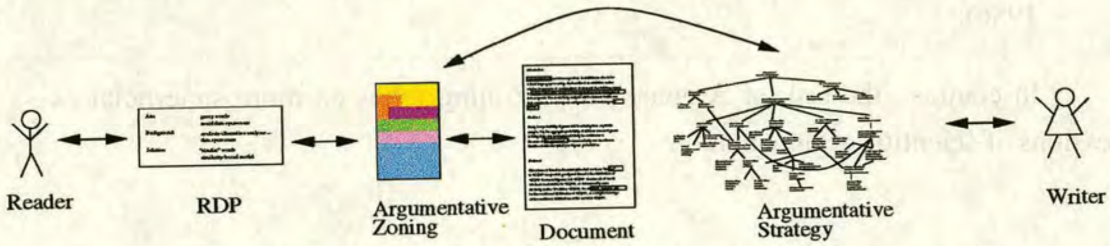


Figure 3.29: Argumentative Zoning and RDPs

This chapter has recast the task of building RDPs as that of Argumentative Zoning. The following questions about Argumentative Zoning now have to be asked:

- How intuitive is Argumentative Zoning? Are the definitions of our categories meaningful to other humans? To answer this question, we observed human annotation with our annotation scheme on naturally occurring, unrestricted text.

We will show in chapter 4 that humans can perform Argumentative Zoning robustly.

- How well can Argumentative Zoning be performed automatically? To answer this question, we built a prototype that applies the scheme automatically, as reported in chapter 5. The results show that Argumentative Zoning can be performed automatically in a robust fashion, although humans are substantially better at the task.

Chapter 4

A Gold Standard for Argumentative Zoning

In the previous chapter, we have introduced a new task: Argumentative Zoning. We will in this chapter define the specifics of the task in such a way that we end up with *gold standards* for it: a definition of what the "right answer" for a set of example documents should look like. For any new task, the right evaluation method is an essential design criterion. Of course, it is essential that the gold standards be defined *before* the experiment, and *independently* of it.

Gold standards are also needed during system development. In chapter 5, we will describe an automatic procedure for determining argumentative zones. We will use our gold standards to determine sentential features and to provide training material. Importantly, gold standards serve for progress evaluation: the evaluation of day-to-day changes to current versions of the system.

Section 4.1 is concerned with finding the right evaluation strategy for Argumentative Zoning. As it is a new task, there is no existing evaluation strategy for it, but the evaluation strategies for similar tasks can inform our decision. We decide to use human judgement; we will then discuss how exactly to define the task in such a way that the similarity of such judgements on the task can be measured objectively.

We will then discuss which numerical evaluation measures to use for the reliability studies (section 4.2). The rest of the chapter is dedicated to describing the reliability studies which measure how much our human annotators agree when they perform Argumentative Zoning.

4.1. Evaluation Strategy

In sections 4.1.1 and 4.1.2, fact extraction and text extraction approaches are contrasted in the light of their evaluation strategies. This contrast will lead to a list of desired properties for our gold standard, and motivate our concrete evaluation strategy for Argumentative Zoning in section 4.1.3.

4.1.1. Evaluation of Fact Extraction

In template-filling tasks like the Message Understanding Conference (MUC; cf. section 2.2.2), the gold standards are called *answer keys*; they are provided by information specialists. Evaluation proceeds by direct comparison of the slot fillers presented by the competing systems with the answer keys.

Answer keys can be of different kinds: they often consist of extracted textual strings, e.g. NPs; sometimes the answer is one of a fixed set of answers ("Was the position newly created, or had it existed before?"). These fixed-choice slots often require inference from subtle linguistic cues. Slots can also be filled by pointers to other templates, or may contain numerical values which the systems have to calculate if the values are not present in the text.

It is easy for humans to assess the correctness of these answer keys after having read the text, as the slot semantics is concrete and domain-specific. However, even though humans can *decide* whether an answer key is correct or not, it is still not an easy task for human experts to *fill* template slots consistently. For more complex slots, there might be two different, but equally appropriate ("correct") keys—superficially different material, coming from different places in the document.

Sometimes, there is an overlap problem, e.g. when one annotator decides to include an apposition of an NP in a slot and the other does not. Annotation guidelines (Chinchor and Marsh, 1998) provide decision criteria for this and other problematic cases.

To measure how often annotators disagree, a subset of the materials (about 30% of the texts), is provided with answer keys by more than one expert. The keys of one annotator are taken as gold standard in turn, and percentage agreement is calculated, i.e. the percentage of identical keys over total keys. A full discussion of evaluation measures for tasks like this is given in section 4.2. In MUC, only reproducibility is reported (e.g. 83% for Scenario Templates); no stability tests are conducted, i.e., it is not measured if the same annotator will annotate in a similar way at a different point in time.

There are disadvantages associated with this type of gold standard. A simple comparison of *one fixed* answer key does not incorporate enough flexibility to deal with cases where the system's answer is different from the answer key. As we cannot perform deep understanding, we need a fair comparison method which deals with *surface* strings. Direct surface comparisons might punish the system unfairly: the answer might be a string which looks different but means something very similar to the given answer key. Fairer system evaluation should give the system a score better than zero in case a second-best answer is retrieved by the system instead of the best answer. What is needed is a gold standard which can provide some kind of fall-back option, i.e. other acceptable—albeit less relevant—answers.

4.1.2. Evaluation of Text Extraction

Gold standards consisting of whole sentences—*target extracts*, i.e. a set of sentences that together constitute the best possible extract from a document—are still the most typical gold standard for text extraction-type summarizers, as target extracts allow for a simple comparison with the machine produced extracts. The problem of evaluation seems to get simpler when gold standards are always full sentences; at least, there is no overlap problem, as there might be with MUC answer keys.

There are different methods whereby one could achieve a target extract, e.g. by asking humans to select important sentences from the text, or by finding other independent, objective criteria for "extract-worthiness", e.g. similarity of document sentences with sentences in a human-written abstract.

4.1.2.1. Free-selecting Sentences from Documents

Early researchers developing corpus resources for summarization work have often defined their own target extracts, relying only on their intuitions (see, e.g. (Luhn, 1958; Edmundson, 1969)). Some have tried a more objective approach by asking unrelated humans to prepare a target extract, i.e. subjects which are not involved in the process of automatic summarization. Several researchers report reasonable agreement between their subjects (Klavans et al., 1998; Zechner, 1995) for free-selecting sentences from newspaper text.

Using unrelated subjects, however, still does not guarantee objectivity: Paice and Jones (1993) reject the use of free-selected sentences for the evaluation of their template-generated summaries, as a small trial showed that their (expert) subjects' selection strategies were very heavily biased towards their individual research interests.

The texts chosen are typically short, so that there are few alternative sentences that *could* have been chosen by the subjects, and the journalistic style makes the selection easier still: the most important sentences will be found in the beginning (Brandow et al., 1995).

For scientific text, the level of subjectivity needed for the task might be higher. Rath et al. (1961) report low agreement between human judges carrying out free selection. If six subjects were asked to select 20 sentences out of *Scientific American* texts ranging from 78 to 171 sentences, all six of them agreed only on 8%, and five agreed on 32% of the sentences. Rath et al. also found that annotators only chose 55% of the sentences they chose six weeks ago. Edmundson (1961) reports similarly low human consistency.

The text extraction evaluation strategy also suffers from surface comparability problems: an ideal gold standard should treat two or more sentences in the text alike, if they express the same semantics. However, target extracts do not account for the cases where two sentences are directly replaceable, or where two sentences taken together contain roughly the same information as another one. There is not a single best target extract for a document:

> [the] lack of inter- and intra subject reliability seems to imply that a single set of representative sentences does not exist for an article. It may be that there are many equally representative sets of sentences which exist for any given article. (Rath et al., 1961, p. 141)

4.1.2.2. Abstracts as Gold Standards

One would ideally want a gold standard which allows different research teams to replicate the gold standard. Asking humans to select sentences does not provide this level of objectivity, of course, as relevance is situational (cf. section 2.1). Researchers have thus looked for an independent, fixed definition of relevance which comes with the text itself and which cannot be influenced anymore, e.g. one that is based on a historic decision of a professional (the indexer or the abstractor). Such a gold standard could be given by a back-of-the-book index (Earl, 1970), or by the human-written abstract (Kupiec et al., 1995).

Earl used a back-of-a-book index to identify all sentences in a book chapter that contained an indexed term; these *indexible* sentences constitute her gold standard. But scientific articles do not typically contain back-of-a-book indexes. Kupiec et al. (1995) use the summary supplied with the article instead to define the gold standard sentences: their gold standard is the set of sentences in the source text that are maximally similar ("align") with a sentence in the summary. An automatic similarity finder is used to identify potential pairs of summary and source text sentences by superficial criteria; subsequently, a human judge (presumably one of the system developers) decides if the alignment is justified on semantic grounds. For alignment to hold, Kupiec et al. allow for minor modifications between sentences; full matches, partial matches and non-matches were possible.

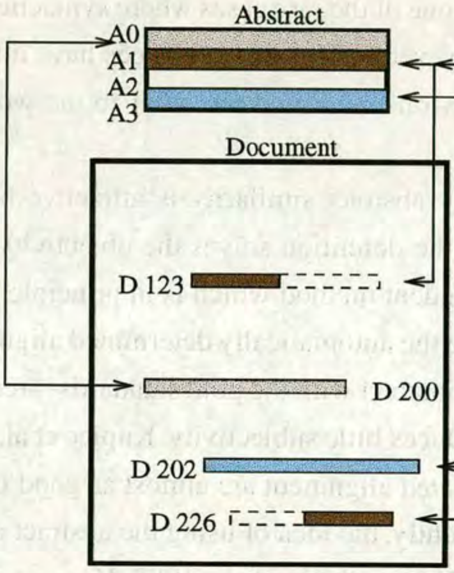


Figure 4.1: Target Extract by Alignment (Kupiec et al., 1995)

In Kupiec et al.'s corpus of 188 engineering articles plus summaries, 79% of the sentences in the summary could be aligned with sentences in the source text. In figure 4.1, for example, document sentences D-200 and D-202 align with abstract sentences A-0 and A-3, respectively. Parts of sentences D-123 and D-226 align with abstract sentence A-1, whereas abstract sentence A-2 does not have a corresponding sentence in the document. Examples for matches and non-matches from our corpus follow; they were obtained in a duplication of Kupiec et al.'s experiment (cf. section 5.3.4.1; also described in Teufel and Moens 1997).

> *Summary:* In understanding a reference, an agent determines his confidence in its adequacy as a means of identifying the referent. (A-3, 9405013)

> *Document:* An agent understands a reference once he is confident in the adequacy of its (inferred) plan as a means of identifying the referent. (S-131, 9405013)

The previous sentence pair illustrated a *match*, the following sentence pair a *non-match*:

> *Summary:* Recent studies in computational linguistics proposed computationally feasible methods for measuring word distance. (S-2, 9601007)

> *Document:* The paper proposes a computationally feasible method for measuring context-sensitive semantic distance between words. (A-0, 9601007)

The last example illustrates one of the rare cases where syntactic similarity does not mirror semantic similarity: however similar, the sentences have different propositional content, as one refers to previous work and the other to the work discussed in the source text itself.

Gold standard definition by abstract similarity is attractive because the machinery is technically simple, and the definition solves the objectivity dilemma: gold standards are defined by an independent method which is in principle outside the system developers' control. Correcting the automatically determined alignment—the only point where the system developers interact with the gold standards—requires relatively little human intervention and introduces little subjectivity. Kupiec et al. even argue that gold standards attained by *uncorrected* alignment are almost as good for system training as the corrected ones. Subsequently, the idea of using the abstract as gold standard has found a number of followers (Mani and Bloedorn, 1998; Hovy and Liu, 1998).

However, Kupiec et al.'s method introduces a dependency on the quality of the abstracts, and on the process of how they were generated. This is an issue with our texts. In our corpus, the abstracts were not written by professional abstractors but by the authors themselves.

While the literature on summarization techniques for professional abstractors is large (cf. section 2.1.1 and 2.3.1.2), there is not much research into how non-information specialists generate abstracts. However, it is indeed commonly assumed that author summaries are of a lower quality when compared to summaries by professional abstractors (Lancaster, 1998; Cremmins, 1996; Rowley, 1982). Rowley says about author abstracts that they are sometimes poorly written, that they often contain too much or too little data, and that there is often undue emphasis on author's priorities.

Borko and Bernier (1975) similarly caution that authors do not necessarily write the best abstracts for their papers, and Dillon et al. (1989) found empirically that journal-scanning readers often ignore author-written summaries if the full article is available too, and reject the summaries as "misleading" or "biased". We see several dangers with author summaries as gold standards for our task:

- We suspected that there is a less systematic relation between the information contained in the author-written summaries and the information contained in the documents. If it is not the case that the abstracts were created predominantly by selecting sentences, but if they were created from scratch, a surface-alignment procedure might provide too few gold standard sentences, and coverage would be too low, and indeed, this is the case in our corpus. Authors tend to reuse less of the document sentences, but *deep generate* new sentences from scratch.

- The papers in our collection come from different presentation styles, academic traditions and cover a wide range of subdomains. As a result, they differ in their internal document structure.

- They also differ in the structure of their abstracts. There is no guarantee that abstracts written by the authors keep to any kind of fixed rhetorical building plan, which abstracts produced by professional abstractors do (Liddy, 1991). Even though the information which ends up in the author abstracts is most certainly *relevant*, there are large individual differences of style and preference with respect to what *kind* of information an abstract contains, particularly if the authors of the abstracts were careless or biased. In a task such as ours it is essential that if there is information which is of comparable rhetorical status across papers, then the gold standard should mark this information similarly, independently of presentation form or where in the paper the information occurs. Comparability of information is hard to obtain with a surface-based method anyway, but if author decisions are taken to define the gold standard, comparability across papers decreases dramatically.

  Indeed, a later analysis (cf. section 4.4.1) reconfirms that the length and structure of our author abstracts vary considerably from paper to paper.

- Abstracts written by professional abstractors are typically self-contained, such that they can be understood without reference to the full paper. In many examples in our materials, this is not the case.

- Even worse, it is not even guaranteed that all the information contained in the abstract will also occur in the main document in *some* form. Writing advice states that the text and the abstract, apart from conveying the same semantics, should be viable texts which can be read on their own. But some of the authors in our collection assumed that the abstract would always be read before the main document, and in order to save time, they "abused" the abstract as an introduction. We found five papers in our collection where information in the abstract is not repeated anywhere else in the main document. Such cases are catastrophic for approaches which derive their gold standard from the abstract.

In early experiments with alignment (Teufel and Moens, 1997), we use a simple surface similarity measure which computes the longest common subsequence (LCS) of non-stop-list words. The results show a much lower alignment rate of 31% in our corpus, in comparison to Kupiec's 79%.

For example, consider the author summary of our example paper and the best-aligned sentences (figure 4.2).

Sentence **A-2** does not align with any document sentence, and alignments **A-1–113** and **A-3–147** were rejected by the human judge (us) as bad matches. The one acceptably aligned abstract sentence (**A-0**) is only partially aligned—with sentences **0** and **164**. Overall, the authors do not seem to have prepared the abstract by sentence extraction: all abstract sentences are at a higher level abstraction level than the corresponding document sentences, cf. the difference between **A-3** and **147**. It is immediately clear from the low level of alignment that this particular target extract cannot be a good representation of the document, even though the author abstract itself is.

Matters get even more complicated when we look at the rhetorical status of sentences, which is essential for Argumentative Zoning. For example, the rhetorical structure of the original abstract consisted of a sequence of Research goal (**A-0**), Solution applied (not invented by Pereira et al.; **A-1**), Further description of the solution (**A-2**), and Description of the evaluation (**A-3**). This summary is most similar in type to the summary for intellectual ancestry for uninformed readers, as discussed in section 2.3.3 (figure 2.20, p. 69). In comparison to the original abstract, the target extract is impoverished with respect to rhetorical structure; it consists of a very general statement about the task, and a statement that a solution was found—only 2 out of the 7 slot fillers available in the author abstract. Even though the aligned document sentences might be superficially similar to the ab-

Abstract sentences		Aligned document sentences	
A-0	We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts.	**0**	`(partial)` Methods for automatically classifying words according to their contexts of use have both scientific and practical interest.
		164	`(partial)` We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words.
A-1	Deterministic annealing is used to find lowest distortion sets of clusters.	**113**	`(bad match)` The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering <REF>Rose et al. 1990</REF>, in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter <EQN/> following an annealing schedule.
A-2	As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data.		—
A-3	Clusters are used as the basis for class models of word coocurrence, and the models evaluated with respect to held-out test data.	**147**	`(bad match)` For each critical value of <EQN/>, we show the relative entropy with respect to the asymmetric model based on <EQN/> of the training set (set train), of randomly selected held-out test set (set test), and of held-out data for a further 1000 nouns that were not clustered (set new).

Figure 4.2: Author Abstract and Target Extract by Alignment for Document 9408011

stract sentences, their rhetorical status is not necessarily similar to that of their aligned sentences. Without context, the rhetorical status of the document sentences cannot be detected anymore, and it is not even clear that it would be of help. Clearly, something got lost on the way.

To sum up, we do not deny that there are cases where abstract alignment can define good gold standards, and that Kupiec et al.'s experiment is probably one such case. However, the reason for this is not that abstracts *per se* provide good definitions of gold standards—rather, it is due to other fortunate circumstances, like the extensive training of professional abstractors and the high homogeneity with respect to paper and abstract structure in some data collections. In our case, alignment with abstracts would probably define a low-quality gold standard.

In general, surface comparability remains a problem for target extracts by abstract similarity. Document sentences which share propositional contents with an abstract sentence but which look different on the surface will not be contained in the gold standard, even though they should be; and a system which correctly determines such a sentence would be unduly punished by a target extract gold standard.

There is an additional problem with the static nature of the gold standard definition. The fact that the gold standard cannot be touched anymore might well make it more "objective", but the process by which the abstract was obtained (described in abstractors' guidelines) does not necessarily provide the specific information needed for a given task. For example, our task demands finding information about the goal of the paper, in relation to previous work: the determination of rival approaches and supporting previous research is essential. Unfortunately, this type information is not traditionally present in abstracts. Instead, this information might be hidden *anywhere* in the texts. An advantage of asking subjects to free-select sentences is that new criteria can be applied to the search as needed, in a *dynamic* way. As these criteria of selection are defined after the creation of the text, they can be changed according to task requirements.

One good point about target extracts in general, no matter by which method they are obtained, is that only a small number of sentences are selected, which are guaranteed to be globally important. This would be an advantage for a gold standard like ours. But most target extracts do not provide fall-back options, as they only make a binary distinction between relevant and non-relevant sentences.

4.1.3. Our Evaluation Strategy for Argumentative Zoning

We have argued that neither target extracts nor MUC-style answer keys can offer us high-quality gold standards for our task. But there is yet another field whose gold standards might be important for us.

Gold standards by total-coverage are traditionally in use in areas where the annotation in the text serves as a long-term resource itself, e.g. in dialogue act coding

(Carletta et al., 1997; Alexandersson et al., 1995; Jurafsky et al., 1997). Humans are asked to classify utterances in a corpus into a finite set of categories (called "dialogue moves"). For this kind of exercise, it is essential that different annotators performing the task independently (e.g. in different places) can create a resource that fits in with already existing resources generated according to the same annotation scheme. High reproducibility of such a scheme is thus important. Training of the annotators is extensive. Guidelines, also referred to as *coding handbooks*, are used to describe the task and the semantics of the categories. Specialized, standardized statistics, borrowed from the content analysis community (Carletta, 1996), exist for testing certain properties of the annotation scheme, most notably stability and reproducibility (cf. section 4.1.1).

Our evaluation strategy is as follows: we elicit judgements from subjects about the argumentative status of sentences in the source text according to our annotation scheme. Subjects perform *full-coverage* annotation, i.e., they give a judgement for *each* sentence in the paper.

We argue that this evaluation strategy will improve the gold standard situation with respect to surface comparability, fall-back options and comparability between papers. System evaluation is no longer a comparison of extracted sentences against a finite set of "good" sentences—this inevitably cannot work because there is not just *one* possible extract for a paper. Instead, *every* sentence in the source text which expresses the main goal will have been identified, and the system's performance is evaluated against *that* classification, providing an evaluation that portrays the real situation better.

We have, in chapter 3, made an implicit claim about the adequacy of our annotation scheme: that its categories provide an intuitive description of certain aspects of scientific texts. But the semantics of our slots are not as simple as the domain-specific MUC slots, which have the advantage that humans can confirm with high confidence if a slot filler is "right". In our scheme more subjective judgements are necessary. If we could prove a high degree of human agreement on the application of argumentative zones, this would also serve to verify our definition of the zones. Learnability of the scheme (and, as a result, reasonable reproducibility) is also important from a practical point of view as we want to use the gold standards as training material if they constitute a reliable resource.

The main difference between our task and other total-coverage annotations is that our task is a document retrieval task, and as a result, relevance *is* an issue for us. Certain items are more important for us than others, and certain errors are more grave than others. We care most about reproducibility in those zones which are particularly important for our task (e.g. AIM zones); we care less about errors in the frequent zones as these sentences are not directly extracted and displayed in RDPs.

Our gold standard should give us sentences which are the *best* slot fillers for each category; it should also define fall-back options. However, total-coverage classification does not readily provide different degrees of relevance. It gives us many "equally relevant" sentences per category, whereas the other gold standards would have given us few "relevant" sentences. In an independent step, the most appropriate slot fillers would have to be determined:

1. Subjects could tell us which sentences are the best fillers, e.g. by ranking their prior classifications.

2. Some external criterion could define relevance independently of the human classification; e.g. sentences alignable with abstract sentences or occurring in the periphery of the paper could be considered "more relevant". The connection between location and the quality of gold standards is explored in section 4.4.2.

3. Slot fillers which are similar to each other could be defined to be more relevant. This approach was suggested in section 2.3.3 where we sketched the generation of tailored summaries from RDP slot fillers.

Apart from not being able to give us the most appropriate slot fillers, total-coverage classification gold standards provide a well-suited evaluation for our task, as such classification is a simple, well-understood cognitive task with a widely accepted evaluation metrics. However, it is a time-consuming task—we consider different ways of reducing the effort, either by reducing the training (cf. section 4.3.2) or by reducing the areas to be annotated (cf. section 4.4.2). Our new gold standard helps us get around some of the problems that other evaluation strategies have:

- *Objectivity:* The new gold standard measures objectivity in terms of stability and reproducibility, i.e. in how far humans will agree on the task (results are reported in section 4.3). One could, however, argue that a static, fixed, independent standard as in Kupiec et al.'s work is intrinsically more objective.

- *Task-flexibility:* Instructions to the annotators can be adjusted according to the requirements of the task.

- *Comparability between papers*: The new gold standard guarantees comparability because *all* sentences are classified. Coverage of all categories should be high, i.e. there should always be enough candidates for each category. As a result, a sensible comparison of information between papers is possible, unlike in the abstract-as-gold-standard strategy.

- *Fall-back options:* The new gold standard provides fall-back options for each category (provided the category was present in the paper), unlike other methods.

- *Best fillers:* The new gold standard still gives too many fillers per category, all of which are judged equally-relevant, in contrast to selection methods. In order to determine the most relevant fillers in our case, an independent measure of relevance is needed.

- *Surface comparability:* The new gold standard has fewer problems with surface comparability than target extracts or answer keys. This is due to the fact that judgements for each sentence are compared.

4.2. Evaluation Measures

In the following experiments, we are particularly interested in two properties of our annotation scheme: Firstly, stability, i.e. the extent to which one annotator will produce the same classifications at different times (Krippendorff, 1980). Stability is important, because in unstable annotation schemes the definition of the categories is not even consistent within one annotator's private understandings, and as a result, such schemes are very unreliable. High stability shows at the very least that there must be *some* consistent definition of semantics in the gold standard, even if we do not know yet if this definition can be communicated to others. The second property is reproducibility, i.e. the extent to which different annotators will produce the same classifications, which measures the consistency of shared understandings (or meaning) held by more than one annotator. As consistent *shared* understandings require consistent *private* understandings, an unstable annotation can never be reproducible; conversely, it is commonly assumed that a proof of the reproducibility of a scheme implies its stability. Thus, many experimentators only measure and report reproducibility (cf. the MUC enterprise, section 2.2.2).

We feel that stability is independently important, and that stability and reproducibility have completely different consequences with respect to our task. Researchers in document retrieval have argued that although stability is important to some degree, if one is interested in user satisfaction, then reproducibility is of little importance. If there are two or more intuitively "good" but different gold standards, two judges might disagree over which one to choose, resulting in a low reproducibility. However, both of these gold standards might have satisfied the user. We subscribe to the argument of theoretical priority of stability over reproducibility in document retrieval, but at the end of the day, only extrinsic evaluation can prove or disprove if the argument is valid.

A related question is how exactly we should establish an *upper bound* for the task. An upper bound is the best measurement that an automatic performance can *theoretically* reach. When humans systematically do not agree beyond a certain degree, this degree must be accepted as the upper bound: it makes no sense to think of a machine as performing better than this level of agreement. We argue that reproducibility constitutes a good upper bound. That is, if the performance stays the same if an automatic approach is added to a pool of independently annotating human annotators, then this approach has reached the theoretical best performance possible.

In many related tasks, definitions of upper bounds are handled less strictly. Kilgarriff (1999), for example, reports an upper bound for word sense disambiguation which is numerically very high. This gold standard was gained by negotiation between the annotators, as is common in lexicography. We also believe that interaction between annotators is important, in order to arrive at a shared understanding of the categories. However, experience has shown that it is often the annotator with the strongest personality which convinces the other annotators of the validity of her annotation.

Another form of improving "reproducibility" would be to ask annotators to *correct* somebody else's output—in other tasks like manual parts-of-speech (POS) assignment, annotators have been shown to agree much more if they do not perform the task from scratch.

However, as we are interested in the properties of the cognitive task, we measure reliability of independent annotation *before* discussions. The real keepers of the semantics of the categories should always be the guidelines. The guidelines for annotation tasks should be written before the experiment and changed as little as possible during the experiment. However, as annotation experiments are long and expensive enterprises, it might be difficult to repeat an experiment after each change (and ideally with new annotators). We had to change the guidelines several times (e.g., the example annotations in figures on p. 327 and 328 were added after those papers had been annotated independently).

Our annotation task is mutually exclusive categorial assignment. There have been different ways in the past to evaluate agreement between humans for such task (cf. the overview in Carletta 1996), using either majority opinion or percentage agreement as measurement. We are opposed to using majority opinion: the average does not reflect anybody's understanding of the categories. We want to treat all our annotator's opinions as a valid judgement. None of these is by definition wrong or right—we are dealing with a difficult "high-level" task, where a certain level of subjective disagreements can be expected.

We use the Kappa coefficient K (Siegel and Castellan, 1988) to measure stability and reproducibility among k annotators on N items (here: sentences). For our task, Kappa has the following advantages:

- It factors out random agreement.
- It allows for comparisons between arbitrary numbers of annotators and items.
- It treats less frequent categories as more important.

The Kappa coefficient controls agreement $P(A)$ for agreement by chance $P(E)$:

$$K = \frac{P(A)-P(E)}{1-P(E)}$$

No matter how many items or annotators, or how the categories are distributed, $K = 0$ when there is no agreement other than what would be expected by chance, and $K = 1$ when agreement is perfect. If two annotators agree *less* than expected by chance, Kappa can also be negative. Chance agreement is defined as the level of agreement which would be reached by random annotation using the same distribution of categories as the real annotators. Kappa is stricter than percentage agreement: its value is always lower or equal to percentage agreement P(A); it is equal in the case of a uniform distribution and lower for skewed distributions. We already know that our category distribution will most likely be very skewed, for example because the category OWN is so predominant. The fact that Kappa is a more sensible measurement for our task than percentage overlap can be easily shown with the following argument about baselines for our task. (This argument anticipates some numerical values which we will obtain later on in this chapter.)

Choosing the most frequent category Own is one possible baseline for our task (Baseline 1). Figure 4.3 shows that percentage agreement makes this baseline look like a good one at 69%, in comparison to human agreement at only 87%. However, if Kappa is used to measure the similarity of this baseline with the annotation of a human annotator, it reveals a negative (K=–.12)—compared to chance agreement, the baseline performs *worse* than random. This agrees with our intuition that always choosing the most frequent category is a bad strategy for our task. For our task it is important to choose the rare categories Aim, Textual, Contrast and Basis.

Baseline	Kappa	P(A)	P(E)
Baseline 1: Most frequent category	-.12	68%	71%
Baseline 2: Random, uniform distribution	-.10	14%	22%
Baseline 3: Random, observed distribution	0	48%	48%

Figure 4.3: Baselines for the Task of Argumentative Zoning

We implemented a random generator, assigning categories based either on a uniform distribution (Baseline 2) or the observed distribution (Baseline 3). Baseline 2 has a slightly better chance agreement; it achieves K=–.10 if compared to the human annotator. The hardest-to-beat baseline is random choice according to the observed distribution of categories (Baseline 3). Kappa for this baseline should theoretically be K=0 which is reconfirmed by our data. Kappa agrees with our intuition that Baseline 3 is better than Baseline 1 whereas the numerical values of percentage agreement contradict our intuition.

Kappa is designed to abstract over the number of annotators as its formula relies on *pairwise* agreement. That is, K for $k = 6$ annotators will be an average of the values of K for $k = m$ where $m < k$, taking all possible m-tuples of annotators from the annotator pool. This property makes it possible to compare between different numbers of annotators, and between groups of annotators and versions of our system. A look at Rath et al.'s awkward way of reporting agreement for different annotator pools (cf. p. 132) makes clear that numerical comparability is a big advantage.

We are also looking for a measurement which will punish disagreement on the rare (= important) categories more than disagreement in the more frequent categories. As a side effect of taking random agreement into account, Kappa treats agreement in a rare category as more surprising, and rewards such agreement more than an agreement in a frequent category.

There are different scales of how to interpret Kappa values. Krippendorff (1980) starts from the assumption that there are *two* independently annotated variables which show a clear correlation. If the agreement of an annotation of one of these is so high that it reaches a value of K=.8 or above on a reasonably-sized dataset, then the correlation between these two variables can be shown with a statistical significance of $p \leq 0.05$. That is, the annotation contains enough signal to be found among the noise of disagreement. If agreement is in a range of $.67 \leq K < .8$, the correlation can be shown with a (marginal) statistical significance of p=0.06, which allows for tentative conclusions to be drawn. Krippendorff's strict scale considers annotations with $K < .67$ as unreliable. More forgiving scales take into account that most practical annotation schemes only mark one dependent variable and assume that K=.6 is still reasonable agreement. However, Krippendorff (1980, p. 147) describes an annotation experiment performed by Brouwer et al. (1969) in which annotators achieved K=.44 with an annotation scheme whose categories were described only by complicated Dutch names with no resemblance to English words. This is disturbing, because Kappa *should* have been zero, due to the lack of semantics attached to the categories (as the annotators did not understand Dutch): any agreement achieved in that experiment can be only considered as chance. Having said this, it is so difficult to achieve *high* Kappa values that one can nevertheless exclude chance in those cases—Kappa is in general accepted in the field as a sensible and rigorous measure.

Whereas researchers using Kappa frequently have developed some intuitions about whether or not not two Kappa values probably are statistically significantly different or not, there still is no statistical formula to calculate if this is the case or not. This is a disadvantage of using Kappa, but we think it is out-weighed by its advantages.

We use our own implementation of Kappa which allows us to vary annotation areas (cf. section 4.4.2), calculate values for single files, subsets of annotators in the pool and to show confusion matrices for pairs of annotators.

4.3. Reliability Studies

4.3.1. Experimental Design

We conducted three studies. The first two, studies I and II, were designed to find out if two versions of our annotation scheme can be learned by human annotators with a significant amount of training. The first version is the *basic* annotation scheme which encodes intellectual ownership (cf. section 3.3). The second version is the *full* annotation scheme with seven (more complicated) categories. A positive outcome of studies I and II would convince us that the human-annotated training material constitutes a good gold standard, and that it can be used for both training and evaluation of our automatic method in chapter 5. The outcome of study II is crucial to the task, as it deals with the full annotation scheme. Some of the categories specific to the full annotation scheme (AIM, TEXTUAL, BASIS and CONTRAST) provide essential information for RDPs.

Study III tries to answer the question if the considerable training effort used in studies I and II can be reduced. If this were the case, i.e. if annotators with no significant task-specific training could produce similar results to highly trained annotators, the training material could be acquired in a more cost and time effective way. A positive outcome of study III would also substantiate claims about the immediate intuitivity of the category definitions.

4.3.2. Study I

4.3.2.1. Method

Subjects: Three annotators participated in this study: Annotator A holds a Master degree in Cognitive Science and Annotator B was a student of Speech Therapy at Queen Margaret's College, Edinburgh. Annotator C is the author of this thesis. The annotators can be considered skilled at extracting information from scientific papers but they are not experts in all of the subdomains of the papers they annotated. Annotator A has some overview knowledge in most of the subfields represented in the corpus; in particular, he is well accustomed to articles in computer science, which Annotator B was not. Annotator B had some knowledge in phonology and phonetics, and to a lesser degree in theoretical linguistics. Annotators A and B were paid for their work at the standard academic student rate of the University of Edinburgh.

Materials: The materials consist of 26 computational linguistics papers from our collection (cf. appendix A.2 for the overall list of articles in our corpus). Figure 4.4 lists the materials used in this study: the papers and their numbers of sentences (abstract sentences and document sentences, but excluding sentences occurring under the heading *Acknowledgements*). We used the first four articles of our collection (papers 0 – 3) for training, and the next 22 papers (papers 4 – 25) for annotation by all three annotators. As we wanted to cover as much variety as possible in writing style, we decided to only include one paper by each first author in each study—subsequent papers by the same authors were discarded. In study I, no paper was excluded on the grounds of authorship, however. During the annotation phase, one of the papers (paper 18) turned out to be a review paper. This paper caused the annotators difficulty as the scheme was not intended to cover reviews. Thus, we discarded this paper from the analysis. For the stability figures (intra-annotator agreement), 5 papers were randomly chosen out of the set of 21 papers.

Type of Material	Paper numbers	Sent.
Training material	0, 1, 2, 3	532
Annotation material	4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 23, 24, 25	3643
Intra-annotator material	0, 7, 10, 23, 24	1115

Figure 4.4: Study I: Materials

Procedure: The training procedure was as follows: the annotators read our written instructions which define the categories of the *basic* version of the annotation scheme in detail (7 pages; reproduced in appendix C.1). For the reader's convenience, figure 4.5 repeats the categories of the basic annotation scheme.

BACKGROUND	Generally accepted background knowledge
OTHER	Specific other work
OWN	Own work: method, results, future work...

Figure 4.5: Study I: Overview of Basic Annotation Scheme

After reading the guidelines, the annotators marked up the first two training papers, followed by a discussion, then the other two training papers, followed by another discussion. In these discussions, we tried to settle disagreements in the annotators' judgements and change unclear passages in the instructions.

The annotation procedure itself was as follows: Annotators marked up the 21 papers, 5–6 papers per week, in the same order. There was no communication between

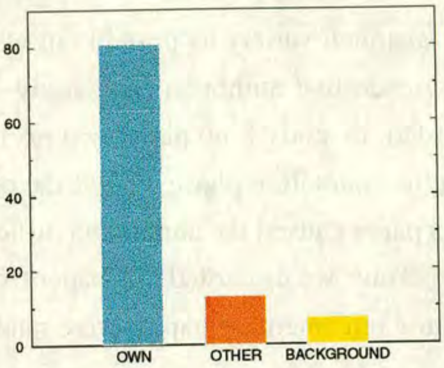


Figure 4.6: Study I: Overall Frequency of Categories

the annotators during the annotation. Annotation included the abstracts as well as all sentences in the document (excluding acknowledgement sentences). Reading and annotating a paper took the annotators 20–30 minutes on average. Weekly discussions between the three annotators took place during the annotation phase. The rationale of the discussions was to increase *future* agreement by clarify unclear passages in the guidelines in the light of unclear annotation cases. However, agreement was measured *before* discussions. As there was no time to implement a specific annotation tool, all annotation reported here was done pencil-on-paper and then edited into an XML version of the documents.

6 weeks after the end of the first annotation phase, stability was measured by an intra-annotator experiment, where annotators were asked to re-annotate randomly chosen papers.

We collected informal comments from our annotators about how natural the task felt, but did not conduct a formal evaluation of subjective perception of the difficulty of the task. Instead, our analysis concentrates on trends in the data as the main information source.

4.3.2.2. Results and Discussion

The results show that the basic annotation scheme is stable (K=.83, .79, .81; N=1115; k=2 for all three annotators) and reproducible (K=.78, N=3643, k=3). This reconfirms that trained annotators are capable of making the basic distinction between own work, specific other work, and general background. To our knowledge, this study is the first to research attribution of intellectual ownership empirically on a corpus.

Categories		Kappa
OWN + OTHER 93.2%	BACKGROUND 6.8%	.58
OWN 80.4 %	OTHER + BACKGROUND 19.6%	.83
OWN + BACKGROUND 87.2%	OTHER 12.8%	.77

Figure 4.7: Study I: Krippendorff's Diagnostics for Category Distinction

Figure 4.6 shows that the distribution is very skewed, as predicted. The relative frequency of the three categories is 80.4% (OWN), 12.8% (OTHER) and 6.8% (BACKGROUND).

Though the reliability values are acceptable, there are some questions that are typically asked in order to improve an annotation scheme:

- Do all annotators perform equally well?
- Are there particular category distinctions that are hard to make?
- Is there a difference between clusters of items (papers)?

The first question is answered easily—the variation between annotators is fairly small. The results for pairwise comparison are K=.74 (A, B), K=.78 (B, C) and K=.82 (A, C). It is important that the results do not change dramatically when the developer of the annotation scheme (Annotator C) is left out of the annotator pool. In this case, they drop a little from K=.78 to .74. This still suggests that the training conveyed the intentions of the developer of the annotation scheme fairly well.

In order to see which category distinctions are hard to make, we use Krippendorff's diagnostic for category distinctions: all other categories but the one(s) of interest are collapsed. The most difficult single distinction is the one that results in the *best* reproducibility values if omitted. In our case, this most difficult distinction is the one between OTHER and BACKGROUND. We are not surprised about this: the distinction between other general work and other specific work concerns only the degree of specificity. Swales (1990) reports similar difficulties with a distinction between his two related moves 1.2 (making topic generalizations; background knowledge) and 1.3 (reviewing previous research). There might not be an easy way to avoid this difficulty; it seems to be part and parcel of the task.

Figure 4.8 shows that the variation in reproducibility across items (papers) is large: there are some papers that are annotated very consistently, and others that are not.

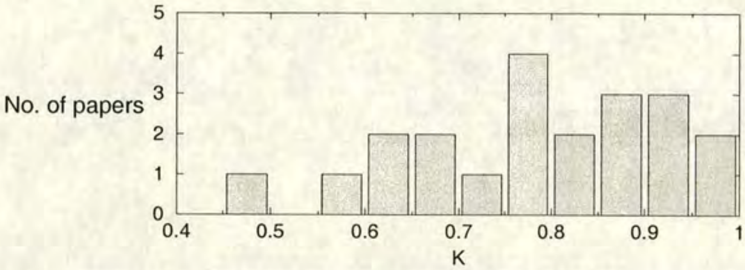


Figure 4.8: Study I: Distribution of Reproducibility Values

We tried to diagnose the reasons for the low reproducibility of some papers. We have several hypotheses of what could be responsible for this:

1. One frequent problem our annotators reported was a difficulty in distinguishing OTHER work from OWN work, due to the fact that some authors did not express a clear distinction between *previous* own work (which, according to our instructions, had to be annotated as OTHER) and *current, new* work. Our annotators reported that in some papers there are long sections that cannot be obviously attributed to either *previous* or *current* work because the authors did not make the distinction clear. This was particularly the case where authors had published several papers about different aspects of one piece of research (cf. the idea of "smallest publishable unit", section 3.2.3).

   We suspected that the effect of mixing descriptions of own and previous research could be gauged by the *self citation ratio*, i.e. the ratio of self citations to all citations in running text. 5 papers contain no self citations and were thus put into one group. We divided the remaining papers into two equally sized groups, one with a high and one with a low self citation ratio (the borderline turned out to be at 18% of all citations).

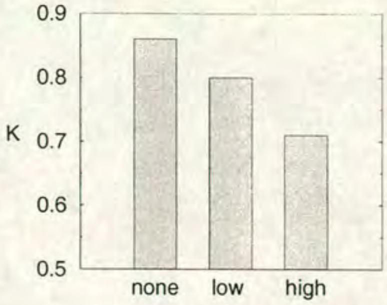


Figure 4.9: Study I: Effect of Self-Citation Ratio on Reproducibility

Figure 4.9 confirms that papers who quote previous own work only rarely or not at all seem to be annotated most consistently in our scheme. Subsequent analysis shows that part of this effect can indeed be attributed to a difficulty in distinguishing the categories OWN and OTHER. In the groups with no self citations or a low self citation ratio, we found that reproducibility does not increase too much (from K=.86 to K=.90 and from K=.8 to K=.83) if OWN and OTHER are collapsed, indicating that this distinction is not too difficult. In the high self citation group, the reproducibility increase was much higher (from K=.71 to K=.85), indicating that the distinction is more difficult in this group. This might be due to the fact that papers in the first group (and to a certain degree, in the second group) are structured in a simpler way, i.e., they might report on some isolated piece of research. However, there might be other reasons why the own new work is well-distinguished from other and own previous work in these cases.

2. There is also a difference in reproducibility between papers from different *conference types*. Out of our 21 papers, 4 were presented in student sessions, 4 came from workshops and the remaining 13 were main conference papers. Figure 4.10 shows that student session papers are the easiest to annotate, which might be due to the fact that they are shorter and have a simpler structure, with fewer mentions of previous research. Main conference papers dedicate more space to describing and criticizing other people's work than student or workshop papers (on average about one fourth of the paper). They seem to be more carefully prepared than workshop papers (and thus easy to annotate); conference authors must express themselves more clearly because they are reporting

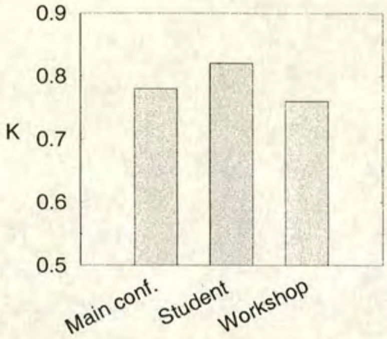


Figure 4.10: Study I: Effect of Conference Type on Reproducibility

finished work to a wider audience.

3. Another persistent problem in some papers was the distinction between OWN and BACKGROUND. This could be a sign that the authors of these papers aimed their writing at an expert audience, and thus thought it unnecessary to signal clearly which statements are commonly agreed in the field, as opposed to their own new claims. If a paper is written in such a way, its understanding requires a considerable amount of domain knowledge, which our annotators did not necessarily have. The problem here seems to be the same that Manning (1990) reports for human abstractors: the production of informative abstracts is difficult, because one needs to contrast the findings of the text with the already-established findings in the field. The recognition of the scientific contribution of a given paper requires a lot of domain knowledge in the field, particularly if it is not signalled well in the paper.

4.3.3. Study II

The only difference introduced in study II is the use of the full annotation scheme instead of the basic one.

4.3.3.1. Method

Subjects: The same annotators as in study I participated in this study.

Materials: In principle, the materials for study II were similar to the materials in study I (cf. figure 4.11). They consisted of 30 chronologically adjacent papers (papers 38–67). Papers were excluded if the first author was already represented in the materials for the given study (this was the case for papers 54, 55, 57). 5 papers were chosen as training material (papers 38, 39, 50, 51, 62). During the annotation phase, another paper turned out to be a review paper; as before, we discarded this paper from the analysis. And finally, in order to compare the performance of the tasked-untrained annotators to be used in study III to our task-trained annotators, we needed their judgement on the materials chosen for study III (papers 4 and 14). This resulted in 23 papers for annotation. For the stability experiment, we randomly chose 7 papers out of these 23.

Type of Material	Paper numbers	Sent.
Training material	38, 39, 50, 51, 62	784
Annotation material	4, 14, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 52, 56, 58, 59, 60, 61, 63, 64, 65, 66, 67	3449
Intra-annotator material	14, 41, 43, 44, 52, 58, 65	1091

Figure 4.11: Study II: Materials

Procedure: Training and annotation procedure was as in study I, except that the annotators were asked to annotate with the full annotation scheme, repeated in figure 4.12. Again, annotators were asked to annotate abstracts as well as all sentences in the document, but not acknowledgement sentences.

BACKGROUND	Generally accepted background knowledge
OTHER	Specific other work
OWN	Own work: method, results, future work...
AIM	Specific research goal
TEXTUAL	Textual section structure
CONTRAST	Contrast, comparison, weaknesses of other solution
BASIS	Other work provides basis for own work

Figure 4.12: Study II: Overview of Full Annotation Scheme

The written instructions for that scheme are reproduced in appendix C.2; they are 20 pages long. As the main decision criterion, they contain the decision tree discussed in section 3.3 (figure 3.21; p. 110). No special instructions about the use of cue phrases were given, although some of the example sentences given in the guidelines contained cue phrases.

The annotators already knew three of the seven categories from study I, and this might might have sped up the learning process with respect to completely untrained annotators; however, as there was a gap of several weeks between the two experiments, it is unlikely that this advantage was substantial.

4.3.3.2. Results and Discussion

The annotation scheme is stable (K=.82, .81, .76 for all three annotators; N=1091, k=2) and reproducible (K=.71, N=3449, k=3). Because of the increased cognitive difficulty of the task in comparison to study I, the decrease in stability and reproducibility is acceptable. Annotation between annotators varies only minimally: K=.70 (A, B); K=.70 (A, C) and K=.72 (B, C).

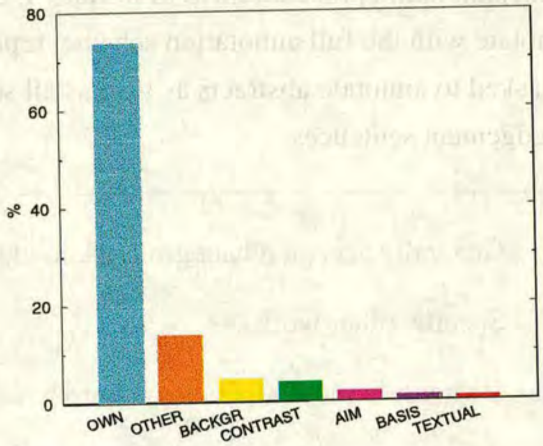


Figure 4.13: Study II: Overall Frequency of Categories

Figure 4.13 shows the relative frequencies of all seven categories. The transition between the basic categories Own, Other and Background on the one hand, and the "non-basic" categories Aim, Textual, Contrast and Basis on the other is not as pronounced as we expected.

Again, variability in reproducibility is large (cf. figure 4.14), as it was in study I. Even more so than in study I, there seems to be a bimodal distribution: there is a cluster of papers with high reproducibility (K in the range of .85), and another cluster of papers with medium reproducibility (K in the range of .6). Similar explanations for this divergence as in study I are true here too: confusion between current and own previous work can be measured by self-citation ratio (cf. figure 4.15), and conference type is a predictor of overall reproducibility (cf. figure 4.16).

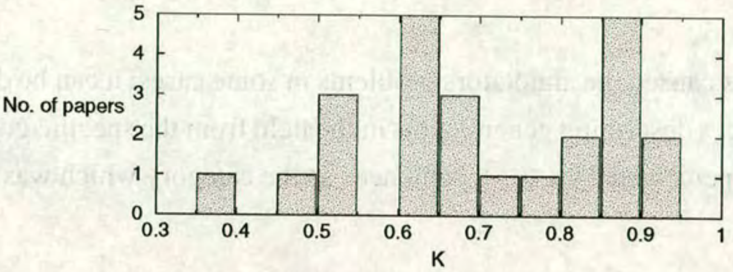


Figure 4.14: Study II: Distribution of Reproducibility Values

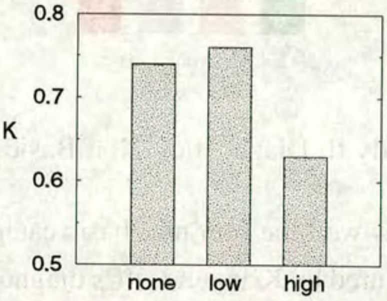


Figure 4.15: Study II: Effect of Self-citation Ratio on Reproducibility

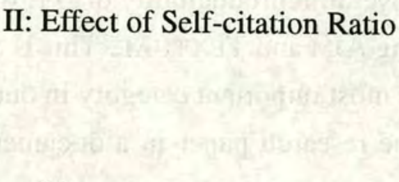


Figure 4.16: Study II: Effect of Conference Type on Reproducibility

There are problems which are specific to the new categories: annotators sometimes find it hard to distinguish neutral descriptions of other work (OTHER) from descriptions of other work which express author stance (CONTRAST and BASIS). Often, contrastive stance was not expressed openly (cf. MacRoberts and MacRoberts's (1984) explanation for this phenomenon in section 3.2.2); in order to decide if a sentence was of category BASIS, annotators needed to interpret possible reasons for the positive evaluation of other work.

AIM sentences caused the annotators problems in some cases; it can be difficult distinguishing sentences describing general aims in the field from the specific goals of a paper. All annotators perceived TEXTUAL sentences as the category which was easiest to annotate.

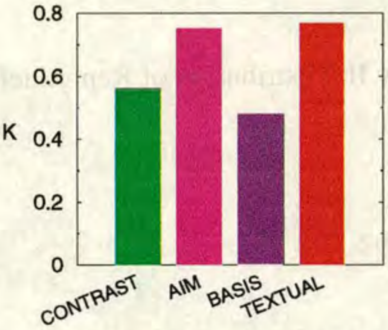


Figure 4.17: Study II: Diagnostics, Non-Basic Categories

Figure 4.17 reports how well the four non-basic categories were distinguished from all other categories, measured by Krippendorff's diagnostics for category distinctions. When compared to the overall reproducibility of .71, we notice that the annotators were good at distinguishing AIM and TEXTUAL. This is an important result: AIM sentences constitute the single most important category in our scheme as they provide the best characterization of the research paper in a document retrieval context. Annotation performance on AIM sentences can be compared to results of free-selecting experiments where subjects were asked to identify "most relevant" sentences from a paper; traditionally, low agreement is reported for such tasks (Rath et al., 1961).

The annotators were less good at determining BASIS and CONTRAST. In section 3.2.5, we saw that there is large variation in the syntactic realization of meta-discourse signalling categories such as BASIS and CONTRAST, which makes it harder to find them. Anther reason might have to do with the location of those types of sentences in the paper: whereas AIM and TEXTUAL are usually found at the beginning or end of the introduction section, CONTRAST, and even more so BASIS, are usually interspersed within longer stretches of OWN. As a result, BASIS and CONTRAST are more exposed to lapses of attention during annotation.

If high reliability was our priority, the annotation scheme could be simplified by creating a new category which collapses CONTRAST, OTHER and BACKGROUND. This would cause the reproducibility of the scheme to increase to K=.75. Structuring our training set in this way seems to be an acceptable compromise for our task as such a scheme would maintain most of the distinctions contained in the basic annotation scheme, while also categorizing AIM, TEXTUAL and BASIS sentences.

Figure 4.18 shows the confusion matrix between two annotators. The diagonal shows the decisions in which they agree, all other cells show decisions where they disagree. The confusion matrix is another tool apart from Krippendorff's diagnostics for detecting weaknesses in annotation schemes. One can see that the only category that AIM sentences are confused with are OWN sentences—what both categories have in common is that they describe own work. The decision of whether or not to assign an AIM label to such a sentence is a type of relevance judgement. CONTRAST sentences are often confused with OWN sentences. This is natural, as contrast sentences often compare own and other work: annotators have to judge which aspect (own or other) is more dominant, which can be hard in some cases. BACKGROUND sentences are confused with OTHER and OWN sentences, as discussed above; we suspect that the confusion with CONTRAST sentences occurs when a failure of some general method in the field is discussed. Confusion between OTHER and CONTRAST is often due to different judgement of author stance vs. neutrality expressed in the sentences. BASIS sentences are most likely to be confused with either OTHER sentences (author stance vs. neutrality), or with OWN sentences, when the annotators disagree as to if an aspect of the own work has been contributed by prior work or is first described in the current article. Appendices B.5 and B.6 show the example paper annotated by Annotators A and B; the previously shown figure 3.23 (p. 113) actually gives Annotator C's annotation of the example paper.

Figure 4.19 shows how well one annotator can predict another annotators' choice of non-basic categories. Taking Annotator B's decisions of a certain category as gold standard, recall reports how many of those instances Annotator C found, and precision reports how many of the instances that Annotator C categorized as that category, really turn out to be of that category (by Annotator B's judgement). That is, precision measures how confident we can be with the result set, whose size is measured by recall.

Annotator C achieves a precision and recall of almost 80% on TEXTUAL sentences, and 72% precision and 56% recall for AIM sentences. These values are much higher than similar values reported in earlier results for overall relevance (Rath et al., 1961). We believe that our task, given detailed guidelines, is indeed easier and better delineated than the direct determination of globally relevant sentences.

		Annotator B							
		AIM	CTR	TXT	OWN	BKG	BAS	OTH	Total
Annotator C	AIM	35	2	1	19	3		2	62
	CTR		86		31	16		23	156
	TXT			31	7			1	39
	OWN	10	62	5	2298	25	3	84	2487
	BKG		5		13	115		20	153
	BAS	2			18	1	18	14	53
	OTH	1	18	2	55	10	1	412	499
	Total	48	173	39	2441	170	22	556	3449

Figure 4.18: Study II: Confusion Matrix between Annotators B and C

	AIM	CTR	TXT	OWN	BKG	BAS	OTH
Precision	72%	50%	79%	94%	68%	82%	74%
Recall	56%	55%	79%	92%	75%	34%	83%

Figure 4.19: Study II: C's Precision and Recall per Category if B is Gold Standard

4.3.4. Study III

Study III uses a different subject pool than studies I and II. The annotators used here are not acquainted with our scheme; they are only given some general descriptions about the semantics of the categories.

4.3.4.1. Method

Subjects: 18 subjects with no prior annotation training were chosen for the second experiment. All of them have a graduate degree in Cognitive Science, with two exceptions: one was a graduate student in Sociology of Science, and one holds a master degree in English and Spanish Literature. It can be assumed that all the subjects are used to reading scientific articles, in the course of their daily work or studies, though the non-Cognitive Scientists might have come across less technical articles.

Materials: We randomly chose three papers (papers 4, 14 and 52) out of the pool of those papers for which our trained annotators had previously achieved good agreement in study I or in study II (at least K=.65). The reasoning behind this was that the task seemed cognitively difficult considering the lack of training, so we wanted to give our annotators less controversial materials. One of the three papers (paper 14) had previously resulted in much lower reproducibility (K=.67,N=205) than the other two (K=.85,N=192 for paper 4; K=.87,N=144 for paper 52).

Procedure: Each annotator was randomly assigned to a group of six, all of whom independently annotated the same single paper: group I annotated paper 4, group II paper 14 and group III paper 52. Subjects were given minimal instructions (1 page; appendix C.3), and the decision tree in figure 3.21 (p. 110).

4.3.4.2. Results and Discussion

The results show that reproducibility varies considerably between groups (K=.49, N=192, k=6 for group I; K=.35, N=205, k=6 for group II; K=.72, N=144, k=6 for group III). As Kappa is designed to abstract over the number of annotators, lower reliability in study III as compared to studies I and II is not an artifact of how K was calculated.

We must conclude that our very short instructions did not provide enough information for consistent annotation; some subjects in groups I and II did not under-

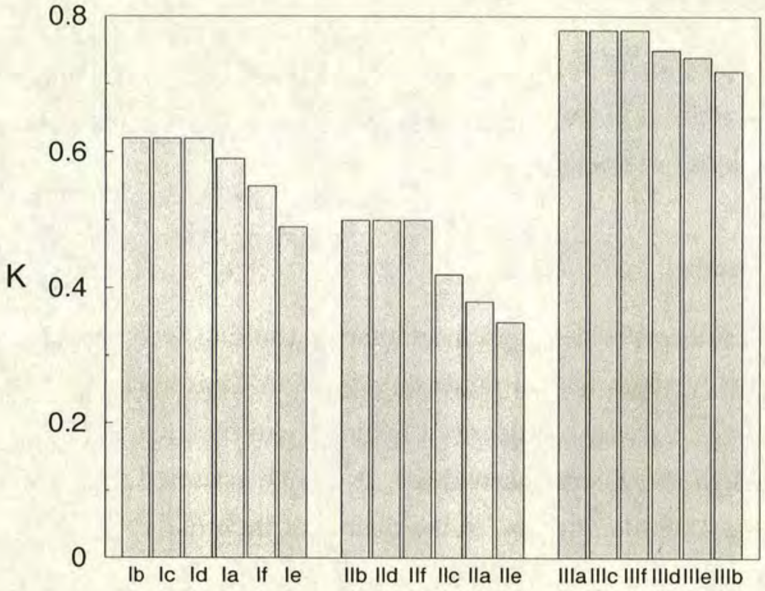


Figure 4.20: Study III: Reproducibility per Group and per Subject

stand the instructions as intended. Part of the low reproducibility results in group I and group II was due to a misunderstanding at a very superficial level. Many subjects misinterpreted the semantics of the TEXTUAL category as including sentences that refer to figures and tables in the text. This misunderstanding is easily rectifiable for future experiments, but still decreased the reliability values in this experiment considerably.

Part of the low reproducibility result can be attributed to the papers themselves: group III, which annotated the paper found to be most reproducible in study II, performed almost as well as trained annotators; group II, which performed worst, also happened to have the paper with the lowest prior reproducibility.

Figure 4.20 shows reproducibility for the most similar three annotators in each group, successively adding the next similar annotator to the pool. We can see that the performance between subjects varies much more in groups I and II than in group III, where all annotators performed more or less similarly well. Within each group, there is a subgroup of "more similar" annotators. In groups I and II, the most similar three annotators reached a respectable reproducibility (K=.63, N=192, k=3 for group I; K=.5, N=205, k=3 for group II). This result, in combination with the good performance of group III, seems to point to the fact that the annotators did have at least *some* shared understanding of the meaning of the categories.

The two subjects in study III who had no training in computational linguistics (subjects Ia and IIa) performed reasonably well: although they were not part of the circle of the most similar three subjects in their groups, their annotation also was not the odd one out.

4.3.5. Significance of Reliability Results

The reproducibility and stability values for Argumentative Zoning measured in these studies do not quite reach the levels found for, for instance, the best dialogue act coding schemes (around K=.80). Our annotation requires more subjective judgements and is possibly more cognitively complex. The reproducibility and stability results achieved with trained annotators are in the range which Krippendorff (1980) describes as giving marginally significant results if two coded variables were correlated. Of course, our requirements are rather less stringent than Krippendorff's because our annotation involves only *one* variable. On the other hand, annotation is expensive enough that simply building larger data sets is not an attractive option. Overall, we find the level of agreement which we achieved acceptable.

The single most surprising result of the experiments is the large variation in reproducibility between papers. Intuitively, the reason for this are qualitative differences in individual writing style—annotators reported that some papers are better structured and better written than others, and that some authors tend to write more clearly than others. It would be interesting to compare our reproducibility results to independent quality judgements of the papers, in order to determine if our experiments can indeed measure the clarity of scientific argumentation.

We are particularly interested in the question if shallow (human and automatic) information extraction methods, i.e. those using no domain knowledge, can be successful in a task such as Argumentative Zoning. The experiments reported in this chapter were in part conducted to establish an *upper bound* for the automatic simulation of the task. We believe that argumentative structure has enough reliable linguistic or non-linguistic correlates on the surface—physical layout being one of these correlates, along with linguistic indicators like *"to our knowledge"* and the relative order of the individual argumentative moves. The fact that the two non-computational linguists in the subject pool performed reasonably well is remarkable as the strategy that they must have used for Argumentative Zoning could not have included any domain knowledge. This result fits in nicely with the reasoning behind our approach: the implementation of Argumentative Zoning introduced in the next chapter is based on our belief that it should be possible to detect the line of argumentation of a text in a shallow, robust way.

In the framework of constructing practical gold standards for our task, the results of study II are positive as they tell us that training material gained by our method of human annotation is in principle reliable. With respect to a reduction of the effort for producing the gold standards, the outcome of study III was disappointing, as it implied that the effort cannot be reduced by simply shortening the training procedure drastically. One of the two post-analyses reported in the next section looks at a different way to reduce the effort. It determines the effect of a reduction of the textual material in each paper which is annotated. The other post-analysis looks at the argumentative structure of the author-written abstracts.

4.4. Post-Analyses

After the reliability studies had reconfirmed that the annotation can in principle be done reliably by trained annotators, Annotator C annotated the rest of the corpus. This annotation is used as system training material in chapter 5, and it also serves for the two post-analyses reported here.

4.4.1. Argumentative Structure of Author Abstracts

We wanted to establish to what extent the author abstracts differed with respect to their rhetorical structure. We therefore looked at different compositions of abstracts in terms of argumentative zones.

In the 80 papers, we found 40 different patterns, 28 of which were unique. Figure 4.21 lists all non-unique argumentative patterns in the abstracts of our corpus. The large variability reconfirms our suspicion in section 4.1.2.2 that the authors did not use a common building plan when they wrote their abstracts, in sharp contrast to how professional abstracts write their abstracts (Liddy, 1991). The composition of author abstracts seems a matter of individual choice.

The combination AIM – OWN is the single most prototypical argumentative structure we found. 29% of the abstracts in our corpus consist of this pattern. Such an abstract gives the main goal of the paper, typically followed by more detailed information about the solution. But the AIM – OWN pattern also appears as part of other abstracts: 73% of all abstracts contain it in direct sequence, and an additional 8% contain it interrupted by one other argumentative zone. A reason for the predominance of

Abstract structure	Count
Aim – Own	23
Background – Aim – Own	6
Other – Aim– Own	3
Aim – Contrast – Own	3
Other – Contrast – Aim	3
Other – Aim	2
Aim – Own – Contrast	2
Aim – Own – Aim	2
Aim – Own – Bas – Own	2
Background – Contrast – Aim – Own	2
Own – Aim – Own	2
Background – Aim	2

Figure 4.21: Typical Abstract Structures

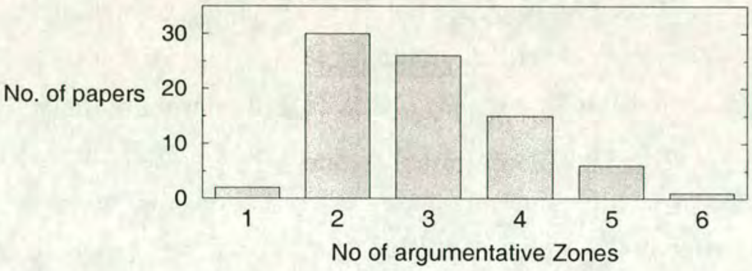


Figure 4.22: Distribution of Number of Argumentative Zones in Abstracts

this pattern might be found in the communicative function of the abstract: it is important for the success of a scientific article that the knowledge claim be established in clear terms at the earliest point of contact with the reader. This also explains the low frequency of zones referring to other researchers' work in the abstract.

Aim sentences on their own have an important function in the abstract; only one of our abstracts does not contain any Aim sentences.

Another phenomenon concerns the length of the abstracts. The average number of sentences per abstract is 4.5; the average zone in the abstract is only 1.5 sentences long. The distribution of abstract length, measured in number of argumentative zones, is given in figure 4.22. Most abstracts contain only 2 or 3 argumentative zones (average: 2.95). That is, the author abstracts in our corpus do not cover enough argumentative zones to be useful for document characterization, apart from the fact that their structure is very heterogeneous. This reconfirms our hypothesis from section 4.1.2.2: author abstracts do not provide good gold standards for Argumentative Zoning.

4.4.2. Reduction of Annotation Areas

Annotating texts with our scheme is time-consuming, so we wanted to test if the annotation of only *parts* of the source texts (which would certainly increase efficiency) would still result in reliable hand-annotated training material.

In general, we expect most of the non-basic categories (which carry the most information for our task) to be located in the periphery of the paper. For example, the TEXTUAL zone makes most sense at the end of the introduction. If an introduction section is rich in non-basic categories, it probably displays a miniature argumentative structure of the whole paper, which is generally held to be a good strategy for writing introductions (Swales, 1990; Manning, 1990). Similarly, the abstract and conclusions of source texts are often considered as "condensed" versions of the contents of the entire paper. It is thus plausible that these sections could contribute more "important" sentences to the gold standard. Additionally, one could expect these areas to be amongst the most clearly written and information rich sections in a paper.

In the following study, sections entitled *Motivation*, *Background* or *Summary* are treated as if they were called *Introduction* or *Conclusions*, respectively. As *Discussion* sections contain more speculative material, we do not treat them like *Conclusions*. Many papers do not contain explicit rhetorical sections, so we also report values for approximations of these sections: the first and last one fifth (and one tenth/twentieth) of the paper.

The abstract has a special status. As it is not clear if the abstract itself would be available for extraction in a typical practical scenario, we also report results for *aligned* abstract sentences, as discussed in section 4.1.2.2.

We test the hypothesis that the reproducibility in these special areas is higher than the overall reproducibility. If it turned out to be the case, we could either reduce annotation to these areas, or use sentences from those areas as "best fillers" to a slot (cf. section 4.1.3).

Results are given in figure 4.23: only some of the supposedly "good" areas for annotation restriction show an increase in reliability, namely only *Abstract* and *Conclusions*. These two sections have the clearest summarization function of the entire article. The effect that abstracts are more consistently annotated is even stronger in the

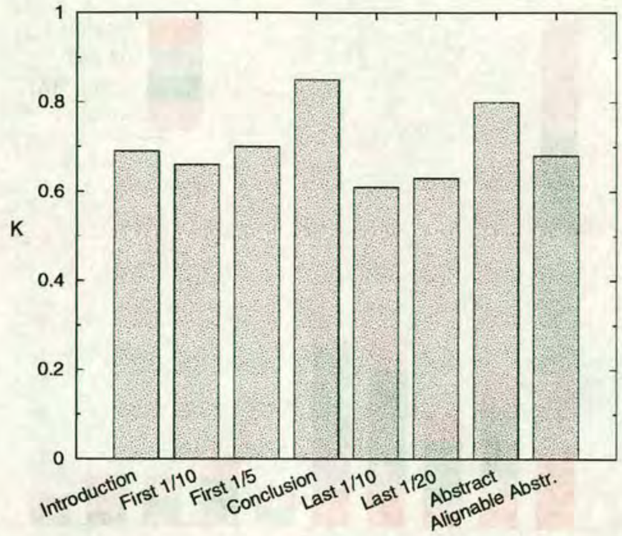


Figure 4.23: Reproducibility by Areas

basic scheme (not shown here): reproducibility within abstracts shows the very high value of K=.92. This means that authors make particularly clear in the abstract what their own contributions are.

All other areas actually show a *lower* reproducibility than the average. This is true in particular for the areas defined by absolute location (e.g. the last 1/20). These areas are therefore *not* a good approximation to *Conclusions* type material. It looks as if the last few lines in papers that do not have an explicitly marked conclusion section should not be considered at all—these sentences do not contribute "summary" type information. The *Introduction* section shows a slight decrease in reproducibility, and location approximations of introduction sections also perform badly. Reproducibility is considerably lower in alignable abstract sentences than in the abstract itself. This is consistent with our observation in section 4.1.2.2 that the rhetorical status of the aligned abstract sentences is often different from the status of the corresponding document sentences.

But there is a second point we have to take into account when restricting the areas for gold sentence selection: it is also necessary to cover all argumentative categories, as discussed in section 4.1.3. Obviously, any strategy of annotation restriction will give us fewer gold standard sentences per paper, so it is an empirical question whether there are still enough candidate sentences for all seven categories.

Some documents do not even contain all argumentative zones. In our data, each document contains at least one AIM sentence (this is required in the guidelines);

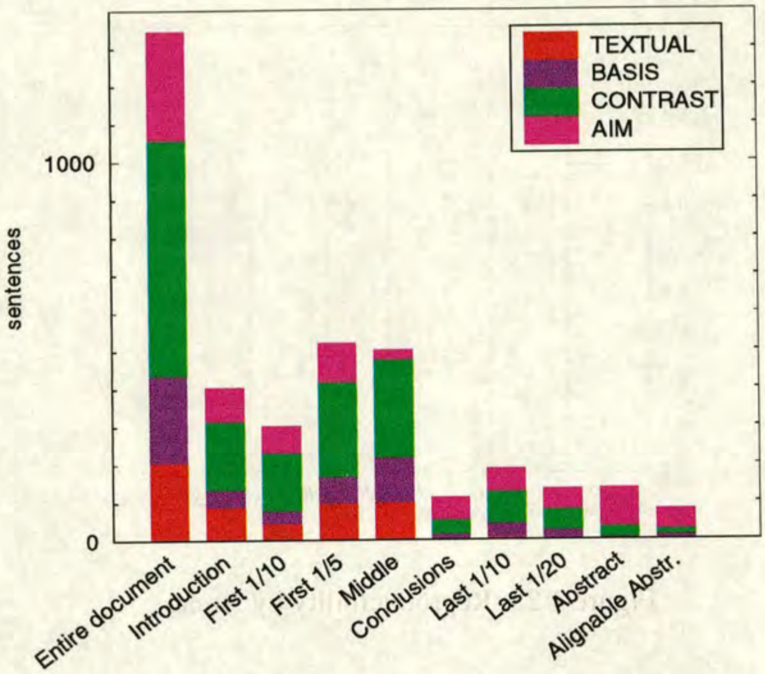


Figure 4.24: Non-Basic Areas by Categories; Absolute Values

almost every document contains at least one CONTRAST sentence (3 documents do not, i.e. 4% of our corpus). However, the use of TEXTUAL zones seems to depend much more on personal writing style. 26 % of the documents do not contain TEXTUAL zones. As the papers are conference papers and thus rather short, authors did not always perceive the function of explicitly previewing the textual presentation as necessary. Similarly, BASIS sentences are not present in 20% of the papers. However, the presence or absence of BASIS sentences seems to have less to do with writing style and more with the type of research done.

The values in figure 4.24 show absolute numbers for the occurrence of non-basic categories in special areas. For example, we can see that there are not many alignable abstract sentences anywhere in the document—a gold standard defined by alignable sentences only would thus result in bad *overall* coverage, as we have argued in section 4.1.2.2.

Figure 4.25 shows which categories can be found in a given area, and figure 4.26 shows in which areas a given category can be found. We see that some areas show a particularly low *variability* with respect to categories. Conclusions, for example, mainly consist of OWN sentences, with occasional AIM and CONTRAST sentences. Conclusions capitalize on the overall research process: they highlight own

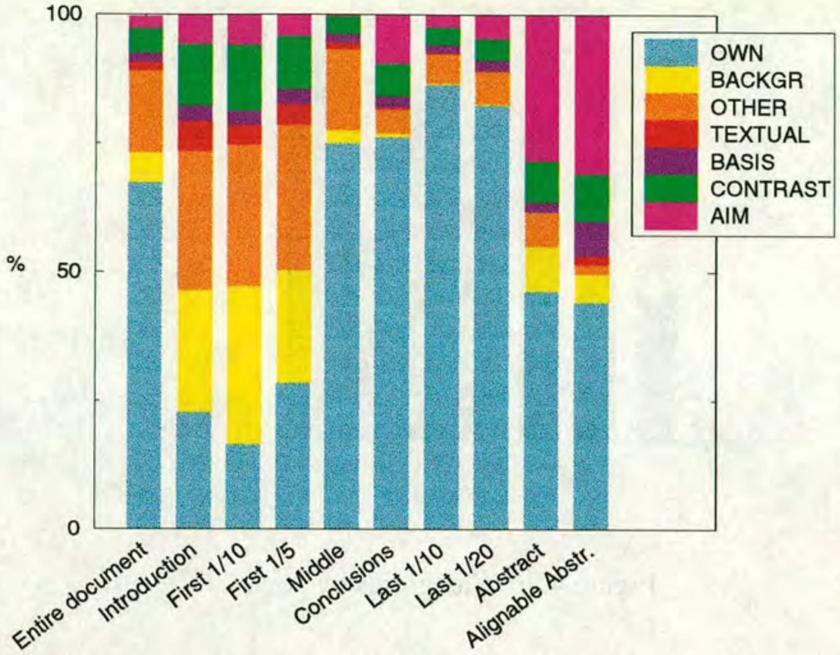


Figure 4.25: Areas by Categories; Relative Values

contribution, relevance of results, limitations, future work, and advantages over rival approaches. For some tasks, this type of information might be enough; however, we predict that it would not be enough for ours.

The relatively high proportion of AIM sentences found in abstracts would be advantageous for our task. However, even if we considered conclusion and (alignable) sentences together, coverage would still be low for certain categories, e.g. BACKGROUND, BASIS and TEXTUAL. All of these categories can be found in the introduction. It is the variety of argumentative categories in the introduction which makes annotation of this section more difficult (cf. the comparatively low reproducibility in figure 4.23), but also more rewarding for our task.

A compromise between time efficiency and quality is to annotate abstracts, introductions and conclusions where available, and first and last paragraphs as a fall back option. The price to be paid for this efficiency is in coverage and comparability. Annotated material occurring in the large area marked "Middle" or "Rest" (all document areas except alignable sentences, introduction and conclusions; black in figure 4.26), including BASIS, would get lost. Also, we cannot be sure that a given paper is written in a modular way, i.e. that it reiterates important material from the middle of the document in the periphery—some do not repeat information introduced from the abstract in the introduction section (cf. section 4.1.2.2). This is another reason why the quality

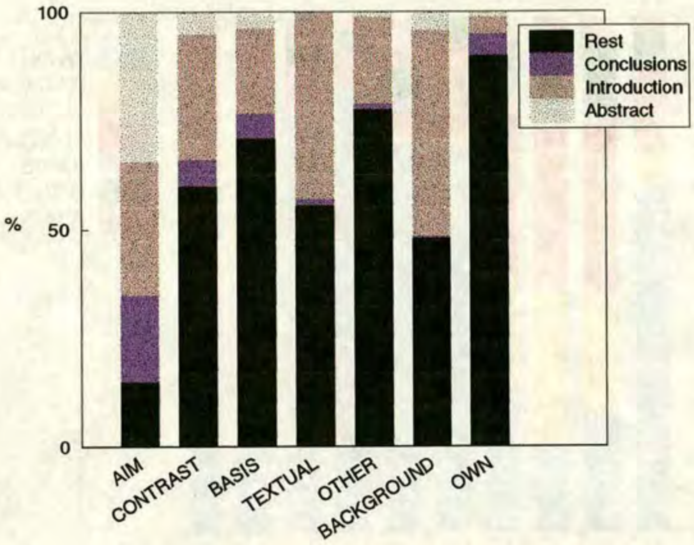


Figure 4.26: Categories by Areas

of area-reduced annotation might be lower than unrestricted annotation.

In sum, the annotation effort can be reduced by restricting the annotation to certain areas within a paper, but such a restriction has its price in quality of the gold standards. One could restrict the annotation to sentences appearing in the introduction section, even though annotators will find them harder to classify, or to all alignable abstract sentences, even if there are not many of them overall, or to conclusion sentences, even if the coverage of different argumentative categories is very restricted. The implications for Argumentative Zoning gold standards are that the advantage of time savings have to be weighed against task considerations in the concrete scenario.

4.5. Conclusion

In the first section of this chapter we discussed the question how a practical gold standard for a task like Argumentative Zoning could be constructed, and how its value could be evaluated. This discussion led to a list of desired properties of a gold standard—some of which are difficult to achieve with a surface-based evaluation strategy like ours. We have discussed why simpler gold standards, such as targets keys and free-selected sentences, are not sufficient in our text type and task. In particular, we have argued that similarity with abstract sentences does not automatically constitute a good gold standard; evidence presented in section 4.4.1 confirms this argument. Our methodology for arriving at a gold standard relies on human judgements of every sentence in the document. We decided to conduct reliability studies to measure the degree of human agreement on the task.

In section 4.2, we advocate the Kappa coefficient as a measure for annotation similarity. The main part of the chapter (section 4.3) presents the experiments: they demonstrate that the annotation scheme can indeed be learned by trained annotators and subsequently applied in a consistent way. In particular, study I shows that the basic annotation scheme, which distinguishes sentences on the basis of attribution of scientific authorship, is particularly reliable, both over time as well as between annotators. This is important, as the concept of intellectual attribution is new and central to our model of argumentation (cf. section 3.2).

Study II examines Argumentative Zoning (i.e. it uses the full annotation scheme). It shows that the two most important additional categories, AIM and TEXTUAL, are annotated reliably, but we identified some minor difficulties with the two categories BASIS and CONTRAST. As the reliability of the full scheme (as used in study II) is still acceptable, we decided to use the annotated corpus as our gold standard. This corpus is to be used for training an automatic Argumentative Zoning system, and also for intrinsic evaluation.

Study III tentatively confirms the intuitivity of the categories of the scheme, but also shows that Argumentative Zoning is a complex task which requires a certain training period in order to be performed consistently. In particular, our results show that very short annotation instructions do not provide enough information for Argumentative Zoning.

In section 4.4.1 we report the results of two post-analyses. One looks at the argumentative zones found in author abstracts and reconfirms that they cannot be directly used as gold standard. The other investigates the possibility of restrictions of the practical annotation effort by annotating only parts of papers. Our hypothesis that the reliability of the annotation in special areas of the paper would be higher in comparison to the reliability achieved overall has not been confirmed in all cases. The best gold standard is achieved when the entire paper is annotated, though we have given some alternatives for cases when such annotation might seem too costly.

Chapter 5

Automatic Argumentative Zoning

In this chapter, we will describe one method for solving the task of Argumentative Zoning automatically. As previously detailed, the task is to determine the best argumentative category for each sentence, out of a fixed list of seven categories. We have already discussed how we collected human judgements about the argumentative category for each sentence in our corpus. In this chapter, we will report on a prototype system which, on the basis of algorithmically determinable features of the sentence, learns the correlation between the human judgements and the features. An alternative system determines argumentative zones in a rule-based way. In the following, we will give an overview of the definition of the features and of the implementation, followed by results of an intrinsic evaluation.

5.1. Overview of Automatic Argumentative Zoning

Figure 5.1 gives an overview of the processes involved in automatic Argumentative Zoning. Before the experiment, the following steps had to be performed:

- *1. Feature definition*: Sentential features had to be determined which we expect to correlate with argumentative status. It is important that these features can be easily determined automatically. Our choice of features is described in section 5.2.

- *2. Human annotation*: As already discussed, a gold standard is needed, in our case in the form of human annotation of argumentative categories (cf. 4). The annotation is used for training and for evaluation.

The statistical system consists of a training and a testing phase. During training, the following steps are performed:

- *3. Preprocessing*: Each document in the training corpus is preprocessed into a machine readable format with minimal mark-up, e.g. divisions and headlines are marked (cf. section 5.3.2).
- *4. Feature determination*: For each sentence in the training corpus, values for each of the sentential features are determined automatically (cf. section 5.3.3).
- *5. Statistical training*. Several statistical classifiers are used for statistical model building, determining the correlation between sentential features and argumentative zones (cf. section 5.3.4).

Testing, i.e. the application of the statistical model to a new (test) document, uses preprocessing and feature determination in the same way as during training. This is followed by a step of

- *6. Statistical classification*: Using the model acquired in the training phase, each sentence is classified by its most likely argumentative status.

Alternatively, there is also a different system for Argumentative Zoning:

- *7. Symbolic rules*: These rules operate on the representation derived in the feature determination step (cf. section 5.3.5).

We compare human-annotated test documents against the output of the symbolic and the statistical Argumentative Zoning systems in the evaluation:

- *8. Intrinsic Evaluation:* Some parts of the training corpus are singled out for testing (i.e. they are *not* used for training). The system output is then compared with the human classification (cf. sections 5.4.1 and 5.4.2).

Finally, the output of the systems has to be displayed:

- *9. Postprocessing*: The output of the automatic and the human annotation, and the output of the automatic feature determination, are transformed into HTML (using cascading style sheets) so that the paper plus all of its annotation can be displayed in an HTML browser, eg. Netscape.

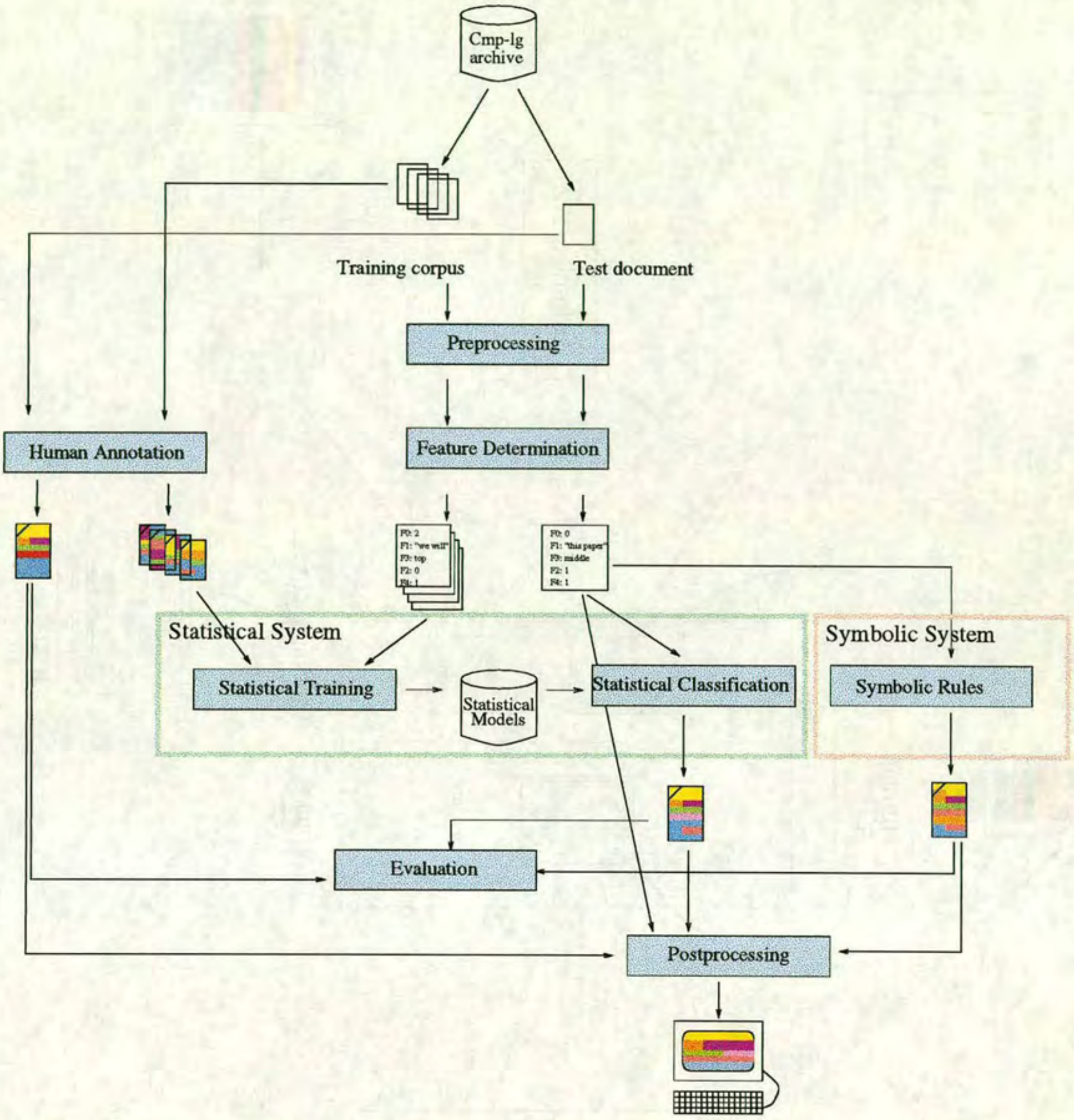


Figure 5.1: Overview of our Implementation of an Argumentative Zoner

Another overview of this rather complex setup is given in figure 5.2, which concentrates on the representations of the corpus at different stages of processing. The documents are taken from the source archive in two formats (LaTeX and PostScript). The PostScript versions are printed out and hand-annotated, the corresponding LaTeX versions are converted into XML. They constitute the training material for automatic Argumentative Zoning. After the training corpus has been automatically annotated, intrinsic evaluation is measured by the Kappa statistics, and postprocessing produces web-browsable HTML representations of the output of seen and unseen papers.

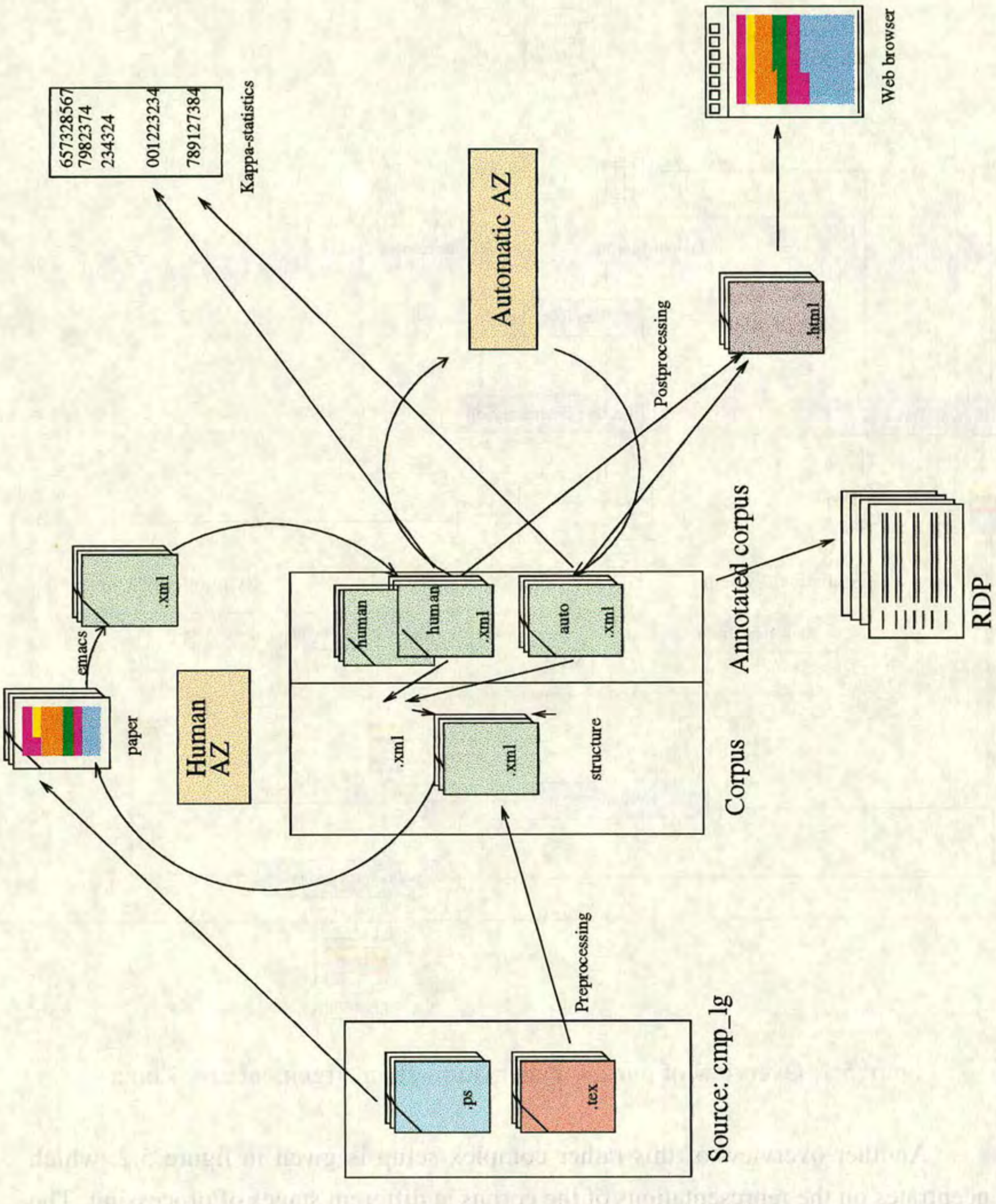


Figure 5.2: Data Flow in the Argumentative Zoner

5.2. Correlates of Argumentative Status

The argumentative status of a sentence is a property that is too difficult to determine directly algorithmically. Instead, we define heuristics which measure how appropriate it is to assign a given argumentative zone to a sentence. For this end, we need to define operationally tractable correlates (sentential features) which capture some characteristic aspect of that sentence's argumentative status.

It is generally assumed that appropriate correlates exist for similar tasks. For example, human summarizers are guided by sentential features like location and the occurrence of certain cue phrases when they determine importance of a textual segment (Cremmins, 1996); and the text extraction literature provides us with a pool of such features (heuristic measures) for sentence *relevance* (Paice, 1990; Luhn, 1958; Baxendale, 1958; Edmundson, 1969; Kupiec et al., 1995).

The task of Argumentative Zoning moves away from the concept of sentence relevance towards a new concept of argumentative status. Our annotation scheme can be interpreted as encoding different *types* of relevance. We have defined four different kinds of sentences which are particularly important for the global argumentation of the paper (the non-basic categories), and three categories which provide background information. All of these are important for different reasons. We have assumed so far that there are correlates of this argumentative status in our texts which can be read off the surface.

It might well be that the features which are useful for our task differ from the ones used for determining global relevance. Figure 5.3 gives an overview of our feature pool. Some of the features we use (the Content, Explicit Structure, Absolute Location, Formulaic and Sentence Length features) are borrowed from the text extraction literature, but in some cases, changes were necessary; the Formulaic feature, for example, is an elaboration of similar, simpler features used previously. We also use features not typically used for text extraction, namely the Syntactic, Citation and Agentivity Features; as far as we know, we are the first to define these for any task.

When defining the features, we tried to make them maximally *distinctive*. In order to do so, we used information provided in *contingency tables*. A contingency table lists the values of a given feature with its counts in the corpus, cf. figure 5.4.

Distinctive features have heterogeneous (skewed) distributions, i.e. distributions which differ as much as possible from the overall distribution of categories. There are statistical measures for this heterogeneity, e.g. g-score (Dunning, 1993). In section 5.3.3, we will provide the contingency tables for each of our features; the use of contingency tables for statistical classification will be discussed in section 5.3.4.

Type	Name	Feature description	Feature values
Content Features	Cont-1	Does the sentence contain "significant terms" as determined by the *tf/idf* measure?	Yes or No
	Cont-2	Does the sentence contain words also occurring in the title or headlines?	Yes or No
Absolute location	Loc	Position of sentence in relation to 10 segments	A-J
Explicit structure	Struct-1	Relative and absolute position of sentence within section (e.g. first sentence in section or somewhere in second third)	7 values
	Struct-2	Relative position of sentence within a paragraph	Initial, Medial, Final
	Struct-3	Type of headline of current section	16 prototypical headlines or *Non-Prototypical*
Sentence length	Length	Is the sentence longer than a certain threshold in words?	Yes or No
Verb Syntax	Syn-1	Voice (of first finite verb in sentence)	Active or Passive or NoVerb
	Syn-2	Tense (of first finite verb in sentence)	9 simple and complex tenses or NoVerb
	Syn-3	Is the first finite verb modified by modal auxiliary?	Modal or no Modal or NoVerb
Citations	Cit-1	Does the sentence contain a citation or the name of an author contained in the reference list?	Citation, Author Name or None
	Cit-2	Does the sentence contain a *self* citation?	Yes or No or NoCitation
	Cit-3	Location of citation in sentence	Beginning, Middle, End or NoCitation
Formulaic expressions	Formu	Type of formulaic expression occurring in sentence	20 Types of Formulaic Expressions + 13 Types of Agents or None
Agentivity	Ag-1	Type of Agent	13 different types of Agents or None
	Ag-2	Type of Action, with or without Negation	20 different Action Types X Negated/Non-negated, or None

Figure 5.3: Overview of Feature Pool

Paragraph (Struct-2)	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
Initial	117	92	267	135	601	2532	73	3817
Medial	56	87	306	289	971	3779	68	5556
Final	34	47	147	172	442	2125	82	3049
Total	207	226	720	596	2014	8436	223	12422

Figure 5.4: A Contingency Table: Paragraph Feature

Another desired property is *coverage* (as opposed to *peakiness*). Some features are strong indicators of a certain category, but occur very rarely in the corpus. For the average sentence, such a feature would not be of help for classification. Moreover, such features lead to *over-fitting*, a problem which occurs when features encode idiosyncrasies of the training data which are accidental to the data. The feature will then not provide useful information for unseen, but similar data. An example for a peaky feature is the occurrence of the phrase "*in this paper*" in a sentence. Evenly distributed features (e.g. verb tense) have a higher coverage, i.e., they can be more reliably estimated from text. They typically do not give strong indications, but many of them in combination might influence the statistical classification into the right direction. We have tried to find a compromise between features that are peaky and those that are evenly distributed.

The choice of the values for the features is not independent of the classification method chosen. We initially followed Kupiec et al. (1995) in using a Naive Bayesian classifier. Later, we used other classifiers, but the original design of the features was influenced by the intention to use them in a Naive Bayesian classifier. This classifier demands that features must have discontinuous values, and in practice it also implies that feature values all fall into a small set of distinct values. Too many values might influence classification results negatively as there might not be enough training material available for the rare values. Thus, we often had to *cluster* values into classes; we did so manually. Another limitation is that Naive Bayes allows only one value of a feature per classified item. Additionally, Naive Bayes assumes that the features are *statistically independent* of each other, so we tried to identify features which would classify sentences into certain categories for reasons different from the other features in the feature pool.

5.2.1. Traditional Features

5.2.1.1. Content Features

The assumption behind the content features is that concepts (approximated by textual strings) are representations of the semantics of the text span in the context of the overall document. Different content features might differ in exactly *how* they determine the most salient concepts in a text span. Content features are used in most of today's sentence extractors, i.e. for determining global sentence relevance.

The two content features we use are different from the other heuristics in our pool in that they concentrate on *subject matter* rather than more structural or rhetoric cues. We hypothesized that content features should be less important for Argumentative Zoning than the other features, as it is not immediately obvious how the fact that a certain sentence contains characteristic subject-matter key words would help determine its argumentative category.

Term Frequency (`Cont-1`): `Cont-1` uses the *tf/idf* (term frequency times inverse-document-frequency) method, which employs lexical frequency to identify concepts that are characteristic for the contents of the document. The *tf/idf* method is successfully used for information retrieval (Salton and McGill, 1983).

tf/idf tries to identify diagnostic units (textual spans) which are frequent in one document but rare in the overall collection. This is achieved by combining the relative frequency weights (*tf*) with a function of the inverse frequency of the diagnostic unit in the overall text collection (the *idf* element), e.g. the number of documents where this term occurs, or the frequency of overall occurrences:

$$tf/idf_w = tf_w * log(\frac{100*N}{df_w})$$

td/idf_w: td/idf weight for diagnostic unit w
tf_w: term frequency of w in document
df_w: number of documents containing diagnostic unit w or number of occurrences of w in document collection
N: number of documents in collection

If a diagnostic unit appears often in the overall collection, it is assumed that it represents a concept which is common in the domain, and which has a low discriminating power—as a result, it is penalized by a low *idf* score. If a diagnostic unit appears only once, it might be noise (e.g. misspelled words); such words can be filtered out by frequency thresholds.

In the first text extraction experiments (Luhn, 1958; Baxendale, 1958), a predecessor of today's *tf/idf* formula was used, which relied only on the *tf* part. There are variations of the formula used in the literature (e.g. Brandow et al. (1995) use the logarithm also for the *tf* part). Other approaches have varied the diagnostic units used. Luhn's (1958) diagnostic units were the most frequent content word *stems* (after function words had been stripped out with a stop list), i.e. *"hypothesis"* and *"hypothesize"* were reduced to the same stem. Nowadays, the simplest implementations use either full words or lemmas (words normalized to their lexicon entries). Other implementations use nominal pairs, or noun groups determined by partial parses, derived by techniques like chunking (shallow parsing of NP and VP complexes; Abney 1990; Grefenstette 1994). Georgantopoulos (1996) improves results achieved by Finch and Mikheev (1995) by using noun groups as diagnostic units.

There has also been criticism of the method, as it cannot handle synonymy, pronominalization, general co-referentiality and conceptual generalizations such as the replacement of a list by its superordinate term (Hovy and Lin, 1999; Mauldin, 1991). This limitation has been referred to in IR as the "keyword boundary".

An additional criticism questions if the application of *tf/idf* measures from document retrieval to text extraction is sensible, i.e. if the transition from *documents* as units of scoring to smaller units like *sentences* actually works. (Hearst, 1997) voices the intuition that *tf/idf* works much better to determine important concepts which distinguish *between* documents rather than between smaller segments *within* a document:

> [...] the estimates of importance that *tf/idf* makes seem not to be accurate enough within the scope of comparing adjacent pieces of text to justify using this measure [...] (Hearst, 1997, p. 44)

Title Words (`Cont-2`**):** `Cont-2` draws its definition of what a good keyword is from occurrences of a word in the title and headline. This feature goes back to Edmundson (1969). The assumption is that words occurring in the title are good candidates for document specific concepts. Particularly in experimental disciplines, titles can be a document surrogate in themselves, as they often summarize the main knowledge claim of the document ("*Low Dose Dobutamine Echocardiography Is More Predictive of Reversible Dysfunction After Acute Myocardial Infarction Than Resting Single Photon Emission Computed Tomographic Thallium-201 Scintigraphy*"; American Heart Journal, 134(5): 822-834, 1997).

Along the same lines, headlines are considered summaries of the major sections of the document—unless they are prototypical headlines such as *Introduction* or *Results*.

However, in other fields, "jokey" titles have become fashionable (*"Four out of five ain't bad"*; Archives of General Psychiatry, 55(10): 865-866, 1998). This practice makes reliance on title heuristics risky as titles do not necessarily express the document's topic anymore.

5.2.1.2. Absolute Location

The next two features use the location of a sentence in text. In many previous experiments, local organization within a section has been correlated with importance. Experiments in text extraction have assumed that more relevant sentences can be found in the periphery of the document (Edmundson, 1969). Indeed, in other genres like newspaper text, location has been shown to be the single most important feature for text extraction (Brandow et al., 1995; Hovy and Lin, 1999).

Absolute location, in terms of absolute spatial organization of information in the linear medium of text, should be a good correlate for Argumentative Zoning. Readers have certain expectations of how the chain of argumentation will proceed and which argumentative components are handled in which areas of the paper.

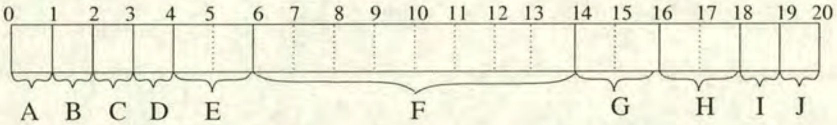


Figure 5.5: Values for Location Feature

We divide the document into 20 equally sized segments; we then collapse some of these (cf. figure 5.5), resulting in 10 differently-sized segments which mimic the structure of ideal documents. Segment size is smaller towards the beginning and the end of the document, where documents are often written more densely, i.e. where we expect the author's rhetorical units to be smaller. In the middle, the segments are larger (cf. segment F in figure 5.5, which covers 40% of the text).

5.2.1.3. Structural Correlates

The structural features seek to exploit the explicit hints given by the author about the structure of the paper.

Section Structure (`Struct-1`)**:** We noticed that apart from global locational structure, there is also a section internal locational organization which might be important for Argumentative Zoning. Introductions usually proceed from the more general to the more specific, with general knowledge typically coming first and statements about own work appearing towards the end. In particular, AIM sentences often occur in a typical position about two-thirds down in introduction sessions.

We also observed that the first and last sentences in other sections often fulfill a summarizing function, and are often associated with text-organization meta-discourse (*"in this section we will"*), which is captured by our TEXTUAL sentences. The second and third or second and third-last sentence also often have a special summarizing function.

The feature `Struct-1` divides the section into three equally sized segments, and additionally singles out the first and the last sentence, and takes together the second and third sentence as a sixth value, and the second-last plus third-last sentence as a seventh value.

Paragraph Structure (`Struct-2`)**:** There is disagreement in the literature whether paragraph information should be considered as a surface indicator of importance and topic boundaries. Are paragraphs regarded as logical units by authors, or rather as layout units?

(Baxendale, 1958) states that due to the hierarchical organization of well-written research papers, sentences at the beginning and end of the paragraph are more likely to be "topic sentences"—in 85% of the paragraphs, the topic sentence was the initial sentence, and in 7% the final. Marcu (1997b) also suggests that paragraph breaks help readers determine the most important textual units in a text.

In contrast, Longacre (1979) holds that the function of many paragraph breaks is purely aesthetic, and Starck (1988) conducted an experiment which confirms the marginal role of paragraphs in higher-level interpretive tasks. The task of human re-introduction of paragraph breaks led to poor results: only nine of the 17 paragraph breaks in a text were correctly identified as such by more than 50% of the subjects. We lean towards the layout argument: we believe that in conference papers, the number and placement of paragraph breaks will be affected by the question whether or not a paper was printed in "two-column" style.

Even if we do find crucial information at the beginning and the end of paragraphs, we still do not know how useful this is for Argumentative Zoning. With respect to other tasks, Hearst (1997) indicates that thematic boundaries do not always occur at paragraph boundaries, but Wiebe (1994) states that the information whether or not a sentence begins a paragraph is useful for her task, namely the determination of private-state sentences in narrative (subjective vs. objective orientation). In our case, it seems sensible to assume that CONTRAST sentences are more likely to occur at the end of a paragraph, but other than that it seems difficult to predict a direct correlation between paragraph boundaries and argumentative flow. We included the feature in our heuristics pool to determine its usefulness empirically.

Headlines (`Struct-3`): Van Dijk (1980) states that in scientific articles, rhetorical sections are marked by fixed headlines. Knowing which rhetorical section a sentence belongs to should be directly useful for Argumentative Zoning. For example, Nanba and Okumura (1999) assume a correlation between rhetorical section and type of citation. They expect CONTRAST citations to occur more often in the sections *Introduction, Discussion*, and *Related* work, and BASIS citations to occur more often in the *Introduction* and the *Method* section.

However, we have argued in section 3.1 that not all articles in our corpus keep to a fixed section structure. As a result, we expect the feature `Struct-3` to be of use only in those cases where prototypical headings are available.

Feature `Struct-3` classifies the headlines into groupings of similarity on semantic grounds and morphological variants, resulting in the following 15 classes: *Introduction, Problem Statement, Method, Discussion, Conclusion, Result, Related Work, Limitations, Further Work, Problems, Implementation, Example, Experiment, Evaluation, Data* and *Solution*. Pattern matching of a range of expressions in the headlines is applied. If no pattern matches, the value *NonPrototypical* is assigned.

5.2.1.4. Sentence Length

At first glance, the criterion of sentence length seems to be a trivial criterion which is not related to relevance or to argumentative zones. For trivial features, we expect a distribution which is near–identical to the global distribution of categories in the corpus, and therefore no help for a statistical classifier.

Kupiec et al. report better results when including the Sentence Length feature, but this point seems to be pertinent to their data coding: captions, titles and headings are not encoded as such and the sentence length feature can filter them out. In our corpus, this information is already directly encoded: sentence length thus cannot fulfill the filtering function.

But there are some other reasons why sentence length might *not* be a trivial feature after all. Sentence length is one indicator of sentence complexity which has been used in extraction experiments before. Earl (1970) argues that short sentences in her material are more likely to contain trivial material. Robin and McKeown (1996) state that complex sentences (conveying a maximal number of facts) are advantageous as a summary. There are, of course, other criteria for complexity apart from sentence length. Some measurements try to determine how contentful the sentence is by calculating the proportion of content words per length, or by measurements of the syntactic complexity of the sentence.

Sentence length might be a useful feature for Argumentative Zoning due to the high number of OWN sentences in our corpus, which describe details of the solution. They contain less meta-discourse than other sentences, and they tend to be less complex and thus shorter.

5.2.1.5. Syntactic Correlates of the Verb

In text extraction, there have been some efforts to use purely syntactic criteria for the indication of overall relevance, but most of these proved unsuccessful. Baxendale (1958) used the objects of prepositions as sole representation for the document. Earl (1970) describes an unsuccessful experiment to correlate global importance to the parts-of-speech (POS) shape of sentences. However, there were too many different POS shapes, and she concludes that:

> it seems fair to say that indexible and non-indexible sentences cannot be distinguished by structure alone. (Earl, 1970, p. 321)

Also interesting are experiments differentiating different linguistic factors per rhetorical sections. These experiments concentrate on the standard four-part fixed structure (*Introduction, Methods, Results, Discussion*), which is, as we have argued before, related to argumentative zones, albeit not in a trivial way (cf. section 3.1).

Verbal syntactic features can be indicators of rhetorical section structure, as studies like Biber and Finegan (1994) and Milas-Bracovic (1987) show. West (1980), for example, manually determined and counted the occurrence of *that*-nominals (e.g. *"the fact that…"*) in different rhetorical sections. That-nominals often indicate knowledge-stating sentences. West found that the density of that-nominals differed significantly between rhetorical sections: there were statistically more that-nominals in the *Introduction* and *Discussion* sections than in the *Results* section. The *Methods* section has fewer that-nominals than any other section.

Myers's (1992) work is particularly relevant to Argumentative Zoning. He describes properties of sentences stating authors' knowledge claims (our AIM sentences). Apart from two non-linguistic features (cue phrases and location), he lists the following linguistic features of the main verb in such sentences:

- Verb: *"to present", "to report"* or similar
- Tense: Present Perfect
- Person: First

We consider only verbal syntactic features here: voice, tense and the existence of a modal auxiliary.

Voice (`Syn-1`**):** Riley's (1991) work shows that there is a correlation between rhetorical roles and the use of the passive tense. The explanation for this is that voice is connected to *authors' perspective*. Prescriptive accounts of academic writing advise writers to avoid the mention of the own person, in order to avoid the impression that they are unduly interested in the success of their own research. This results in a high proportion of passive sentences, and often makes texts less readable and more difficult to understand. If a text is written in this style, it is sometimes difficult to tell who performed a certain research action. Many authors in our collection use the active voice instead to describe their own work, but nevertheless, there are also articles which use the passive voice frequently.

Tense (`Syn-2`**):** It has been hypothesized that authors use different tenses for different rhetorical segments (Biber and Finegan, 1994; Milas-Bracovic, 1987) or for certain argumentative tasks. Aspect and tense have been shown to correlate with discourse structures (Salager-Meyer, 1992; Hwang and Schubert, 1992; Malcolm, 1987). The connection between aspectual information (which is predominantly expressed by tense in English) and argumentation is that aspect signals the state of an activity (*"has the problem been solved or is it unsolved yet?"*). For example, the present perfect, being used for unfinished states, is often associated with pending problems, whereas the use of past tense, particularly in combination with statements of solution-hood, signal an accomplishment, i.e. the fact that an end state has been reached.

Another reason why tense should be an interesting feature for Argumentative Zoning is that many formal guidelines for publication, e.g. in certain journals, require authors to use past tense for descriptions of previous work, including own previous work, and present tense for current work. This distinction, as it is connected to the attribution of ownership, is particularly important for Argumentative Zoning. On the other hand, many of the authors in our collection are non-native speakers and might use tense in an idiosyncratic way.

Modality (`Syn-3`)**:** The use of modal auxiliaries is one of the correlates for a phenomenon called *hedging* (cf. Hyland's (1998) hedging category in figure 3.13, p. 100). Hedging occurs when authors distance themselves from a scientific statement (Salager-Meyer, 1994). Other correlates of hedging are adverbials like *likely, possibly, maybe* which formed part of Edmundson's negative cue phrases. Hedging has been proposed as a signal for rhetorical sections, as it is associated with speculative statements in *Discussion* sections. Wiebe (1994) also uses the occurrence of a modal other than "*will*" for her subjective/objective distinction.

5.2.1.6. Citation Features

Type of Citation (`Cit-1`)**:** Citations are a good indication that the topic of the sentence is somebody else's work; our human annotators use this factor to distinguish between OTHER and BACKGROUND categories. Thus, the existence or non-existence of formal citations should prove useful for Argumentative Zoning. We also believe that mentions of other authors' names in the text, even if these do not occur in a formal citation context, have a status similar to full citations. Consider sentence **8** of our example article:

> *In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events.* (S-8, 9408011)

The full citation was used in sentence **5**; similarly to the use of pronominal reference, use of the author's name avoids repetitiveness. We think that in this sentence should be logically treated as if it had read "*In Hindle's (1993) proposal*", i.e. as if a formal citation had been present.

Self Citations (`Cit-2`)**:** If some own previous work is mentioned in a paper, it is very likely that the authors mention it because they base their own work on it (BASIS). Therefore, the fact that previous work is the author's own should be recognized.

Citation Location (`Cit-3`)**:** Citations are authorial if they form a syntactically integral part of the sentence, or parenthetical if they do not (Swales, 1990). We believe that the attribution of intellectual ownership is more often expressed by authorial citations, and that parenthetical citations are often there for other reasons ("piety, policy, politeness" cf. Ziman (1969)). If this is true, the syntactic type of a citation might prove useful for Argumentative Zoning.

As authorial citations form the subject of the sentence, they typically occur in the beginning, whereas most of the parenthetical uses of citations occur in the end of the sentence. Citation location (`Cit-3`) captures exactly this aspect.

5.2.2. Meta-Discourse Features

Meta-discourse represents one of the most reliable indicators of rhetorical status and is potentially very useful for Argumentative Zoning. Other computational approaches (Marcu, 1997a; Litman, 1996) also exploit meta-discourse, but meta-discourse of a different kind: short cue phrases belonging to a closed-class vocabulary (e.g. adverbials, sentence connectives or general relevance markers like "*in sum*"). As a result, the linguistic realization of such meta-discourse phrases tends to be invariant between disciplines and authors.

But when we looked at realizations of scientific meta-discourse in section 3.2.5, we found that apart from formulaic, fixed meta-discourse (*"to my knowledge", "in this paper"*), there is another kind of meta-discourse which shows a wide range of syntactic variation—recall the different ways of expressing intellectual ancestry exemplified in figure 3.14 (p. 102). It is difficult to see how this type of meta-discourse could be captured with a fixed list; a more flexible way of analyzing it is needed.

We suggest that one way out of the dilemma of linguistic variation is to discover prototypical *agents* and *actions* individually in a wider range of syntactic contexts, e.g. in passive and active constructions (Teufel and Moens, In Prep.). Looking at the examples for the argumentative moves in figures 3.7, 3.9, 3.10, 3.12 and 3.15, one cannot help noticing that scientific argumentative text abounds in prototypical agents and actions, which recur in different syntactic disguises. We argue that it should be enough for Argumentative Zoning to recognize these prototypical actions and agents, while reading over all agents and actions that are not understood (and which are likely to refer to the science in the paper). As the patterns themselves are rather prototypical (*"our approach"*), pattern matching and syntactic heuristics should be able to find a large part of these agents and actions.

This would provide a simple profile of the agent/action structure of the document: the information of "who-does-what". We assume that the agent/action structure is an integral part of the kind of document structure that we are looking for, and should help us perform Argumentative Zoning. We also believe that the agent/action structure provides a deeper, more semantic–oriented kind of text representation than the text strings themselves. Such intermediate representations have been called for by Spärck Jones (1999) as a prerequisite for better text summarization strategies.

One last caveat: the phrases we call meta-discourse *can* have a meta-discourse interpretation—but they do not always have this interpretation. Litman (1996) uses machine learning to address the problem that the phrase *"so"* can function as meta-discourse or as propositional contents. There are some ambiguity problems associated with our approach, which we discuss in section 6.2.

5.2.2.1. Formulaic Expressions (`Formu`)

The Formulaic Expressions Feature is designed to determine and classify explicit meta-discourse statements of a fixed kind.

Indicator or cue phrases have a long history as features for text extraction, i.e. for determining global sentence importance. In Edmundson's (1969) approach, sentences containing positive cue phrases like superlatives or explicit markers of importance or confidence ("*important*", *"definitely"*) were considered fit for extraction, whereas other sentences containing *stigma* words like *"hardly", "unclear", "perhaps", "for example"* (belittling expressions, expressions of insignificant detail or speculation/hedging) were discouraged from extraction. Edmundson's list was statistically acquired and manually corrected. A similar but much more extensive list containing 777 terms (called the Word Control List or WCL) was used in ADAM, the first commercially used automatic abstracting system (Pollock and Zamora, 1975).

More recent work on longer indicator phrases has been done by Paice and colleagues (Paice, 1981; Paice and Jones, 1993; Johnson et al., 1993), whereby sentences containing explicit rhetorical markers like "*the purpose of this research is*" or "*our investigation has shown that*" are considered fit for extraction. Paice (1981) describes the first implementation of a pattern-matching extraction mechanism relying on indicator phrases. Paice and Jones (1993) make the method more flexible by supplying a finite state grammar for indicator phrases specific to the agriculture domain; however, Oakes and Paice (1999) state that importance cues are often not reliable.

All these approaches use indicator phrases which indicate *global sentence relevance*—again, using indicator phrases for the determination of argumentative status is different. For example, the phrase "*in this paper, we have ...* " is a very good overall relevance indicator: it is quite likely that a sentence or paragraph starting with

Formu: Formulaic Expression Types

Type	Example	Type	Example
GAP_INTRODUCTION	*to our knowledge*	PREVIOUS_CONTEXT	*elsewhere, we have*
OUR_AIM	*main contribution of this paper*	FUTURE	*avenue for improvement*
TEXTSTRUCTURE	*then we describe*	AFFECT	*hopefully*
DEIXIS	*in this paper*	PROBLEM	*drawback*
CONTINUATION	*following the argument in*	SOLUTION	*insight*
SIMILARITY	*similar to*	IN_ORDER_TO	*in order to*
COMPARISON	*when compared to our*	POSITIVE_ADJECTIVE	*appealing*
CONTRAST	*however*	NEGATIVE_ADJECTIVE	*unsatisfactory*
DETAIL	*this paper has also*	THEM_FORMULAIC	*along the lines of*
METHOD	*a novel method for X-ing*	GENERAL_FORMULAIC	*in traditional approaches*

Figure 5.6: Formulaic Expression Types (Feature Formu)

it will carry important discourse-level information. However, without knowing the following verb, we cannot be sure about the argumentative status of the sentence. It could continue with "... *used machine learning techniques for* ... ", in which case the sentence is likely to be a description of solution/methodology; with a different verb, it might also be a conclusion ("... *argued that* ... ") or a problem statement ("... *attacked the hard problem of* ... ").

Our argumentative model in section 3.2 describes typical statements about the problem-solving processes in research. Our method for finding meta-discourse is to use pattern-matching on expressions that are expected by the model of argumentation introduced in section 3.2. We particularly concentrate on those meta-discourse expressions which have become formulaic expressions of scientific writing (cf. Hyland 1998; Swales 1990).

Our formulaic expressions are bundled into 20 major semantic groups. Figure 5.6 gives examples for the types of formulaic expressions used in feature Formu. For example, a marker like *"our goal in this paper"* is expected to co-occur frequently with the AIM category, whereas *"in the following section"* is a good marker for TEXTUAL. On the other hand, if we find a negative polarity item in the sentence e.g. "*however*"; *"no method has...* "; *"none of the approaches...* ", this raises the probability that we are dealing with a sentence which indicates a flaw of some other work (CON-

TRAST). Another good indication of a gap in knowledge is the phrase *"to our knowledge"*. The full list of 396 formulaic patterns is given in appendix D.1.

5.2.2.2. Agentivity Features (Ag-1 and Ag-2)

The recognition of prototypical agents and actions serves to identify scientific meta-discourse which is less fixed than the phrases covered by the Formu feature. For writing styles that do not use much meta-discourse it might be particularly advantageous to determine agents and actions, because they might provide the only superficially marked correlates of argumentative status. For data collections with large variations in meta-discourse like ours, it makes sense to *classify* the agents and actions. Then it does not matter which particular term the authors use (e.g. *"we"*, *"I"* or *"one of us"*)—these expressions are represented as the same entity (US_AGENT), and automatic processing can generalize over the same concept.

Possibly the closest related work with respect to agents and actions is that of Barzilay et al. (1999), which uses overlap of actions and agents to detect the similarity of events in newspaper paragraphs. However, whereas in our text type *prototypical* agents are particularly relevant, in their text type (news stories), any potential agent needs to be matched.

In our approach, agents and actions are expressed separately and modularly; their syntactic context is recognized (passive vs. active), and negation is automatically taken into account. Such an approach is more robust and less error-prone than standard pattern matching methods which are string-based, as individual subject–verb combinations might easily be forgotten from such lists.

Using syntactic constraints in Agentivity features (i.e. agents and actions) also increases the precision of pattern matching. As an example, GAP_AGENT pattern are designed to find statements expressing the lack of a solution (*"no papers/articles/studies describe a solution to the problem...* "). But when GAP_AGENT patterns (e.g. *"no articles"*) are applied without syntactic restrictions (i.e. anywhere in the text), the error rate is high: 5 out of the 13 GAP_AGENT occurrences in our corpus were erroneous. The problem is polysemy: *"article"* can mean article-in-a-journal (the interpretation intended here), or it can also mean the grammatical article (*"a"* or *"the"*). If we, however, search for GAP_AGENT patterns only in subject positions (as determined by our heuristics), we reduce the error due to polysemy completely, and we get 9 out of 9 occurrences with the correct meaning.

For the practical implementation, we made the decision to give grammatical subjects (or by-objects in passive sentences) a special status by encoding them in feature `Ag-1`; we disregard grammatical patients (typically direct objects) even though in many cases the information contained in objects is potentially relevant too (*"we solve the problem of..."*). However, we feel that the robust recognition of subjects (agents) and semantic verbs (agents), as in our approach, is a workable middle ground between shallow and deep text representation.

Agents (`Ag-1`): Agent-hood should be a good indicator of Argumentative Zoning, as it is related to attribution of authorship, which is a defining factor in basically all of our categories. The main agent groups are US_AGENT, GENERAL_AGENT and THEM_AGENT.

Authors often have to refer to themselves; we call this agent class US_AGENT. The terms *"I"*, *"we"* and *"the first author"* all refer to this class. Personal pronouns in 1st person ("*I*" and "*we*") are an important help. The Roman number 1, can, however, be mistaken for the pronoun "*I*", as in the following erroneous example:

> <AGENT TYPE="US_AGENT"> **I** </AGENT> **is an interpretation iff** <AGENT TYPE="US_AGENT"> **I** </AGENT> **is a triple** <EQN/>
>
> (S-21, 9408003).

As we do not check for subject-verb agreement, such errors cannot be avoided in our processing, but they do occur only rarely.

There are also cases where the explicit marking of agenthood might be deceptive. A sentence starting with "*we*" might occasionally have a different function from describing own work. It might be used to clarify notation, to draw preliminary conclusions, to direct the attention of the reader to some non-obvious fact or to explain the presentational form in which an idea (possibly attributed to somebody else) will be presented in the article.

For example, authors might state in one sentence that researcher X has introduced a particular algorithm. The next sentence might state that "*We will demonstrate how the algorithm works by way of example*"—followed by a long (unmarked) description of the algorithm. It is clear to humans that these sentences are attributed to X, and not to the authors. A simple algorithm which assumes that non-marked sentences always carry the status that the last marked sentence displayed will, however, lead to the wrong guess that the long segment is attributed to the authors.

Distinguishing previous own work from the current approach is a difficult case. After such previous own work has been introduced with a self citation, most authors use a 1st person pronoun to refer to it, but some authors use a 3rd person pronoun (particularly if the cited paper is co-authored). However, we found no 3rd *singular* pronominal reference to own previous work in our corpus. The use of 3rd person pronomina might have to do with the instructions for double-blind reviewing of papers: The instructions specifically state that citations of own previous work should not reveal the identity of the author, and many authors obviously did not not change the pronomina after the paper was accepted.

There is a real problem if the description of own previous work is directly followed by a description of the current work in the paper, and if the authors do not use an explicit formulaic signal ("*in this paper*"). In this case, it is almost impossible to guess where in the text "*us*" stops to mean "*us, previously*" and begins to mean "*us, now*".

Noun phrases with a possessive 1st person determiner ("*our*" or "*my*") also indicate own work, if the head of that noun phrase is a prototypical solution (e.g. *"theory, approach, method, algorithm"*), as the authors' approach or solution is often equated with the players "US". The solution type list is also used for the METHOD pattern above in Formu. Our list of solution nouns is given in appendix D.4.

When trying to find mentions of "THEM_AGENT" in text, the following patterns lend themselves well:

- Authorial citations are the best indication of a THEM_AGENT.

- The names of other researchers is an equally good indication of a THEM-AGENT. In our implementation, author names are recognized and are annotated before processing.

- 3rd person possessive pronoun plus solution nouns ("*their system*").

- Personal 3rd pronouns *can* refer to THEM_AGENTs, particularly after formal references (and if the grammatical number is right). However, 3rd person personal pronouns might just as well refer to other things: Singular pronouns often refer to fictional characters in the example sentences. The plural pronoun "*they*" can refer to any plural object in the research world, e.g. rules, formulae or trees.

- A demonstrative pronoun plus a solution noun (*"this approach"*) is ambiguous between a reference to US_AGENT and to THEM_AGENT.

When trying to find mentions of "THEM-GENERAL" in text, the patterns we are looking for are quite formulaic.

- Some expressions follow the pattern "*general people in the field*". We use a list of professions, e.g. "*workers, linguists, computer scientists, researchers...*" and allow for syntactic variations, e.g. modification with typical adjectives.
- Other expressions follow the pattern "*previous papers*". We use a list of entities like "*article, paper, work, research*" and allow for syntactic variations. All these groups of nouns can be found in appendix D.4.
- Yet other expressions are variations of the pattern "*traditional solutions in the field*". We use the aforementioned list of solution types.

Figure 5.7 lists the agent types we distinguish. Rather than just the agent types US_AGENT, THEM_AGENT and GENERAL_AGENT and a fourth type US_PREVIOUS_AGENT, there are altogether 13 types. Some of these are non-personal (pseudo) agents like aims, problems, solutions, absence of solution, or textual segments: OUR_AIM_AGENT; PROBLEM_AGENT; SOLUTION_AGENT; GAP_AGENT; TEXTSTRUCTURE_AGENT (*"this section"*). In other agent types the syntactic form does not allow to determine the referent unambiguously, e.g. because of pronominal

Ag-1: Agent Types

Type	Example
US_AGENT	*we*
REF_US_AGENT	*this paper*
OUR_AIM_AGENT	*the point of this study*
AIM_REF_AGENT	*its goal*
US_PREVIOUS_AGENT	*the approach given in* <REF SELF=YES/>
REF_AGENT	*the paper*
THEM_PRONOUN_AGENT	*they*
THEM_AGENT	*his approach*
GAP_AGENT	*none of these papers*
GENERAL_AGENT	*traditional methods*
PROBLEM_AGENT	*these drawbacks*
SOLUTION_AGENT	*a way out of this dilemma*
TEXTSTRUCTURE_AGENT	*the concluding chapter*

Figure 5.7: Types of Agents (Feature Ag-1)

or deictic anaphora (*"this approach"*). Such forms are clustered together into ambiguity classes with a lower confidence level: REF_US_AGENT, THEM_PRONOUN_AGENT, AIM_REF_AGENT and REF_AGENT. The 168 agent patterns we use are given in appendix D.2 (p. 339).

It is possible that the agent patterns appear in a position other than subject position, in which case they still carry some information, even if they are not the agents. In this case, they are reported under the `Formu` feature; the 13 `Ag-1` classes are thus added as values to the 20 `Formu` types, resulting in a total of 33 values for the feature `Formu`.

Actions (`Ag-2`**):** This section discusses a classification of verbs into semantic classes which assist Argumentative Zoning. Verbs are not frequently used in NLP experiments, in contrast to nouns. Klavans and Kan (1998) are an exception in that they use verbal classes for document classification according to text type and event. They use Levin's (1993) alternation classes and found that occurrence of communication verbs and agreement verbs correlated with text type and/or event (e.g. opinion pieces vs. documents about legal cases or mergers). In contrast to ours their work looks at large text units (documents) whereas we are interested in using verb information per sentence.

Negation is a phenomenon which should be recognized—there is an essential difference between the action of "*does not solve*" and "*solves*". Not understanding this difference would deliver the opposite interpretation to the one intended and thus undermine the core of our shallow selective text-understanding task. We heuristically determine if a verb is negated or not.

We use a manually constructed verb lexicon for verb classification, cf. figure 5.8. The semantics of these verbs mainly comes from the argumentative moves defined in section 3.2, which are concerned with similarity, contrast, competition, presentation, argumentation and textual structure. We will describe them in the following:

PRESENTATION_ACTIONs include verbs like *present, report, state*, often referred to as communication verbs. Myers (1992) performs a pragmatic analysis of such verbs in combination with knowledge claims; Thomas and Hawes (1994) analyze such verbs in medical texts, and Thompson and Yiyun (1991) look at presenting verbs in the context of citations and positive/negative evaluation.

Explicit signalling of the research process ahead is another frequent phenomenon. Research goals can be introduced by stating an interest in a certain research question (INTEREST_ACTION; *"aim to", "attempt to"*) or by stating some involvement or affect towards the solving of a problem (AFFECT_ACTION; *"seek", "want"* and *"wish"*). Direct argumentation verbs (ARGUMENTATION_ACTION) include *"argue"*, *"disagree"* and *"object to"*.

In statements about problem-solving processes (cf. section 3.2.4), verbs of problem introduction abound (PROBLEM_ACTION). These are the ones which state that a situation is problematic. Examples for verbs in this class are *"fail"*, *"degrade"*, *"overestimate"*, and *"waste"*. If there is a lack or need of something, this often has the same semantics (NEED_ACTION; verbs like *"lack"*, *"need"*, *"be void of"*). Problem-solving actions (SOLUTION_ACTION) indicate that a solution has been found (*"solve"*, *"circumvent"*, *"mitigate"*). Contrast between approaches might be expressed overtly with CONTRAST_ACTION verbs like *"clash"*, *"contrast with"*, and *"distinguish"*. BETTER_SOLUTION_ACTIONs state that one solution solves the problem better than another. Examples include *"outperform"* and *"increase"*). Comparison actions (COMPARISON_ACTION) draw a direct comparison between own and rival approaches (*"compare with"*, *"test against"*). Display-of-awareness verbs (AWARENESS_ACTION) like *"know"* can be used to show that there is a gap in the literature, or that the own task is done for the first time, as in the phrase *"we know of no approach which… "*.

There is a range of ways of stating that aspects of a solution are borrowed from another one. CONTINUATION_ACTIONs include *"base on"*, *"borrow"*, *"take as our starting point"*. Another way of stating research continuity is to state the simple use of another solution (USE_ACTION; *"employ"*, *"use"*); this can be combined with a statement of which aspect of the other solution was changed (CHANGE_ACTION; *"transform"*, *"change"*). In some cases, similarity between solutions (SIMILARITY_ACTION) is stated as a signal for intellectual ancestry (*"resemble"*, *"be similar"*).

There are generic, prototypical RESEARCH_ACTIONS which can be predicted from the discipline (e.g. *"analyze"*, *"conduct"*, *"define"* and *"observe"*). Many other such actions are document specific, describing the creative inventive step of the article. They can therefore not be predicted. We also look for TEXTSTRUCTURING_ACTIONS such as *"outline"* and *"structure"*.

The action lexicon contains a total of 365 verbs; it is reproduced in appendix D.3 (p. 343). This lexicon also contains phrasal verbs and longer idiomatic expressions (e.g., "*have to*" is a NEED_ACTION; "*be inspired by*" is a CONTINUE_ACTION).

Ag-2: Action Types

Type	Example	Type	Example
AFFECT	*we hope to improve our results*	NEED	*this approach, however, lacks...*
ARGUMENTATION	*we argue against a model of*	PRESENTATION	*we present here a method for...*
AWARENESS	*we are not aware of attempts*	PROBLEM	*this approach fails...*
BETTER_SOLUTION	*our system outperforms ...*	RESEARCH	*we collected our data from...*
CHANGE	*we extend <CITE/>'s algorithm*	SIMILAR	*our approach resembles that of*
COMPARISON	*we tested our system against...*	SOLUTION	*we solve this problem by...*
CONTINUATION	*we follow Sag (1976) ...*	TEXTSTRUCTURE	*the paper is organized...*
CONTRAST	*our approach differs from ...*	USE	*we employ Suzuki's method...*
FUTURE_INTEREST	*we intend to improve ...*	COPULA	*our goal is to...*
INTEREST	*we are concerned with ...*	POSSESSION	*we have three goals...*

Figure 5.8: Types of Actions (Feature Ag-2)

5.3. A Prototype System

We have implemented a statistical and a symbolic Argumentative Zoning prototype system. Our corpus is encoded in XML (eXtensible Markup Language). XML, which provides a universally recognized platform for data representation, also allows the definition of customized semantic labels. This helps in the encoding of the document's semantics, rather than just layout information.

Processing is based on a Unix pipeline. Different phases of the pipeline add different information (in the form of XML elements and attributes) to an intermediate XML representation of the document.

The corpus collection and conversion work was initially conducted in summer 1996 by myself and Byron Georgantopoulos, as a joint effort to provide data for different projects with the summarization of academic papers. The final conversion pipeline uses a different implementation, based on the TTT tools available from the HCRC Language Technology Group (Grover et al., 1999). A version of the corpus collected during the current work is now available from Tipster SUMMAC (1999).

5.3.1. Corpus Encoding

The first step in the endeavour to collect a corpus is the design of a corpus encoding format. On the one hand, one wants to encode as much of the original information as possible. It is desirable to standardize the encoding such that it expresses the document semantics, and abstract away from the physical and typesetting information the data comes mixed with. Our XML encoding provides rich information about structural information, e.g. sentences, paragraphs and division structure. The author-written summary is marked as such. Additional mark-up includes titles, headlines, sentences, formal citations, author names and the reference list at the end.

Another criterion is data consistency. LaTeX, the source encoding of our data, is unfortunately a very powerful language, offering a wide range of syntactic constructs. Therefore, similar document semantics might be expressed syntactically differently in different papers (in the worst case even in the same paper), but our encoding should treat them alike.

The two goals of information-richness and data consistency often work against each other. For example, citation handling can be automated in LaTeX with the command `\cite`, but authors could decide to just type the author name and year. Similarly, cross references can be expressed with the command `\cref`; however, some authors prefer to directly state the actual numerical cross reference. Ideally, our representation should mark up both facts: the fact that the string "2.2" refers to a cross reference (type information), and that its identity is "2.2" (string information). However, if authors used `\cref`, we do not have the identity of the string (as it is only determined at runtime of the LaTeX system), whereas the textual variant does not give us the information that the string's type is a cross reference. We decided to use the structural information in preference to the string information—in general, we preferred consistency above informativeness in conflict cases. This means that in our encoding type/structural information is captured consistently, however sometimes at the price of a small information loss.

There are some design decisions which were influenced by the fact that corpus collection took place in collaboration with a project that was less interested in structural features than the current thesis is. The loss of captions is an example of a wrong but non-reversible design decision. It was decided in an early processing stage to remove captions of images and tables. Part of the reason for doing so was data consistency, as captions cannot always be determined automatically. We realized only later that captions often contain information particularly useful for summarization.

It was also decided to remove footnotes, a decision which we do not regret. As textual material contained in footnotes is marked by the author as less central to the overall flow of the argumentation, a summarization system might decide to ignore it. However, for a full representation of a paper, which is not attempted here, footnote text should be kept. Footnote information might be important if one tries to assess relative importance of citations, as some marginal references appear only in footnotes.

Appendix B.1 shows the example paper in XML format after preprocessing, before feature determination. We will now describe in detail how the document semantics of the papers are encoded in XML. Appendix A.1 gives the DTD (*Document Type Definition*) for our corpus. A DTD is a BNF-style description of the hierarchical and logical structure of an XML file. As DTD syntax is cryptic and might be unknown to the reader, the following list explains the components in English.

- *Title, authors and bibliographic information* is marked by elements <TITLE>, <AUTHOR>, <AUTHORS>, <FILENO>, <APPEARED>

- A *unique citation form* is assigned to the document and marked as <REFLABEL>. The citation form is a mnemonic label consisting of name and date, and of an optional letter to distinguish references which are ambiguous within the corpus, if needed. The provision of unique citation forms is important for disambiguation of citations (e.g. for clustering of documents by bibliographic chaining).

- *Divisions:* The hierarchical embedding of text segments is encoded by the <DIV> element, which is recursive. The DEPTH attribute indicates the depth of embedding of a division. Each division must start with a <HEADER> element.

- *Headlines* are marked as <HEADER> elements, containing (tokenized and POS-tagged) text.

- *Appendices:* If appendices occur at some other place in the paper, they are physically moved to the point directly before the reference list. They do not receive preferential treatment; instead, they are treated like all other divisions. The fact that they are appendices can only be read off the headline.

- *Paragraphs:* Paragraphs are marked as element <P>.

- *Sentences:* Sentences are separated and marked as <S> elements. This is important, as sentences are the base level selection and analysis unit.

- *Abstract:* The abstract is marked as <ABSTRACT>, and sentences of the abstract are marked as elements <A-S>.

- *Correspondences* between abstract and document sentences are marked by a double link: attribute DOCUMENTC in abstract sentences, and attribute ABSTRACTC in document sentences. This correspondence is determined by a similarity finding algorithm and manual checking (cf. section 4.1.2.2).

- *Images*: Images are removed and the place is marked by an empty <IMAGE/> element. In cases where the LaTeX `verbatim` environment was used, it was manually decided whether or not such material counts as an image or as text.

- *Tables:* Tables are removed (often automatically, sometimes manually), and their position is marked by an empty <IMAGE/> element.

- *Bullet point lists*: Bullet items are manually marked up as such by as an optional attribute of sentences (TYPE=ITEM). Paragraphs as well as sentences can be bullet items.

- *Cross references:* Cross references are automatically or manually marked as empty elements <CREF/>. Manual effort was needed to find corresponding numbers (*"figure 1"*) and replace them by <CREF/>. For consistency reasons, we erased the numbers themselves, as they were not in all cases available.

- (Linguistic) *example sentences* are manually marked up as <EXAMPLE>.

- *Equations:* any kind of mathematical formula that could not be expressed in ASCII was manually (sometimes automatically) replaced by empty element <EQN/>. There might be cases of inconsistencies with formulas like P(A,B) which might be expressed as ASCII or as as <EQN/>, depending on whether the author used the LaTeX math mode or not.

- *Bibliography list:* During bibliographic processing, the bibliography list at the end is marked as <REFERENCE>. It consists of single <REFERENCE> items, each referring to a formal reference. Within these reference items, names of authors are marked as <SURNAME> elements, and years as <YEAR>.

- Formal *citations:* During preprocessing, formal citations are marked automatically as <REF> wherever the latex command `\cite` was used; otherwise, bibliographic processing automatically marks them. Self references are automatically recognized by comparing the names of the author(s) of the paper with all author names associated with the reference. They are marked using the attribute `SELF`.

- *Names of other authors:* Author names occurring in running text without a data are marked up as <REFAUTHOR> during the bibliographic processing step.

- *Formulaic expressions:* if formulaic expressions are recognized during feature determination, they are marked as <FORMULAIC>, with an attribute specifying the formulaic expression type.

- *Agents:* if prototypical agents are recognized during feature determination, they are marked as <AGENT>, with an attribute specifying the agent type.

- *Actions:* if prototypical actions are recognized during feature determination, they are marked as <ACTION>, with an attribute specifying the action type.

5.3.2. Preprocessing

We chose all papers from CMP_LG which fulfilled the following criteria:

- *Date:* We collected all papers put on the archive between 04/94 and 05/96.

- *Format:* The LATEXsource had to be available (in addition to a PostScript version of the paper), and the paper had to pass our conversion pipeline automatically; about 20% did not pass or showed too many errors such that manually correction would have been too inefficient.

- *Abstract:* The papers had to have an abstract.

- *Type:* The papers had to be published in the proceedings of the main or student session, or of a workshop of one of the following conferences: *The Annual Meeting of the Association for Computational Linguistics* (ACL), *The Meeting of the European Chapter of the Association for Computational Linguistics* (EACL), the *Conference on Applied Natural Language Processing* (ANLP), and the *International Conference on Computational Linguistics* (COLING).

As a result of being published in conference or workshop proceedings, the length of the papers was restricted by the publishing rules of the corresponding proceedings. The PostScript versions of the papers are between 3 and 10 pages long; most papers are between 6 and 8 pages long.

The corpus consists of 333,634 word tokens (counting punctuation as a token), the average number of tokens per paper was 4170, ranging from 1301 to 7635 tokens. The total number of document sentences is 12471, average per paper is 156, ranging from 45 to 322. The total number of abstract sentences is 356, average per paper is 4.5, ranging between 2 and 13 sentences.

Our papers' original format was LaTeX source. The first processing steps are a text format conversion from LaTeX source to XML format: LaTeX source is converted into HTML with the program Latex2html (Drakos, 1994; Latex2Html, 1999); the resulting HTML format is then transformed into XML format with a range of `perl` scripts. The pipeline is fully implemented, but some manual correction effort is still needed as the pipeline works imperfectly. This is due to the difficulty of deducing semantic markup from layout information:

- LaTeX is a rich language, offering a wide range of syntactic constructs which are difficult to standardize.
- Latex2html has certain weaknesses, e.g. the inability to deal with LaTeX macros.
- Our XML encoding contains some information which no automatic processing can perform yet (e.g. the determination of (linguistic) example sentences in text).

As a result of the preprocessing/conversion step, text is in a format in which paragraphs are marked up, but words are not separated yet, and sentences are not marked either. The next step is a pipeline to provide linguistic mark-up, and to determine the values of the features, as described in the next section.

5.3.3. Feature Determination

We will now describe how features are automatically determined in running text. Figure 5.9 shows the single steps of processing; it also shows which feature values each processing step provides.

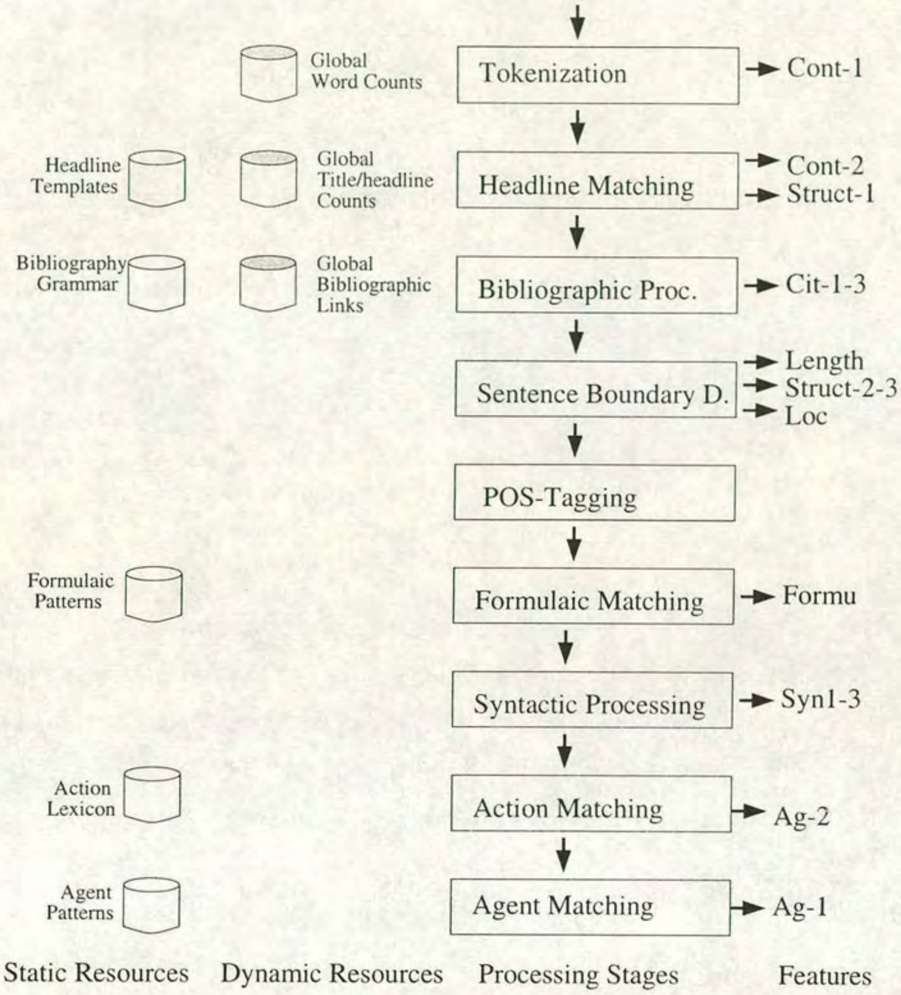


Figure 5.9: Feature Determination Steps

We will describe the practical algorithm for determining the value for each feature. We will also give contingency tables for each feature. Whenever 100% correctness of a feature cannot trivially be assumed, we have also performed an evaluation of the reliably of the heuristics used.

5.3.3.1. Tokenization

Tokenization is the first step in our feature determination pipeline. We used software distributed as the TTT (Text Tokenization) System by the HCRC Language Technology Group Grover et al. (1999). The tokenization grammar was written by Claire Grover; it performs separation of word tokens from the ASCII stream. Tokenization provides information needed for feature `Cont-1`.

Cont-1	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
0	129	193	658	537	1801	7517	172	11007
1	78	33	62	59	213	919	51	1415
Total	207	226	720	596	2014	8436	223	12422

Figure 5.10: Contingency Table for *tf/idf* Feature (Cont-1)

In order to calculate the *tf/idf* score $w_{i,j}$, we use the following formula:

$$w_{i,j} = f_{i,j} * log(\frac{N}{n_i})$$

$w_{i,j}$:	weight for a word k_i in document d_j
n_i:	number of documents containing word k_i
$f_{i,j}$:	frequency of word k_i in document d_j
N:	number of documents in collection

The n top-scoring words according to the *tf/idf* method are chosen as content words; sentence scores are then computed as a weighted count of the content words in a sentence, meaned by sentence length. The m top-rated sentences obtain score 1, all others 0. We received best results with $n = 10$ and $m = 40$. The contingency table is given in figure 5.10.

5.3.3.2. Headline Matching

Headlines are used for two features in our implementation, Struct-3 and Cont-2 (cf. figures 5.11 and 5.12 for contingency tables).

For the feature Struct-3, we pattern match the headline against 89 patterns which correspond to 16 prototypical headlines. If there is a hierarchical nesting of divisions, the headlines of the deeper embedded sections are considered first. If no pattern matches, the value Non-Prototypical is assigned. We can see that more than 45% of all sentences (5576/12422) are not covered by prototypical section headings, i.e. they cannot be easily associated with a rhetorical section. This is in agreement with our argumentation in section 3.1.

Cont-2 is the title method. In our implementation, title scores are determined as the mean frequency of n (or less) title word occurrences (excluding stop-list words). If the title contains more than n non-stoplist words, the n top-scoring words according to the *tf/idf* method are chosen. Again, the m top-scoring sentences receive the value 1, all other sentences 0. Best results in this case were received with n=10 and m=18. One

Struct-3	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
Introduction	102	48	382	185	434	368	89	1608
Implementation	1	18	5	24	262	791	9	1110
Example	1	10	16	27	112	459	6	631
Conclusion	62	14	4	39	27	454	3	603
Result		2		7	33	480	6	528
Evaluation	4	3	1	10	27	427	5	477
Solution	1	7	18	21	78	280	4	409
Experiment		11	4	9	19	306	1	350
Discussion	4	4	3	19	19	277	7	333
Method	1	7	4	26	40	163	6	247
Problems	3	7	14	9	20	95	1	149
Related Work	2	3	5	41	75	19	1	146
Data		1			6	102		109
Further Work					1	71		72
Problem Statement	1	1	5	1	2	42		52
Limitations		1	1	4	9	5	2	22
Non-Prototypical	25	89	258	174	850	4097	83	5576
Total	207	226	720	596	2014	8436	223	12422

Figure 5.11: Contingency Table for Headline Feature (Struct-3)

Cont-2	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
0	128	161	571	437	1546	6201	178	9222
1	79	65	149	159	468	2235	45	3200
Total	207	226	720	596	2014	8436	223	12422

Figure 5.12: Contingency Table for Title Feature (Cont-2)

variant of the method additionally takes words occurring in all headlines into account, but we received better results using only title words.

5.3.3.3. Bibliographic Processing

Bibliographic processing determines information important for features Cit-1, Cit-2 and Cit-3. For the bibliographic processing we used a grammar written in the specific syntax of the program fsgmatch, which is provided with TTT. The grammar was originally written by Colin Mattheson; we changed it to suit our purposes. Bibliographic processing includes the following processing:

- The reference list at the end is parsed according to a grammar for bibliographic entries. This grammar anticipates typical citation styles. Author names and dates are marked up as such, and a <REFLABEL> element is constructed for each bibliographic entry, based on this information.

- The last names of all cited authors are put into a special lexicon, and the body of the text is searched in a second pass for these names.
- If the last names appear in a typical citation context (i.e. with a year, with or without brackets), they are wrapped as XML-elements <REF>. If they occur on their own, they are marked as <REFAUTHOR>. If the LaTeX command `\cite` was used, nothing needs to be done, as <REF> elements are already marked.
- Each reference is checked for overlap of one of the cited authors with the authors of the article (by comparison of all cited authors with the <AUTHOR> field). If such an overlap is determined, the reference is marked as a *self citation*. That means that the common abbreviation *"et al."* in citations in running text is resolved into all cited author names. This piece of information is only available from the reference list (even for human interpretation).

After all <REF> and <REFAUTHOR> in a sentence have been marked up, `Cit-1` reports the existence of either of these (if a sentence contains both <REF> and <REFAUTHOR>, the value `Citation` is chosen, cf. contingency table in figure 5.13). `Cit-2` reports whether or not a reference is a self reference, cf. contingency table in figure 5.14). In cases where a self citation and a non-self-citation appear in one sentence, the self citation is given preference. `Cit-3` gives the location of the reference(s) in order to distinguish authorial from parenthetical citations, cf. contingency table in figure 5.15. In cases of more than one reference in a sentence, "Citation-Beginning" is given preference over both "Citation-Middle" and "Citation-Ending", and "Citation-Ending" is given preference over "Citation-Middle".

`Cit-1`	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
Citation	17	163	79	96	482	290	5	1132
Author name	7	18	1	52	128	71	2	279
No Citation	183	45	640	448	1404	8075	216	11011
Total	207	226	720	596	2014	8436	223	12422

Figure 5.13: Contingency Table for Citation Feature (`Cit-1`)

`Cit-2`	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
Citation to Other Work	12	112	75	78	391	240	3	911
Citation to Own Previous Work	5	51	4	18	91	50	2	221
No Citation	190	63	641	500	1532	8146	218	11290
Total	207	226	720	596	2014	8436	223	12422

Figure 5.14: Contingency Table for Citation Type Feature (`Cit-2`)

`Cit-3`	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
Citation-Beginning		11	7	16	110	24		168
Citation-Middle	5	61	13	50	153	97		379
Citation-Ending	12	91	59	30	219	169	5	585
No Citation	190	63	641	500	1532	8146	218	11290
Total	207	226	720	596	2014	8436	223	12422

Figure 5.15: Contingency Table for Citation Location Feature (`Cit-3`)

5.3.3.4. Sentence Boundary Disambiguation

Determining sentence boundaries is important for each single feature, as sentences are our units of classification. However, some feature values can be determined directly after this step, namely the features `Length` (Sentence Length), `Struct-1` (Position in Section), `Struct-2` (Position in Paragraph), and `Loc` (Absolute Location).

We use the sentence boundary disambiguator provided with TTT (`ltstop`) and add some perl code to assign identifiers to sentences. We also had to write some code to mend some of the systematic mistakes the automatic method performed. We fixed such errors with symbolic rules. For example, in the following sentence the system failed to recognize a sentence break after a variable consisting of a single letter:

> <S> [...] *we make use of parameters ("dependency parameters") <EQN/> for the probability, given a node h and a relation r, that w is an r-dependent of* **h. Under** *the assumption that the dependents of a head are chosen independently from each other, the probability of deriving c is:< /S>*
>
> (S-190, 9408014)

Figures 5.16, 5.17, 5.18 and 5.19 give the contingency tables for features `Length`, `Struct-1`, `Struct-2` and `Loc`, respectively. For feature `Length`, the value 0 means that the sentence was shorter than a fixed threshold (here: 15 tokens including punctuation), 1 means that it was longer than the threshold.

`Length`	Aim	Bas	Bkg	Ctr	Oth	Own	Txt	Total
0	31	41	190	105	554	2507	102	3530
1	176	185	530	491	1460	5929	121	8892
Total	207	226	720	596	2014	8436	223	12422

Figure 5.16: Contingency Table for Sentence Length Feature (`Length`)

`Struct-1`	Aim	Bas	Bkg	Ctr	Oth	Own	Txt	Total
First_third	24	23	195	104	366	1174	22	1908
Second_third	36	48	190	169	736	2518	25	3722
Last_third	22	25	64	118	307	1600	27	2163
First_sentence	57	35	92	19	89	332	32	656
Last_sentence	15	14	7	25	51	487	40	639
Second_or_third_sentence	33	43	129	55	205	793	26	1284
Second-last_or_third-last_sentence	20	38	43	106	260	1532	51	2050
Total	207	226	720	596	2014	8436	223	12422

Figure 5.17: Contingency Table for Section Structure Feature (`Struct-1`)

`Struct-2`	Aim	Bas	Bkg	Ctr	Oth	Own	Txt	Total
Initial	117	92	267	135	601	2532	73	3817
Medial	56	87	306	289	971	3779	68	5556
Final	34	47	147	172	442	2125	82	3049
Total	207	226	720	596	2014	8436	223	12422

Figure 5.18: Contingency Table for Paragraph Feature (`Struct-2`)

For the feature `Struct-1`, the section is separated into three equally sized portions (measured in sentences). In those cases where a sentence is in a specific position within the section, the resulting values are "overwritten" over the tri-section values.

As far as feature `Struct-2` is concerned, if a paragraph contains only one sentence, that sentence receives the value `Initial`. If a paragraph contains only two sentences, the first sentence receives the value `Initial` and the second the value `Final`.

Values of the feature `Loc` are determined by dividing the sentence number of the document by 20, and assigning values according to the diagram in figure 5.5. Document areas corresponding to A, B, C, D, I, J are one twentieth of the document in length, E, G, H one tenth, and value F two fifth.

Loc	Aim	Bas	Bkg	Ctr	Oth	Own	Txt	Total
A	51	18	261	69	167	70	22	658
B	30	18	114	94	186	146	29	617
C	24	20	83	55	199	216	24	621
D	12	12	82	41	160	289	27	623
E	17	25	60	52	363	682	38	1237
F	7	81	104	178	680	3864	66	4980
G	2	11	12	21	121	1052	10	1229
H	6	19	1	30	62	1130	4	1252
I	23	11		31	43	514	2	624
J	35	11	3	25	33	473	1	581
Total	207	226	720	596	2014	8436	223	12422

Figure 5.19: Contingency Table for Absolute Location Feature (Loc)

5.3.3.5. POS-Tagging

Part of speech tagging provides vital information for complex pattern matching algorithms further on in the pipeline (Formulaic pattern matching, Agent Matching, Action Matching). It is performed using the program ltpos, distributed with TTT and written by Andrei Mikheev. It assigns one of the tags of the BROWN tagset (Francis and Kucera, 1982) to each token in text.

As later processing heuristics depend on the correct determination of finite verbs, we needed to determine the error rate of POS tagging. We manually checked the assignment of finite verbs, i.e. the tags VBP, VBZ and VBD on a random sample of 100 sentences containing finite verbs. We compared the automatic POS-tag with the POS-tag we thought should have been assigned. In the 100 sentences, there were 184 finite verbs, 174 of which the system recognized (recall of 95%). Most of the non-recognition errors were present verbs which the system erroneously tagged as singular or plural nouns. The system erroneously tagged an additional 14 tokens as finite verbs (precision of 93%). These words were mostly past participles in reduced relative clause constructions. We feel that this is a solid tagging performance, stable enough to base our further heuristic processing on it.

5.3.3.6. Formulaic Pattern Matching

We have determined a total of 396 formulaic patterns (cf. appendix D.1). As we use a finite-state replace mechanism, these patterns multiply out to many more actual strings. The lexical group of @TRADITIONAL_ADJECTIVES for example includes 37 adjectives like *classic* or *long-standing*, and this lexical group is contained in 29 patterns. There are 44 different lexical groups (cf. the concept lexicon appendix D.4). Some of the patterns use POS place-holders which are checked against the POS-tags of words in running text.

Additionally, the 168 agent patterns are also considered as formulaic patterns, wherever they do *not* occur as the subject of the sentence. The decision to include these into the Formu feature was explained in section 5.2.2.2.

Pattern matching procedures on such a large scale are slow. We reduce the number of comparisons necessary with a trigger mechanism: only to those sentences containing a trigger (a rare word which covers as many patterns as possible) are searched, and they are searched only for those patterns which do contain the trigger. Triggers are marked by the signal ↑ directly in the pattern.

Figure 5.20 gives the contingency table for Formu. It lists *first occurrence* of a formulaic pattern in the text. The restriction to one value per sentence is necessary for the Naive Bayes classifier.

5.3.3.7. Syntactic Processing

Syntactic processing determines the verbal features (Syn-1, Syn-2, Syn-3) and negation. It also determines the base form of the semantic verb, to be used for feature Ag-2. The first step of the algorithm is the determination of finite verbs in the sentence, information which is made available by the POS-Tagging. The next step is a finite state algorithm which checks left and right context of the finite verb for verbal forms of interest which might make up more complex tenses. Such forms are searched within the assumed clause boundaries, and additionally within a fixed window of 6 to the right of the finite verb. Negation is determined by a simple heuristic that searches for a list of 32 negation-items in the surrounding window of 5 items. The list of negation-items is given in appendix D.4 (p. 345).

The syntactic heuristics can contain errors, either due to errors in our algorithm or due to wrong POS-Tagging. We performed an evaluation on the aforementioned 100 sentences. Counting success and failure on the 174 finite verbs correctly determined by POS-Tagging, we found that the heuristics for negation and modality worked without any errors in our sample (100% accuracy), that there were 2 errors in the tense heuristics (99% accuracy) and 7 errors in the voice heuristics, 2 of which are due to POS-Tagging errors (where a past participle was not recognized in a passive sentence). The remaining 5 voice errors correspond to a 98% accuracy. Voice errors are particu-

Formu	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
GAP_INTRODUCTION				1	1	6		8
OUR_AIM	6					2		8
DEIXIS	1	1		2	3	45	3	55
SIMILARITY	2	3	1	1	7	4		18
COMPARISON		1		9	6	6		22
CONTRAST			11	41	17	100		169
DETAIL	1	1			1	36		39
METHOD	28	17	16	14	57	117	10	259
PREVIOUS_CONTEXT		1			2			3
FUTURE					1	20		21
AFFECT						6		6
PROBLEM			10	3	12	62		87
SOLUTION		1	7	4	29	81	3	125
IN_ORDER_TO	2	1	3	1	10	51		68
POSITIVE_ADJECTIVE	27	23	86	88	185	936	16	1361
NEGATIVE_ADJECTIVE	11	9	65	133	143	680	2	1043
THEM_FORMULAIC					4	1		5
AIM_REF_AGENT	13	2	20	7	26	121	2	191
TEXTSTRUCTURE_AGENT	2	3			5	21	83	114
GAP_AGENT				1		3		4
REF_AGENT	9	27	31	43	138	468	44	760
GENERAL_AGENT		2	19	14	50	49	1	135
THEM_PRONOUN_AGENT	3	2	25	22	56	210	4	322
US_PREVIOUS_AGENT		2			1			3
REF_US_AGENT	59	16	2	8	6	63	6	160
US_AGENT	21	21	40	32	74	959	24	1171
COMPARISON_FORMULAIC		1		9	6	6		22
THEM_AGENT	5	53	16	29	169	86	4	362
—	17	40	364	142	987	4262	21	5833
Total	207	226	720	596	2014	8436	223	12422

Figure 5.20: Contingency Table for Formulaic Expressions Feature (Formu)

larly undesirable, as they have knock-on effects on agent determination. An example for such a voice error is the following sentence (underlined; syntactic information about clause-like units is attached to the respective finite verb):

> **At the point where John** <FINITE TENSE=“PRESENT” VOICE=“ACTIVE” MODAL=“NOMODAL” NEGATION=“0” ACTIONTYPE=“0”> **knows** </FINITE> **the truth** <FINITE TENSE=“PRESENT_PERFECT” VOICE=“PASSIVE” MODAL=“NOMODAL” NEGATION=“0” ACTIONTYPE=“0”> **has** </FINITE> **been processed, a complete clause**

> <FINITE TENSE="FUTURE_PERFECT" VOICE="ACTIVE" MODAL="NOMODAL" NEGATION="0" ACTIONTYPE="0"> **will** </FINITE> **have been built.** (S-15, 9502035)

This error was caused by the fact that the threading of auxiliaries in our algorithm did not foresee this particular combination of voice and tense. Note that apart from the voice error, everything else is correct. The high level of accuracy achieved in the syntactic processing is not a trivial result, as the processing encompasses complicated combinations of voice, complex tenses and modal auxiliaries, as exemplified by the following corpus example:

> **The actor** <FINITE TENSE="PRESENT_CONTINUOUS" VOICE="ACTIVE" MODAL= "NOMODAL" NEGATION="0" ACTIONTYPE="0"> **is** </FINITE> **always running and** <FINITE TENSE="PRESENT" VOICE="ACTIVE" MODAL="NOMODAL" NEGATION="0" ACTIONTYPE="AFFECT"> **decides** </FINITE> **at each iteration whether to speak or not (according to turn-taking conventions); the system** <FINITE TENSE="PRESENT" VOICE="ACTIVE" MODAL="NOMODAL" NEGATION="NEGATED" ACTIONTYPE="NEED"> **does** </FINITE> **not need to wait until a user utterance** <FINITE TENSE="PRESENT" VOICE="PASSIVE" MODAL="NOMODAL" NEGATION="0" ACTIONTYPE="RESEARCH"> **is** </FINITE> **observed to invoke the actor, and** <FINITE TENSE="PRESENT" VOICE="ACTIVE" MODAL="MODAL" NEGATION="NEGATED" ACTIONTYPE="0" > **need** </FINITE> **not respond to user utterances in an utterance by utterance fashion.** (S-137, 9407011)

Contingency tables for features Syn-1, Syn-2 and Syn-3 can be found in figures 5.21, 5.22 and 5.23, respectively.

It can be the case that more than one finite verb occurs in a sentence, but our main classification method allows only one feature value per feature. All other factors being equal, we prefer verbs in the beginning of the sentence, for two reasons: in the case of coordination, we assume that the more important material might have been presented first; in the case of subordination, we assume that matrix verbs carry more information with respect to meta-discourse. We choose the values associated with the first verb for which Ag-1 and Ag-2 returns a non-zero value, or, if not applicable, those for which Ag-1 returns a non-zero value, or, if not applicable, those for which Ag-2 returns a non-zero value. Failing all of these alternatives, we chose the values of the first verb in the sentence.

Syn-1	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
Active	175	149	407	446	1214	5079	168	7638
Passive	20	62	109	76	363	1286	39	1955
NoVerb	12	15	204	74	437	2071	16	2829
Total	207	226	720	596	2014	8436	223	12422

Figure 5.21: Contingency Table for Voice Feature (Syn-1)

Syn-2	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
Present Tense	134	158	444	410	1265	5033	177	7621
Present Continuous		4	8	6	18	99	1	136
Past Tense	15	35	23	66	182	819	6	1146
Past Continuous					2	7		9
Past Perfect					1	7		8
Present Perfect	35	10	33	27	88	185	3	381
Future	11	4	8	13	21	211	20	288
Future Continuous						3		3
Future Perfect						1		1
NoVerb	12	15	204	74	437	2071	16	2829
Total	207	226	720	596	2014	8436	223	12422

Figure 5.22: Contingency Table for Tense Feature (Syn-2)

Syn-3	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
Non_Modal	186	195	422	462	1437	5545	200	8447
Modal	9	16	94	60	140	820	7	1146
NoVerb	12	15	204	74	437	2071	16	2829
Total	207	226	720	596	2014	8436	223	12422

Figure 5.23: Contingency Table for Modal Feature (Syn-3)

5.3.3.8. Action Matching

Action Matching determines the value of feature Ag-2 (contingency table in figure 5.24). It relies on the processing done in the syntactic processing, which determines the *semantic* verb along with the *finite* verb, and also determines whether or not negation was present. Depending on the tense, semantic and finite verb can be the same word. Our algorithm thus performs a distinction between auxiliary and full verb sense for *"have", "be"* and *"do"*. The base form of the semantic verb is determined and it is checked if it is contained in the action lexicon.

Ag-2	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
Positive								
AFFECT		2	5	3	11	68		89
ARGUMENTATION	4	2	2	6	26	62	6	108
AWARE				1	1	2		4
BETTER_SOLUTION	1	1	3	9	5	38		57
CHANGE	4	11	13	11	58	187	5	289
COMPARISON	3	1	2	8	5	50	2	71
CONTINUE	2	21	8	1	20	54		106
CONTRAST		1		5	1	19	1	27
COPULA	24	28	156	112	410	1675	6	2411
FUTURE_INTEREST				1	4	21		26
INTEREST	35	4	27	19	56	209	11	361
NEED		2	19	21	42	186		270
POSSESSION	2	2	25	16	43	204		292
PRESENTATION	78	25	38	39	196	533	105	1014
PROBLEM	1		10	26	18	86	1	142
RESEARCH	11	29	47	38	181	831	17	1154
SIMILAR		10	2	2	8	17		39
SOLUTION	11	16	31	50	135	455	11	709
TEXTSTRUCTURE	1	3	2	3	14	66	27	116
USE	3	22	26	21	98	341	3	514
Negated								
AFFECT			2		1	10		13
ARGUMENTATION			2		2	12		16
AWARE				3		1		4
BETTER_SOLUTION			1	1		1		3
CHANGE			2	3	1	10		16
COMPARISON				2		1		3
CONTINUE			1	3		1		5
CONTRAST						1		1
COPULA	3		18	28	34	209		292
FUTURE_INTEREST						1		1
INTEREST				4	1	18		23
NEED			1	4	5	26		36
POSSESSION			5	3	3	46	1	58
PRESENTATION			3	4	2	17		26
PROBLEM				2	1	8		11
RESEARCH			4	5	3	53		65
SOLUTION			4	13	4	46		67
USE			2	5	2	14		23
0	24	46	259	124	623	2857	27	3960

Figure 5.24: Contingency Table for Action Feature (Ag-2)

If the base form is found in the lexicon, its Action Type is returned; otherwise ActionType 0 is returned (examples for this can be seen in the example sentences on p. 209, where no negation was detected, and where the only two Actions recognized were a (negated) NEED_ACTION—*"the system does not need to wait"* and a (passive) RESEARCH_ACTION—*"a user utterance is observed"*).

In our sample of 100 sentences containing finite verbs, there were no errors introduced in the action type determination step. Appendix B.7 (p. 300) gives an impression of the output of our algorithm on the example article. Recognized actions are shown in light blue boxes; the table on p. 301 gives the corresponding action types.

5.3.3.9. Agent Matching

Agent Matching determines the value of feature `Ag-1` (contingency table in figure 5.25). The algorithm is as follows:

1. Start from the next (initially, the first) finite verb in the sentence;

2. Search for the agent either as a by-PP to the right, or as a subject-NP to the left, depending on the voice associated with the finite verb. The search algorithm tries to stay within the clause that belongs to the finite verb, i.e. it will not cross assumed clause boundaries (e.g. commas or other finite verbs).

3. If one of the Agent Patterns matches within that area in the sentence, return the Agent Pattern and its Agent Type. Else return Agent 0.

4. Repeat Steps 1, 2, 3 until there are no more finite verbs left.

We first evaluated the correctness of the algorithm by randomly taking 100 sentences which contain agent patterns. These 100 sentences contained 111 agents. Apart from erroneous voice determination (cf. section 5.3.3.7), errors could also potentially be introduced by our heuristic for clauses, which never steps over commas and is stopped by appositions, for example.

But in 105 of our sample cases, the agent pattern was syntactically correct: the pattern was matched as prescribed in the pattern, and the matched string agent covered the entire subject of the sentence (active case) or the by-PP with the agent-interpretation (passive case). In 5 of the 111 sentences, the pattern was only *part* of a subject NP (typically the NP in a post-modifying PP), as in the following examples (recognized patterns underlined):

the relations in the models (S-131, 9408014)
the problem with these approaches (S-12, 9504017)

We argue that these cases should not be counted as errors, as they still give an indication of which type of agents the NP should be associated with. In the one sentence with a complete error, this error was due to a mistagging at the POS-Stage (100% precision). No agent pattern that should have been identified was missed (100% recall). Appendix B.7 also shows the output of the agent recognition for the example paper (pink boxes).

Ag-1	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
US_AGENT	107	85	53	71	114	1456	93	1979
OUR_AIM_AGENT	10			1		5		16
THEM_AGENT		24	9	56	224	59		372
THEM_PRONOUN_AGENT	2		31	24	57	232	1	347
GENERAL_AGENT		1	13	15	28	34	1	92
US_PREVIOUS_AGENT		2		3	37	10		52
REF_AGENT	10	22	20	56	95	374	9	586
REF_US_AGENT	34	3	2	3	1	20	4	67
AIM_REF_AGENT	7		10	1	9	42		69
TEXTSTRUCTURE_AGENT	2	1			4	6	59	72
GAP_AGENT				5		3		8
SOLUTION_AGENT		1	3	5	14	45	3	71
PROBLEM_AGENT			6	2	8	60		76
—	35	87	573	354	1423	6090	53	8615
Total	207	226	720	596	2014	8436	223	12422

Figure 5.25: Contingency Table for Agent Feature (Ag-1)

5.3.4. Statistical Classifiers

There are many machine learning algorithms which are able to classify items into predefined categories, given a set of sentential features. *Supervised* methods take information into account which can only be provided externally (the "correct" answer) whereas unsupervised techniques learn without such external provision of the correct answer.

For our task, we use a set of supervised methods because we only have a small set of data (unsupervised methods typically need much more data), and because supervised learning provides the convenient built-in feature of a simple intrinsic evaluation. Also, we follow Kupiec et al. (1995) who have received good results with a simple classifier for the task of determining global sentence relevance (text extraction).

$$P(s \in S|F_1,\ldots,F_k) = \frac{P(F_1,\ldots,F_k|s\in S)P(s\in S)}{P(F_1,\ldots,F_k)} \approx \frac{P(s\in S)\prod_{j=1}^{k} P(F_j|s\in S)}{\prod_{j=1}^{k} P(F_j)}$$

$P(s \in S|F_1,\ldots,F_k)$: Probability that sentence *s* in the source text is included in summary *S*, given its feature values;
$P(s \in S)$: Probability that a sentence *s* in the source text is included in summary *S* unconditionally; compression rate of the task (constant);
$P(F_j|\ s \in S)$: probability of feature-value pair occurring in a sentence which is in the summary;
$P(F_j)$: probability that the feature-value pair occurs unconditionally;
k: number of feature-value pairs;
F_j: j-th feature-value pair.

Figure 5.26: Kupiec et al.'s (1995) Naive Bayesian Classifier

After having determined a baseline performance with a Naive Bayesian classifier, we then use a more sophisticated method to improve the results of classification. It estimates a better prior probability from the context in terms of the surrounding categories.

5.3.4.1. Naive Bayes

Kupiec et al. were the first to report extraction experiments using a statistical classification method for heuristic combination for determination of global sentence relevance.

Kupiec et al. use the Naive Bayesian Classifier given in figure 5.26. The target value is an estimate of the probability of a sentence to be contained in the abstract, given its feature values. $P(F_j|s \in S)$. In order to estimate this value, probabilities associated with individual events (features) are accumulated; $P(F_j)$ and $P(F_j|s \in S)$ can be estimated from the corpus by raw frequencies. The feature combination applied in a Naive Bayesian model is extremely simple: all conditional probabilities are multiplied.

Kupiec et al. use cross-validation for measuring the success of their classifier: the system extracts sentences from a test document, using a model which was acquired not using any information in the test document. Evaluation can then be measured in precision and recall by the simple criterion of *co-selection* between gold standard and extracted material. Precision gives the percentage of all sentences selected correctly (co-selected with the gold standard) over the total number of sentences selected. Recall gives the percentage of sentences selected correctly (co-selected with the gold standard) over all sentences in the target extract.

In Kupiec et al.'s evaluation, the numerical values for precision and recall are always identical: they use the information of how many gold standard summaries each test document has (though this information would not be available for completely new test documents without abstracts), and their method then extracts the same number of sentences. The method Kupiec et al. chose is a less time consuming way to get an estimation of the cross-over point. (To measure the cross-over point, compression rates are manipulated such that the function of precision and recall can be plotted; the cross-over point of the two functions is then reported.) Another commonly accepted combination of precision and recall is F-measure (van Rijsbergen, 1979).

In (Teufel and Moens, 1997), we report a duplication of Kupiec et al.'s experiment for text extraction. With different data and two types of gold standards, but with similar features to Kupiec et al., we achieved favourably comparable results (cf. the left two columns in figure 5.27). In Kupiec et al.'s case, the best precision and recall of 44% was reached by combining location, cue phrase and sentence length features; in ours, the best result of 68% was achieved using all five features.

	Kupiec et al.		Our replication	
Heuristics	Individual	Cumulative	Individual	Cumulative
Cue Phrases	33%	33%	55%	55%
Location	29%	42%	32%	65%
Sentence Length	24%	44%	29%	66%
tf/idf	20%	42%	17%	67%
Capitalization + *tf/idf*	20%	42%		—
Title		—	21%	68%
Baseline	24%		28%	

Figure 5.27: Results of our Duplication of Kupiec et al.'s (1995) experiment

But here we adapt Kupiec et al.'s Naive Bayesian formula (figure 5.26) for Argumentative Zoning, resulting in the formula given in figure 5.28. As far as the notation is concerned, let us assume we have n features F_0 to F_{n-1}; a feature is then known as F_j, with $0 \leq j < n$. Each of the features F_j has k_j different values V_{jr}, with $0 \leq r < k_j$. There are m target categories C^0 to C^{m-1}; a target category is then known as C^i, with $0 \leq i < m$. In our case, m is 7 (whereas Kupiec et al. perform binary classification; $m = 2$), n is 16, and the k_j vary from 2 for j= 0,1,6 (Cont-1, Cont-2, Length) to 40 for j=15 (Ag-2).

F_4=`Struct-2`	C^0= Aim	C^1= Bas	C^2= Bkg	C^3= Ctr	C^4= Oth	C^5= Own	C^6= Txt	Total
$V_{4,0}$=Initial	$n^0_{4,0}$= 117	$n^1_{4,0}$= 92	$n^2_{4,0}$= 267	$n^3_{4,0}$= 135	$n^4_{4,0}$= 601	$n^5_{4,0}$= 2532	$n^6_{4,0}$= 73	$n_{4,0}$= 3817
$V_{4,1}$=Medial	$n^0_{4,1}$= 56	$n^1_{4,1}$= 87	$n^2_{4,1}$= 306	$n^3_{4,1}$= 289	$n^4_{4,1}$= 971	$n^5_{4,1}$= 3779	$n^6_{4,1}$= 68	$n_{4,1}$= 5556
$V_{4,2}$=Final	$n^0_{4,2}$= 34	$n^1_{4,2}$= 47	$n^2_{4,2}$= 147	$n^3_{4,2}$= 172	$n^4_{4,2}$= 442	$n^5_{4,2}$= 2125	$n^6_{4,2}$= 82	$n_{4,2}$= 3049
Total	n^0= 207	n^1= 226	n^2= 720	n^3= 596	n^4= 2014	n^5= 8436	n^6= 223	N= 12422

Figure 5.29: Contingency Table for Paragraph Feature

$$P(C^i|V_{0,x},...,V_{n-1,y}) = P(C^i)\frac{P(V_{0,x},...,V_{n-1,y}|C^i)}{P(V_{0,x},...,V_{n-1,y})} \approx P(C^i)\frac{\prod_{j=0}^{n-1}P(V_{j,r}|C^i)}{\prod_{j=0}^{n-1}P(V_{j,r})}$$

$P(C^i|V_{0,x},\ldots,V_{n-1,y})$: Probability that a sentence has target category C^i, given its feature values $V_{0,x},\ldots,V_{n-1,y}$, with $0 \leq x < k_0$ and $0 \leq y < k_{n-1}$;

$P(C^i)$: Probability that a sentence has target category C^i (prior);

$P(V_{j,r}|C^i)$: Probability of feature-value pair $V_{j,r}$ occurring with target category C^i;

$P(V_{j,r})$: Probability of feature value $V_{j,r}$ (rth value of Feature F_j);

Figure 5.28: Our Adaptation of Kupiec et al.'s (1995) Naive Bayesian Classifier

The first part of the second formula, $P(C^i)$, is called the *prior* probability, and the second part $\frac{P(V_{0,x},...,V_{n-1,y}|C^i)}{P(V_{0,x},...,V_{n-1,y})}$ is called the *posterior* probability. The first derivation is due to Bayes' Theorem; the second is specific to the Naive Bayesian formula and only legal under the Independence Assumption, i.e. the assumption that all features are statistically independent $(P(F_1,F_2) = P(F_1) * P(F_2))$. If, however, the data show that certain features are statistically dependent on each other—and to a certain degree this can be expected, as it is difficult to define features that are statistically independent—the Naive Bayes method will not result in an absolutely accurate language model.

We will now describe how the conditional probability $P(V_{j,r}|C_i)$ needed for Naive Bayesian classification can be calculated from the contingency tables.

For example, in figure 5.29 (repeated from figure 5.4), the vertical totals $n_{j,r}$

give the occurrence counts of feature value $V_{j,r}$ ($n_{j,r}$ is a short notation for frequency $f(V_{j,r})$); the horizontal totals n^i (or $f(C^i)$) give the occurrence counts of category C_i, and the data cells $n^i_{j,r}$ (or $f(Vj,r,C^i)$) give the number of occurrences of category C_i with feature value V^r_j. N is the number of all items.

Then the desired probability $P(V_{4,1}|C^0)$, i.e. the probability that a sentence displays the feature value $V_{4,1}$ (`Medial`) of feature `Struct-2` , given that the target class of the sentence is AIM, with $i = 0$, $j = 4$ and $r = 1$ (C^0=`Aim`; F_4 = `Struct-2`; and $V_{4,1}$=`Medial`), can be estimated by corpus frequencies $f(Vj,r,C^i)$ and $f(C^i)$ as follows:

$$P(V_{jr}|C^i) = \frac{f(Vj,r,C^i)}{f(C^i)} = \frac{|n^i_{j,r}|}{|n^i|}$$

$$P(Medial|Aim) = P(V_{4,1}|C^0) = \frac{|n^0_{4,1}|}{|n^0|} = \frac{56}{207} = 0.27.$$

It is obvious that for each category C^i and for each feature F_j, the following equality holds:

$$\sum_{r=0}^{k_j-1} P(V_{j,r}|C^i) = 1$$

Naive Bayes estimates the prior probability $P(C^i)$ by simple unigram frequency:

$$P(C^i) = \frac{|n^i|}{|N|}$$

$$P(Aim) = \frac{207}{12422} = 0.0166$$

The reverse probability is $P(C^i|V_{j,r})$: the probability that, on the basis of a given observed feature $V_{j,r}$, the sentence will be classified as C^i. This probability is not used in our calculation.

Naive Bayes estimates the posterior under the independence assumption, but we suspect that our features are not really independent. Intuitively it is clear that they must be related to each other: certain agents, for example GENERAL_AGENT, tend to occur more often in initial locations in the document. This interaction is highly relevant for our experiment. However, it is less obvious which of the features (if any) is directly related to sentence length. A more sophisticated classifier for the posterior probability

$\frac{P(V_{0\_x},...,V_{n-1.y}|C^i)}{P(V_{0\_x},...,V_{n-1.y})}$ does not simply derive the posterior by multiplication of the single probabilities; it determines which features are independent and only multiplies their conditional probabilities. Because of this, we expect better classification results for more sophisticated classifiers. We use two such algorithms, the rule-learning classifier RIPPER (Cohen, 1995, 1996) and a Maximum Entropy-based classifier (Mikheev, To Appear).

5.3.4.2. N-Gram Modelling

In Naive Bayes, not only the posterior, but also the prior is estimated in a very simple manner: it is constant all over the document. However, our model of the typical flow of argumentation predicts typical patterns in our texts. We know that a sentence is more likely to be of category AIM, for example, if the previous sentence was a CONTRAST (introducing a gap), than if the previous sentence was an OTHER sentence (neutrally describing other work)—even if we do not know anything about the features of the sentence to be classified yet. The simple Bayesian classifier, however, does not exploit this fact, i.e. it does not use the context.

N-gram models estimate a more accurate prior by taking the context of a sentence, in terms of surrounding categories, into account. N-gram models are typically used over letters in statistical language processing, but we apply them to *whole sentences* instead. The prior can then be written as $P(C^i_m|C_{m-1},...,C_{m-o})$, for the m-th sentence in the document, instead of $P(C^i)$. The index $o+1$ is called the *order* of the ngram model. A system of order $o+1$ takes o items before the one to be classified into account—a bigram model ($o+1=2$) uses the formula $P(C^i_m|C_{m-1})$.

We ran experiments with N-gram models of order 2, 3 and 4 to estimate the priors, after we first determined the posterior probabilities with the Naive Bayesian model.

$$P(C^i_m|V_{0,x},...,V_{n-1,y}) \approx \; P(C^i_m|C_{m-1},...,C_{m-o}) \; P(C^i)\frac{\prod_{j=0}^{n-1} P(V_{j,r}|C^i)}{\prod_{j=0}^{n-1} P(V_{j,r})}$$

For parameter estimation, we use the Edinburgh Speech Tools Library (Taylor et al., 1999), which use the Viterbi algorithm to maximize the prior probabilities.

5.3.5. Symbolic Rules

We have provided a set of symbolic rules for the determination of the four non-basic categories AIM, TEXTUAL, BASIS and CONTRAST. The rules rely on the sentential features (mainly the Agentivity features), and provide a high-precision, low-recall extraction. For many applications, precision is more important than recall: few sentences might be sufficient, provided that they can be determined with a high level of confidence.

The first step in the algorithm is to assign each sentence scores for each of the categories, whereby several factors are taken into account. These scores are assigned by symbolic rules. Figures 5.30 and 5.31 give the rules for AIM scores. We use two different algorithms for choosing sentences: Method I takes all sentences whose score is above threshold, whereas Method II only takes two sentences who are above threshold: one in the beginning, and one in the end (i.e., one from the introduction and one from the conclusions). Method II is only used for AIM sentences.

We empirically established good threshold values for the scores assigned in the symbolic processing. Figure 5.32 shows how the thresholds relate to precision and recall values achieved with both algorithms on AIM sentences. For high thresholds, Method II achieves a very high precision, albeit a little lower recall than Method I. This might be the method of choice for determining AIM sentences with a high level of certainty. For example, with Method II, the score of 11 gives us a 96% precision and a 23% recall. For lower thresholds (this might be good for determining "second best" candidates), Method I is advantageous, as Method II cannot achieve recall higher than 48% in our case (not all AIM sentences occur in the beginning and end of a document, and some documents contain more than two AIM sentences).

5.4. Intrinsic Evaluation

Evaluation of the systems relies on 10-fold cross-validation: the model is trained on a training set of 72 documents, leaving 8 documents out at a time (the test set). The model is then used on the test set to assign each sentence a probability for each category *R*, and the category with the highest probability is chosen as answer for the sentence. This is repeated for all ten folds. The baselines for this task were discussed in section 4.2.

5.4.1. Naive Bayes Model

As Naive Bayes does not automatically ignore useless features, and as performance with bad features decreases, the first question is if all of our features are good disambiguators, or if some of the features do not contribute any useful information. Figure 5.33 shows the results of a 10-fold cross-validation.

Condition	Score
Start	Score = 0
If sentence in beginning	Score + 1
If sentence not in beginning	Score – 1
If Ag-1 = OUR_AIM_AGENT and Ag-2 = COPULA (non-negated) and first action in sentence and beginning (i.e. Loc = A, B, C, D or E	Score = 8
If Ag-1 = OUR_AIM_AGENT and Ag-2 = COPULA (non-negated) and first action in sentence and not beginning	Score = 6
If Ag-1 = OUR_AIM_AGENT and Ag-2 = COPULA (non-negated) and not first action in sentence and beginning	Score = 6
If Ag-1 = OUR_AIM_AGENT and Ag-2 = COPULA (non-negated) and not first action in sentence and not beginning	Score = 4
If Ag-1 = US_AGENT and Ag-2 = PRESENTATION_ACTION (non-negated) and first action in sentence and beginning	Score = 6
If Ag-1 = US_AGENT and Ag-2 = PRESENTATION_ACTION (non-negated) and first action in sentence and not beginning	Score = 4
If Ag-1 = US_AGENT and Ag-2 = PRESENTATION_ACTION (non-negated) and not first action in sentence and beginning	Score = 4
If Ag-1 = US_AGENT and Ag-2 = PRESENTATION_ACTION (non-negated) and not first action in sentence and not beginning	Score = 2
If Ag-1 = US_AGENT and Ag-2 = INTEREST_ACTION (non-negated) and first action in sentence and beginning	Score = 5
If Ag-1 = US_AGENT and Ag-2 = INTEREST_ACTION (non-negated) and first action in sentence and not beginning	Score = 3
If Ag-1 = US_AGENT and Ag-2 = INTEREST_ACTION (non-negated) and not first action in sentence and beginning	Score = 3
If Ag-1 = US_AGENT and Ag-2 = INTEREST_ACTION (non-negated) and not first action in sentence and not beginning	Score = 1
If Ag-1 = (REF_)US_AGENT and Ag-2 = SOLUTION_ACTION (non-negated) and first action in sentence and beginning	Score = 3
If Ag-1 = (REF_)US_AGENT and Ag-2 = SOLUTION_ACTION (non-negated) and first action in sentence and not beginning	Score = 2
If Ag-1 = (REF_)US_AGENT and Ag-2 = SOLUTION_ACTION (non-negated) and not first action in sentence and beginning	Score = 1
If Ag-1 = (REF_)US_AGENT and Ag-2 = SOLUTION_ACTION (non-negated) and not first action in sentence and not beginning	Score = 0
If Ag-1 = (REF_)US_AGENT and Ag-2 = ARGUMENTATION_ACTION (non-negated) and first action in sentence	Score = 3
If Ag-1 = (REF_)US_AGENT and Ag-2 = ARGUMENTATION_ACTION (non-negated) and not first action in sentence	Score = 2
If Ag-1 = REF_AGENT and Ag-2 = INTEREST_ACTION (non-negated) and first action in sentence	Score = 4
If Ag-1 = REF_AGENT and Ag-2 = INTEREST_ACTION (non-negated) and first action in sentence	Score = 3
If Ag-1 = REF_AGENT and Ag-2 = PRESENTATION_ACTION (non-negated) and first action in sentence	Score = 3
If Ag-1 = REF_AGENT and Ag-2 = PRESENTATION_ACTION (non-negated) and not first action in sentence	Score = 2

Figure 5.30: Symbolic Scores for AIM Sentences (1 of 2)

Condition	Score
If Ag-1 = AIM_REF_AGENT and Ag-2 = COPULA (non-negated) and first action in sentence	Score = 4
If Ag-1 = AIM_REF_AGENT and Ag-2 = COPULA (non-negated) and not first action in sentence	Score = 3
If Ag-1 = (REF_)US_AGENT and Ag-2 = RESEARCH_ACTION (non-negated)	Score = 1
If Formu = HERE_FORMULAIC and beginning	Score + 5
If Formu = METHOD_FORMULAIC and Ag-2 = (PRESENTATION_ACTION or INTEREST_ACTION) and Ag-1 = (REF_US_AGENT or REF_AGENT or *AIM*_AGENT)	Score + 5
If Struct-3 = Introduction	Score + 2
If Struct-3 = Conclusion	Score + 2
If Struct-1 = First-sentence	Score + 2
If very first sentence in document	Score + 1
If the previous sentence contained contrastive material (GAP, PROBLEM_ACTION, AWARE_ACTION, CONTRAST_FORMULAIC, negated SOLUTION ACTION), and beginning	Score + 2
If Ag-1 = US_AGENT	Score + 1
If there was a textstructure sentence in the past 3 sentences	Score – 1
If there is a DETAIL_FORMULAIC in the sentence	Score – 1
If Ag-1 = REF(_US?)_AGENT and Ag-2 = TEXTSTRUCTURE_ACTION	Score – 2
If last sentence was classified as TEXTUAL	Score – 3
If Ag-1 = (ref_)?us_agent and Ag-2 = PRESENTATION_ACTION and Syn-2 = Present and not beginning	Score – 2
If Ag-1 = TEXTSTRUCTURE_AGENT and Ag-2 = (TEXTSTRUCTURE_ACTION or PRESENTATION_ACTION or INTEREST_ACTION or RESEARCH_ACTION) or Formu = TEXTSTRUCTURE_FORMULAIC or formu = TEXTSTRUCTURE_AGENT	Score = 0
If there is a US_PREVIOUS_FORMULAIC in the sentence	Score = 0
If there is a FUTURE_FORMULAIC in the sentence	Score = 0

Figure 5.31: Symbolic Scores for AIM Sentences (2 of 2)

Feature	Alone	Left out	Feature	Alone	Left out
Cont-1	K=–.12	.37	Syn-2	K=–.12	.37
Cont-2	K=–.12	.37	Syn-3	K=–.12	.37
Struct-1	K=–.12	.36	Cit-1	K=+.18	.38
Struct-2	K=–.12	.37	Cit-2	K=+.13	.38
Struct-3	K=+.05	.35	Cit-3	K=+.12	.38
Loc	K=+.17	.34	Formu	K=+.06	.35
Length	K=–.12	.37	Ag-1	K=+.07	.36
Syn-1	K=–.12	.37	Ag-2	K=–.11	.35

Figure 5.33: Performance of Individual Features (Naive Bayes)

The first column in figure 5.33 ("Alone") corresponds to classification with a model using only the given feature, whereas the second column ("Left out") corresponds to a model using all other features but the given one. Some of the weaker

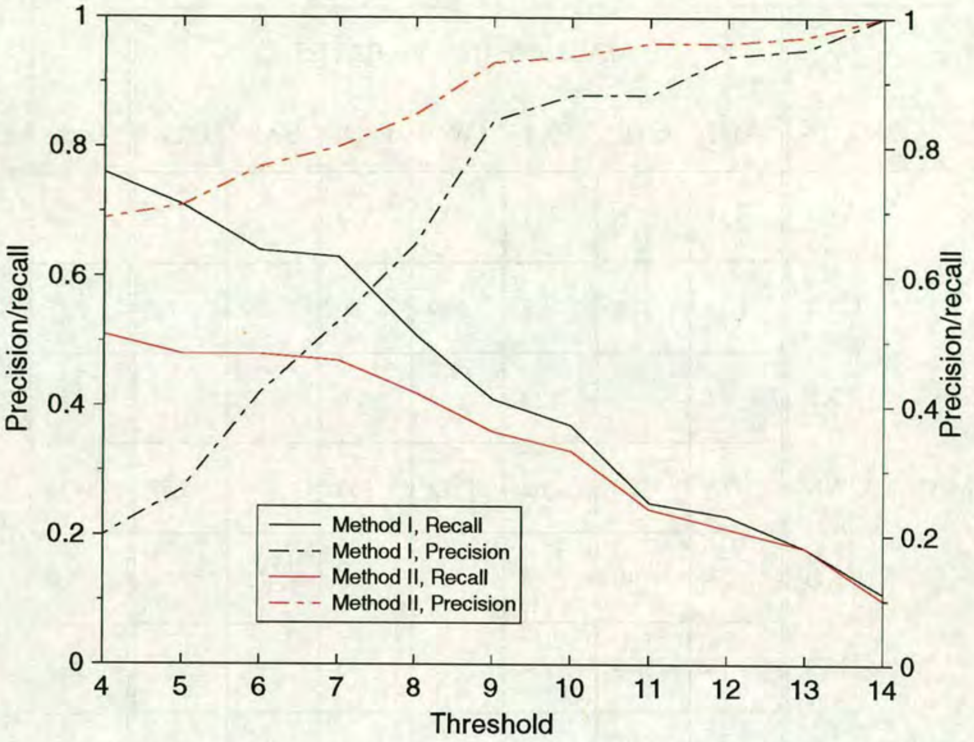


Figure 5.32: Effect of Threshold on Symbolic AIM Sentence Extraction

features are not predictive enough on their own to break the dominance of the prior; in that case, they behave just like Baseline B1 (K=–.12). A distinctive feature has a good classification on its own, and leads to a decreased performance if left out. The numbers show that some of the weaker features contribute some predictive power in combination with others, even if not on their own.

We measured the best performance using the features `Cont-1`, `Cont-2`, `Loc`, `Struct-1`, `Struct-2`, `Struct-3`, `Length`, `Syn-1`, `Syn-2`, `Syn-3`, `Cit-1`, `Formu`, `Ag-1` and `Ag-2`. Results only decreased when combinations of the citation features were used together; we assume this is due to the fact that these features encode redundant information with respect to each other; they are not independent. Appendix B.8 shows the output of the Naive Bayesian model on the example paper. The system's annotation achieved a Kappa value of K=0.41 on the example paper.

In an experiment between one annotator (C) and the statistical method, the observed reproducibility is K=.39 (N=12421, k=2), which corresponds to percentage

		MACHINE (NAIVE BAYES)							
		AIM	CTR	TXT	OWN	BKG	BAS	OTH	Total
HUMAN	AIM	131	8	11	33	14	7	5	209
	CTR	22	124	2	259	80	24	86	597
	TXT	13	3	138	51	6	5	7	223
	OWN	116	116	62	7623	163	96	257	8433
	BKG	28	40	3	257	305	11	76	720
	BAS	14	9	4	48	5	91	56	227
	OTH	8	71	10	1115	198	122	489	2013
	Total	332	371	230	9386	771	356	976	12422

Figure 5.34: Confusion Matrix: Human vs. Automatic Annotation, Naive Bayes

accuracy of 71.2%.

Note here that the system is not asked to annotate abstract sentences, so that N is lower than it would have been in a comparable experiment involving only human annotators. This number cannot be directly compared to experiments like Kupiec et al.'s because in their experiment a compression of around 3% was achieved whereas we classify each sentence into one of the categories.

When the Naive Bayesian Model is added to the pool of 3 coders, the reproducibility drops from K=.71 to K=.54 (N=3446, n=4). This reproducibility value is equivalent to the value achieved by 6 human annotators with no prior training, as in Study III.

Figure 5.34 depicts the confusion matrix for the classification. We can see that the system guesses too few OTHER and CONTRAST sentences, but overestimates the

	AIM	CTR	TXT	OWN	BKG	BAS	OTH
Precision	39%	33%	60%	81%	40%	26%	50%
Recall	63%	21%	62%	91%	42%	40%	24%

Figure 5.35: Precision and Recall per Category, Naive Bayes

number of BASIS sentences.

Figure 5.35 shows that the system performs well on AIM sentences, which can be determined with a recall of 63% and a precision of 39%. These values are more directly comparable to Kupiec et al.'s results of 44% precision and 44% recall for extracted sentences, even though not all of the sentences extracted by their method would have fallen into our AIM category. The other easily determinable category for the automatic method is TEXTUAL (p=60%; r=62%), whereas the results for the other non-basic categories are relatively lower—as are the human annotation results.

The results achieved with the more complicated statistical techniques were not much better. RIPPER (Cohen, 1995, 1996) achieved an error rate of 27.66% +/- 0.35% (a bit better than our error rate of 29%) in a ten-fold cross-validation. When the classifier described in Mikheev (To Appear) was used on our data, the classification was minimally better than both the Naive Bayes model and RIPPER, but training this model is very time consuming.

5.4.2. N-Gram Model

We measured performance of different n-gram models as before by 10-fold cross-validation. The best performance was achieved with a bigram model. This model achieved K=.41 (n=2,N=12422) when compared to Annotator C alone (P(A)=0.703, P(E)=0.492), and K=.56 (N=3334, n=4, P(A)=0.795, P(E)=0.537) when added to the pool of three annotators. Thus, adding the bigram model does improve performance. Appendix B.9 (p. 303) shows the output of the bigram model on the example paper. If we compare it to the output of the Naive Bayes model (p. 302), we notice that the contextual information introduced by the bigram model has added useful aspects to the annotation. For example, the Naive Bayes model did not annotate the two sentences dealing with Hindle's approach (bottom of the first column) as either OTHER or CONTRAST; instead, it just left them as BACKGROUND. Because of the high probability of CONTRAST sentences preceding AIM sentences, the Viterbi algorithm chose to mark

		MACHINE (BIGRAM)							
		Aim	Ctr	Txt	Own	Bkg	Bas	Oth	Total
HUMAN	Aim	124	10	12	27	25	3	8	209
	Ctr	20	122	3	208	138	15	91	597
	Txt	13	4	133	51	11	3	8	223
	Own	107	138	68	7220	459	99	342	8433
	Bkg	9	20	3	141	454	5	88	720
	Bas	18	14	4	69	12	80	30	227
	Oth	3	97	7	797	395	117	597	2013
	Total	294	405	230	8513	1494	322	1164	12422

Figure 5.36: Confusion Matrix: Human vs. Automatic Annotation, Bigram Model

them as Contrast; the fact that the posterior probability for Contrast was slightly lower than the posterior probability for Aim was overridden by the prior probabilities. Similarly, the erroneously tagged Textual sentence at the end of the introduction is corrected by the bigram model into Contrast.

In general, the bigram model tends to annotate longer segments; posterior probabilities have to be high to break this preference, i.e., to start new segments. This also introduces errors, e.g., the long Contrast segment at the end of the second column which was not perceived to be there by either human annotator. Overall, the bigram model's annotation reached a Kappa value of 0.35 on this particular paper, i.e. performance *decreased* when compared to the Naive Bayesian model.

For the case of human vs. bigram model, the confusion matrix in figure 5.36 was recorded. Figure 5.37 shows precision and recall values for individual categories. In contrast to the Naive Bayesian model, the recognition results for the categories Aim,

	AIM	CTR	TXT	OWN	BKG	BAS	OTH
Precision	42%	30%	58%	85%	30%	25%	51%
Recall	59%	20%	60%	86%	63%	35%	30%

Figure 5.37: Precision and Recall per Category, Bigram Model

OTHER and OWN are higher, and those for the categories CONTRAST, TEXTUAL, BASIS and BACKGROUND lower.

5.4.3. Symbolic Rules

The symbolic rules do not aim at a full-coverage recognition of all categories. Rather, they provide a high-precision, low-recall coverage of the four non-basic categories AIM, TEXTUAL, BASIS and CONTRAST. The evaluation of the success of these rules can therefore not be measured by Kappa (which would require a full-coverage classification), but only by precision and recall of these four categories. Precision and recall was varied by changing the threshold.

Figure 5.38 presents precision and recall plots for the non-basic categories. The results show that it is possible to determine AIM and TEXTUAL sentences in a scientific article with high precision, albeit with considerably lower recall. This is a good result, which in itself justifies the Agentivity features. The result is also in agreement with our results from chapter 4 which showed that AIM sentences (and to a lesser degree TEXTUAL sentences) are also recognized most robustly of all categories by humans. They state knowledge claims—it is important for authors to bring the own knowledge claims across—or organize the text. Typically, they are expressed in a formalized way. BASIS and CONTRAST sentences have a less prototypical syntactic realization, and they also occur at less predictable places in the document. Therefore, it is far more difficult for both machine and human to recognize such sentences.

Figure 5.38 also shows the best stochastic results for the non-basic categories (dots) for comparison. The results for AIM and CONTRAST are better with the symbolic system, whereas the reverse is the case for the categories BASIS and TEXTUAL.

5.5. Results of System Run on Unseen Material

An ad-hoc test was performed on a paper randomly drawn from the archive. It was preprocessed with minimal manual intervention and then put through the argumentative

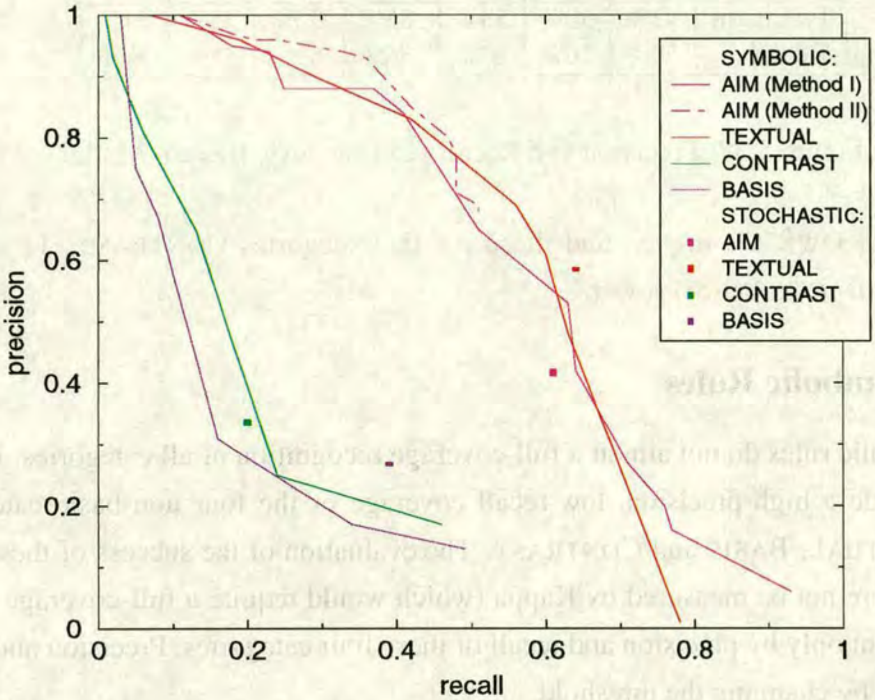
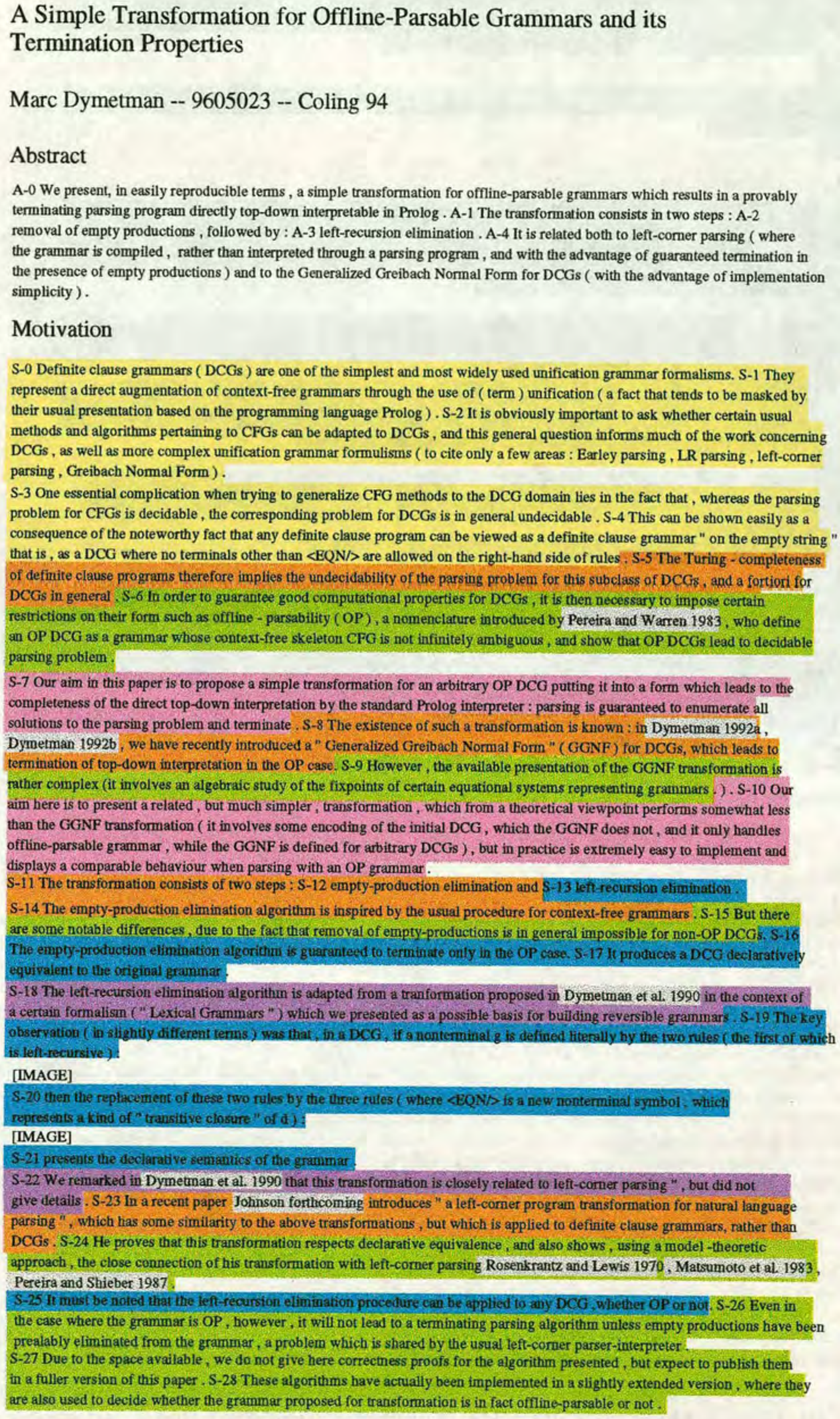


Figure 5.38: Precision and Recall of Symbolic Sentence Extraction

zoner. The output of the Naive Bayesian model is given in figures 5.39 and 5.40, and the output of the bigram model is given in figures 5.41 and 5.42 so that the reader can inspect the result.

The only difference in performance which can be expected when moving from seen to unseen text has to do with the features based on meta-discourse (`Formu, Ag-1` and `Ag-2`), as the list of expressions was expanded manually during system development, whenever the system's results showed phrases not previously contained in the lists. All other features are rather independent of the question whether or not the system developer sees more data. One would hope that the common meta-discourse phrases are covered by the list, and that expressions not encountered in the first 80 papers would be rather specialized and infrequent.

It is difficult to assess to what extent our features treat unseen text adequately, because there are no gold standards for the unseen test. We report an experiment with a predecessor of the three meta-discourse features in Teufel and Moens (1997). We divided our corpus (then 123 articles, including articles which did not appear in ACL, EACL, COLING or ANLP conferences) into three parts. We pretended that one third was "unseen", by using only those 1423 formulaic expressions for extraction which



A Simple Transformation for Offline-Parsable Grammars and its Termination Properties

Marc Dymetman -- 9605023 -- Coling 94

Abstract

A-0 We present, in easily reproducible terms , a simple transformation for offline-parsable grammars which results in a provably terminating parsing program directly top-down interpretable in Prolog . A-1 The transformation consists in two steps : A-2 removal of empty productions , followed by : A-3 left-recursion elimination . A-4 It is related both to left-corner parsing (where the grammar is compiled , rather than interpreted through a parsing program , and with the advantage of guaranteed termination in the presence of empty productions) and to the Generalized Greibach Normal Form for DCGs (with the advantage of implementation simplicity) .

Motivation

S-0 Definite clause grammars (DCGs) are one of the simplest and most widely used unification grammar formalisms. S-1 They represent a direct augmentation of context-free grammars through the use of (term) unification (a fact that tends to be masked by their usual presentation based on the programming language Prolog) . S-2 It is obviously important to ask whether certain usual methods and algorithms pertaining to CFGs can be adapted to DCGs , and this general question informs much of the work concerning DCGs , as well as more complex unification grammar formulisms (to cite only a few areas : Earley parsing , LR parsing , left-corner parsing , Greibach Normal Form) .

S-3 One essential complication when trying to generalize CFG methods to the DCG domain lies in the fact that , whereas the parsing problem for CFGs is decidable , the corresponding problem for DCGs is in general undecidable . S-4 This can be shown easily as a consequence of the noteworthy fact that any definite clause program can be viewed as a definite clause grammar " on the empty string " , that is , as a DCG where no terminals other than <EQN/> are allowed on the right-hand side of rules . S-5 The Turing - completeness of definite clause programs therefore implies the undecidability of the parsing problem for this subclass of DCGs , and a fortiori for DCGs in general . S-6 In order to guarantee good computational properties for DCGs , it is then necessary to impose certain restrictions on their form such as offline - parsability (OP) , a nomenclature introduced by Pereira and Warren 1983 , who define an OP DCG as a grammar whose context-free skeleton CFG is not infinitely ambiguous , and show that OP DCGs lead to decidable parsing problem .

S-7 Our aim in this paper is to propose a simple transformation for an arbitrary OP DCG putting it into a form which leads to the completeness of the direct top-down interpretation by the standard Prolog interpreter : parsing is guaranteed to enumerate all solutions to the parsing problem and terminate . S-8 The existence of such a transformation is known : in Dymetman 1992a , Dymetman 1992b , we have recently introduced a " Generalized Greibach Normal Form " (GGNF) for DCGs, which leads to termination of top-down interpretation in the OP case. S-9 However , the available presentation of the GGNF transformation is rather complex (it involves an algebraic study of the fixpoints of certain equational systems representing grammars .) . S-10 Our aim here is to present a related , but much simpler , transformation , which from a theoretical viewpoint performs somewhat less than the GGNF transformation (it involves some encoding of the initial DCG , which the GGNF does not , and it only handles offline-parsable grammar , while the GGNF is defined for arbitrary DCGs) , but in practice is extremely easy to implement and displays a comparable behaviour when parsing with an OP grammar .

S-11 The transformation consists of two steps : S-12 empty-production elimination and S-13 left-recursion elimination .

S-14 The empty-production elimination algorithm is inspired by the usual procedure for context-free grammars . S-15 But there are some notable differences , due to the fact that removal of empty-productions is in general impossible for non-OP DCGs. S-16 The empty-production elimination algorithm is guaranteed to terminate only in the OP case. S-17 It produces a DCG declaratively equivalent to the original grammar .

S-18 The left-recursion elimination algorithm is adapted from a tranformation proposed in Dymetman et al. 1990 in the context of a certain formalism (" Lexical Grammars ") which we presented as a possible basis for building reversible grammars . S-19 The key observation (in slightly different terms) was that , in a DCG , if a nonterminal g is defined literally by the two rules (the first of which is left-recursive) :

[IMAGE]

S-20 then the replacement of these two rules by the three rules (where <EQN/> is a new nonterminal symbol , which represents a kind of " transitive closure " of d) :

[IMAGE]

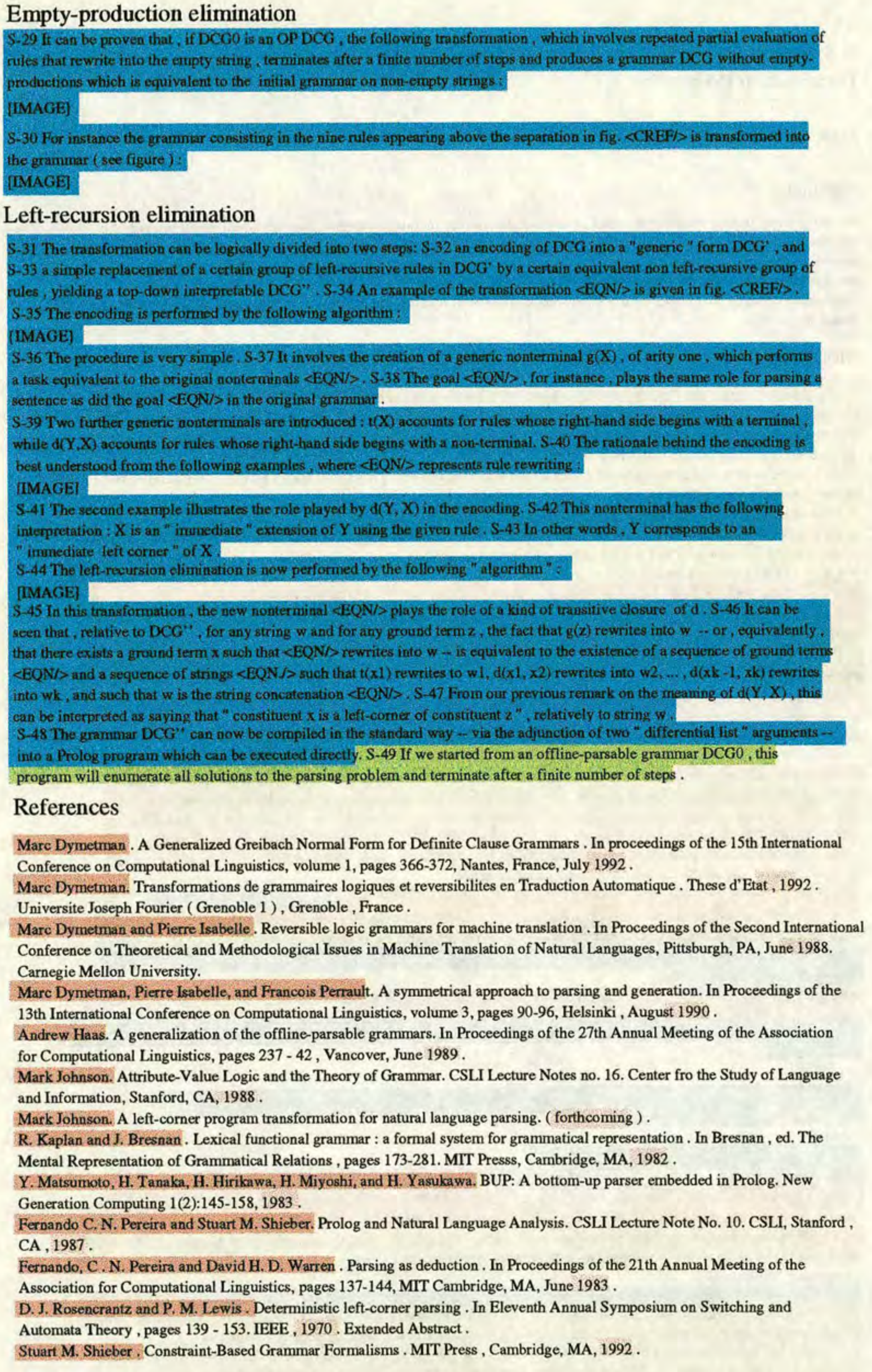
S-21 presents the declarative semantics of the grammar .

S-22 We remarked in Dymetman et al. 1990 that this transformation is closely related to left-corner parsing " , but did not give details . S-23 In a recent paper Johnson forthcoming introduces " a left-corner program transformation for natural language parsing " , which has some similarity to the above transformations , but which is applied to definite clause grammars, rather than DCGs . S-24 He proves that this transformation respects declarative equivalence , and also shows , using a model -theoretic approach , the close connection of his transformation with left-corner parsing Rosenkrantz and Lewis 1970 , Matsumoto et al. 1983 , Pereira and Shieber 1987 .

S-25 It must be noted that the left-recursion elimination procedure can be applied to any DCG ,whether OP or not. S-26 Even in the case where the grammar is OP , however , it will not lead to a terminating parsing algorithm unless empty productions have been prealably eliminated from the grammar , a problem which is shared by the usual left-corner parser-interpreter .

S-27 Due to the space available , we do not give here correctness proofs for the algorithm presented , but expect to publish them in a fuller version of this paper . S-28 These algorithms have actually been implemented in a slightly extended version , where they are also used to decide whether the grammar proposed for transformation is in fact offline-parsable or not .

Figure 5.39: Unseen Document 9605023, Automatic Argumentative Zoning by Naive Bayes (1 of 2)



Empty-production elimination

S-29 It can be proven that , if DCG0 is an OP DCG , the following transformation , which involves repeated partial evaluation of rules that rewrite into the empty string , terminates after a finite number of steps and produces a grammar DCG without empty-productions which is equivalent to the initial grammar on non-empty strings :

[IMAGE]

S-30 For instance the grammar consisting in the nine rules appearing above the separation in fig. <CREF/> is transformed into the grammar (see figure) :

[IMAGE]

Left-recursion elimination

S-31 The transformation can be logically divided into two steps: S-32 an encoding of DCG into a "generic " form DCG' , and S-33 a simple replacement of a certain group of left-recursive rules in DCG' by a certain equivalent non left-recursive group of rules , yielding a top-down interpretable DCG'' . S-34 An example of the transformation <EQN/> is given in fig. <CREF/> .

S-35 The encoding is performed by the following algorithm :

[IMAGE]

S-36 The procedure is very simple . S-37 It involves the creation of a generic nonterminal g(X) , of arity one , which performs a task equivalent to the original nonterminals <EQN/> . S-38 The goal <EQN/> , for instance , plays the same role for parsing a sentence as did the goal <EQN/> in the original grammar .

S-39 Two further generic nonterminals are introduced : t(X) accounts for rules whose right-hand side begins with a terminal , while d(Y,X) accounts for rules whose right-hand side begins with a non-terminal. S-40 The rationale behind the encoding is best understood from the following examples , where <EQN/> represents rule rewriting :

[IMAGE]

S-41 The second example illustrates the role played by d(Y, X) in the encoding. S-42 This nonterminal has the following interpretation : X is an " immediate " extension of Y using the given rule . S-43 In other words , Y corresponds to an " immediate left corner " of X .

S-44 The left-recursion elimination is now performed by the following " algorithm " :

[IMAGE]

S-45 In this transformation , the new nonterminal <EQN/> plays the role of a kind of transitive closure of d . S-46 It can be seen that , relative to DCG'' , for any string w and for any ground term z , the fact that g(z) rewrites into w -- or , equivalently , that there exists a ground term x such that <EQN/> rewrites into w -- is equivalent to the existence of a sequence of ground terms <EQN/> and a sequence of strings <EQN./> such that t(x1) rewrites to w1, d(x1, x2) rewrites into w2, ... , d(xk -1, xk) rewrites into wk , and such that w is the string concatenation <EQN/> . S-47 From our previous remark on the meaning of d(Y, X) , this can be interpreted as saying that " constituent x is a left-corner of constituent z " , relatively to string w .

S-48 The grammar DCG'' can now be compiled in the standard way -- via the adjunction of two " differential list " arguments -- into a Prolog program which can be executed directly. S-49 If we started from an offline-parsable grammar DCG0 , this program will enumerate all solutions to the parsing problem and terminate after a finite number of steps .

References

Marc Dymetman . A Generalized Greibach Normal Form for Definite Clause Grammars . In proceedings of the 15th International Conference on Computational Linguistics, volume 1, pages 366-372, Nantes, France, July 1992 .

Marc Dymetman. Transformations de grammaires logiques et reversibilites en Traduction Automatique . These d'Etat , 1992 . Universite Joseph Fourier (Grenoble 1) , Grenoble , France .

Marc Dymetman and Pierre Isabelle . Reversible logic grammars for machine translation . In Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Pittsburgh, PA, June 1988. Carnegie Mellon University.

Marc Dymetman, Pierre Isabelle, and Francois Perrault. A symmetrical approach to parsing and generation. In Proceedings of the 13th International Conference on Computational Linguistics, volume 3, pages 90-96, Helsinki , August 1990 .

Andrew Haas. A generalization of the offline-parsable grammars. In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, pages 237 - 42 , Vancover, June 1989 .

Mark Johnson. Attribute-Value Logic and the Theory of Grammar. CSLI Lecture Notes no. 16. Center fro the Study of Language and Information, Stanford, CA, 1988 .

Mark Johnson. A left-corner program transformation for natural language parsing. (forthcoming) .

R. Kaplan and J. Bresnan . Lexical functional grammar : a formal system for grammatical representation . In Bresnan , ed. The Mental Representation of Grammatical Relations , pages 173-281. MIT Presss, Cambridge, MA, 1982 .

Y. Matsumoto, H. Tanaka, H. Hirikawa, H. Miyoshi, and H. Yasukawa. BUP: A bottom-up parser embedded in Prolog. New Generation Computing 1(2):145-158, 1983 .

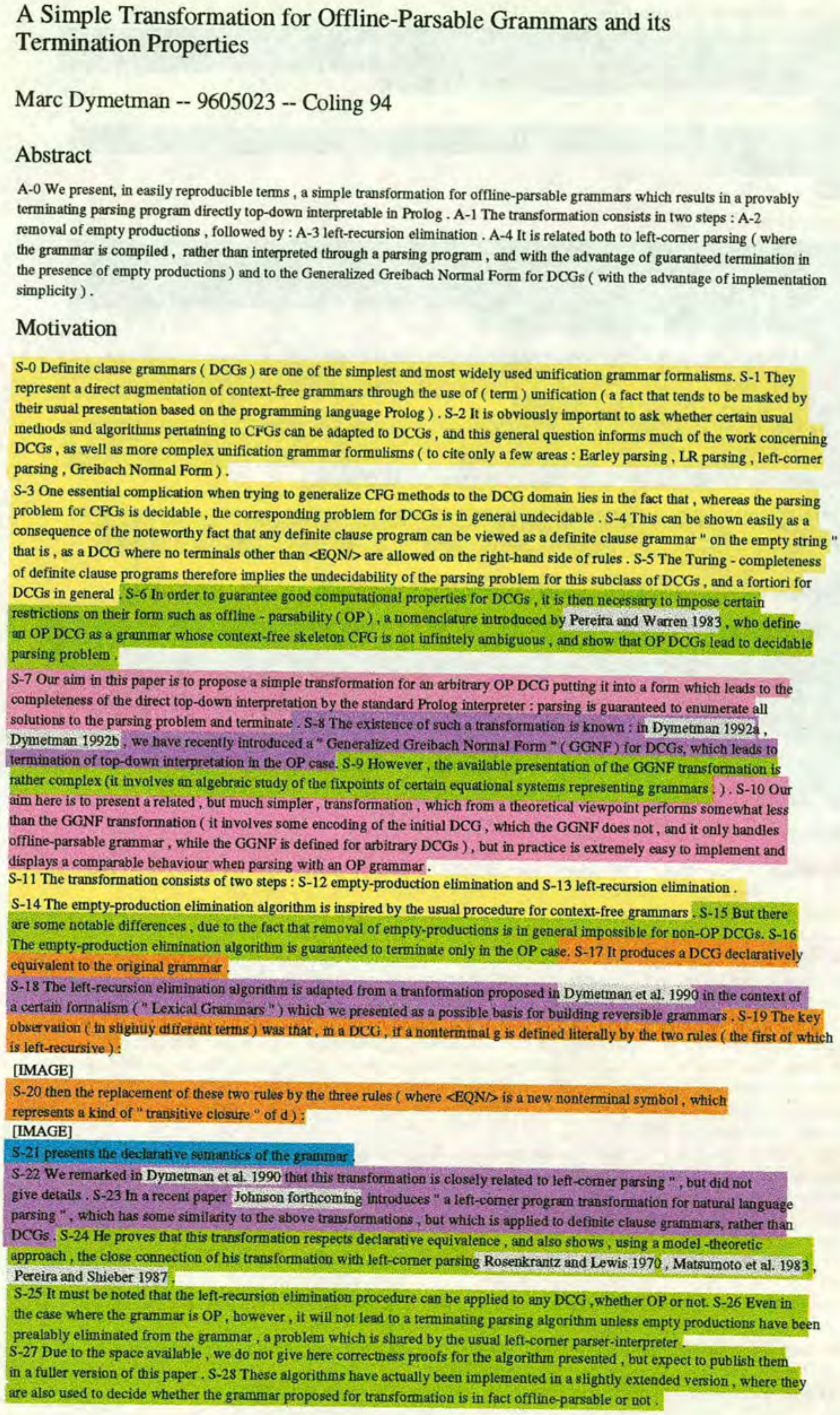
Fernando C. N. Pereira and Stuart M. Shieber. Prolog and Natural Language Analysis. CSLI Lecture Note No. 10. CSLI, Stanford , CA , 1987 .

Fernando, C . N. Pereira and David H. D. Warren . Parsing as deduction . In Proceedings of the 21th Annual Meeting of the Association for Computational Linguistics, pages 137-144, MIT Cambridge, MA, June 1983 .

D. J. Rosencrantz and P. M. Lewis . Deterministic left-corner parsing . In Eleventh Annual Symposium on Switching and Automata Theory , pages 139 - 153. IEEE , 1970 . Extended Abstract .

Stuart M. Shieber . Constraint-Based Grammar Formalisms . MIT Press , Cambridge, MA, 1992 .

Figure 5.40: Unseen Document 9605023, Automatic Argumentative Zoning by Naive Bayes (2 of 2)



A Simple Transformation for Offline-Parsable Grammars and its Termination Properties

Marc Dymetman -- 9605023 -- Coling 94

Abstract

A-0 We present, in easily reproducible terms , a simple transformation for offline-parsable grammars which results in a provably terminating parsing program directly top-down interpretable in Prolog . A-1 The transformation consists in two steps : A-2 removal of empty productions , followed by : A-3 left-recursion elimination . A-4 It is related both to left-corner parsing (where the grammar is compiled , rather than interpreted through a parsing program , and with the advantage of guaranteed termination in the presence of empty productions) and to the Generalized Greibach Normal Form for DCGs (with the advantage of implementation simplicity) .

Motivation

S-0 Definite clause grammars (DCGs) are one of the simplest and most widely used unification grammar formalisms. S-1 They represent a direct augmentation of context-free grammars through the use of (term) unification (a fact that tends to be masked by their usual presentation based on the programming language Prolog) . S-2 It is obviously important to ask whether certain usual methods and algorithms pertaining to CFGs can be adapted to DCGs , and this general question informs much of the work concerning DCGs , as well as more complex unification grammar formulisms (to cite only a few areas : Earley parsing , LR parsing , left-corner parsing , Greibach Normal Form) .

S-3 One essential complication when trying to generalize CFG methods to the DCG domain lies in the fact that , whereas the parsing problem for CFGs is decidable , the corresponding problem for DCGs is in general undecidable . S-4 This can be shown easily as a consequence of the noteworthy fact that any definite clause program can be viewed as a definite clause grammar " on the empty string " , that is , as a DCG where no terminals other than <EQN/> are allowed on the right-hand side of rules . S-5 The Turing - completeness of definite clause programs therefore implies the undecidability of the parsing problem for this subclass of DCGs , and a fortiori for DCGs in general . S-6 In order to guarantee good computational properties for DCGs , it is then necessary to impose certain restrictions on their form such as offline - parsability (OP) , a nomenclature introduced by Pereira and Warren 1983 , who define an OP DCG as a grammar whose context-free skeleton CFG is not infinitely ambiguous , and show that OP DCGs lead to decidable parsing problem .

S-7 Our aim in this paper is to propose a simple transformation for an arbitrary OP DCG putting it into a form which leads to the completeness of the direct top-down interpretation by the standard Prolog interpreter : parsing is guaranteed to enumerate all solutions to the parsing problem and terminate . S-8 The existence of such a transformation is known : in Dymetman 1992a , Dymetman 1992b , we have recently introduced a " Generalized Greibach Normal Form " (GGNF) for DCGs, which leads to termination of top-down interpretation in the OP case. S-9 However , the available presentation of the GGNF transformation is rather complex (it involves an algebraic study of the fixpoints of certain equational systems representing grammars .) . S-10 Our aim here is to present a related , but much simpler , transformation , which from a theoretical viewpoint performs somewhat less than the GGNF transformation (it involves some encoding of the initial DCG , which the GGNF does not , and it only handles offline-parsable grammar , while the GGNF is defined for arbitrary DCGs) , but in practice is extremely easy to implement and displays a comparable behaviour when parsing with an OP grammar .

S-11 The transformation consists of two steps : S-12 empty-production elimination and S-13 left-recursion elimination .

S-14 The empty-production elimination algorithm is inspired by the usual procedure for context-free grammars . S-15 But there are some notable differences , due to the fact that removal of empty-productions is in general impossible for non-OP DCGs. S-16 The empty-production elimination algorithm is guaranteed to terminate only in the OP case. S-17 It produces a DCG declaratively equivalent to the original grammar .

S-18 The left-recursion elimination algorithm is adapted from a tranformation proposed in Dymetman et al. 1990 in the context of a certain formalism (" Lexical Grammars ") which we presented as a possible basis for building reversible grammars . S-19 The key observation (in slightly different terms) was that , in a DCG , if a nonterminal g is defined literally by the two rules (the first of which is left-recursive) :

[IMAGE]

S-20 then the replacement of these two rules by the three rules (where <EQN/> is a new nonterminal symbol , which represents a kind of " transitive closure " of d) :

[IMAGE]

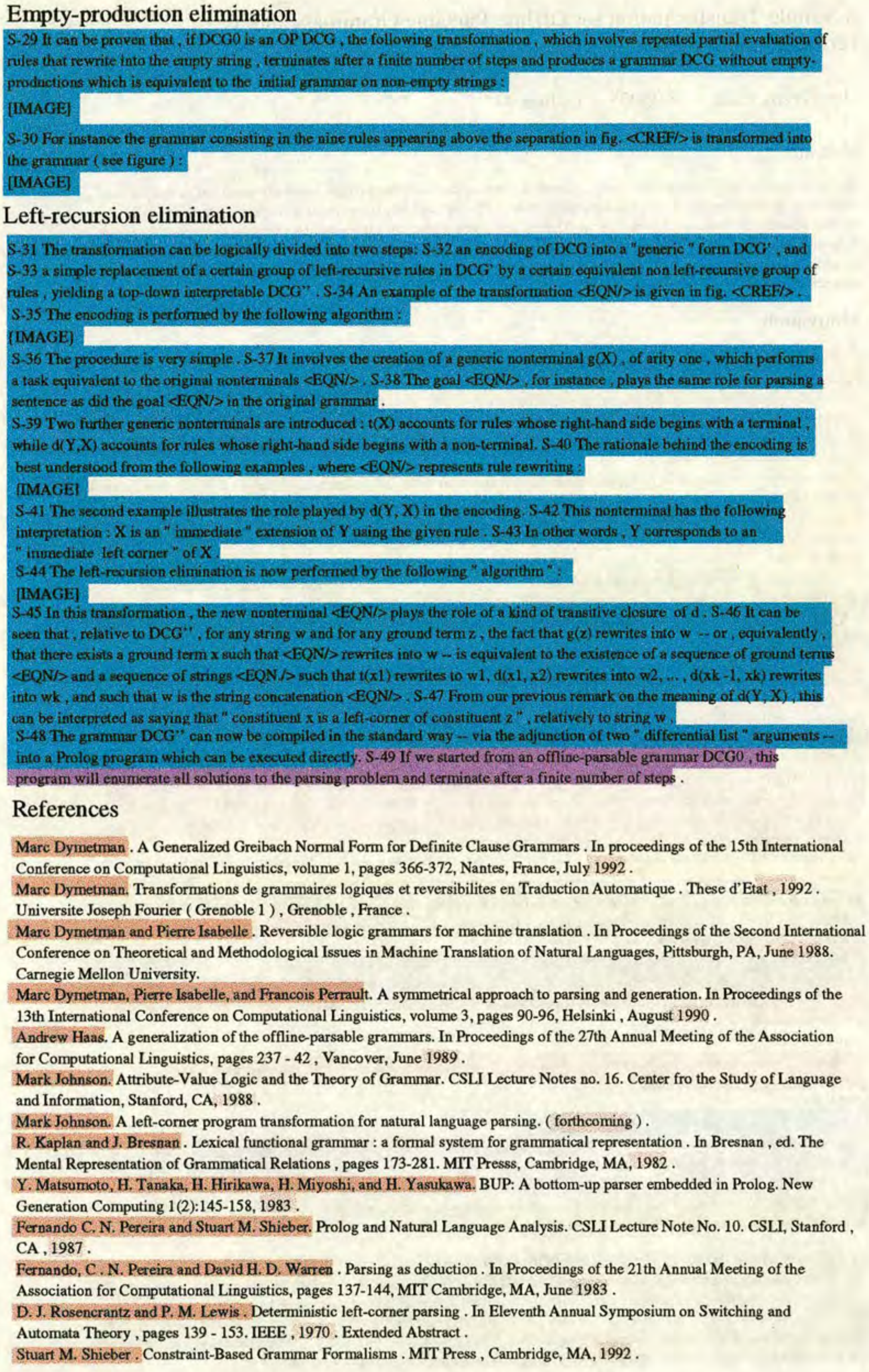
S-21 presents the declarative semantics of the grammar .

S-22 We remarked in Dymetman et al. 1990 that this transformation is closely related to left-corner parsing " , but did not give details . S-23 In a recent paper Johnson forthcoming introduces " a left-corner program transformation for natural language parsing " , which has some similarity to the above transformations , but which is applied to definite clause grammars, rather than DCGs . S-24 He proves that this transformation respects declarative equivalence , and also shows , using a model -theoretic approach , the close connection of his transformation with left-corner parsing Rosenkrantz and Lewis 1970 , Matsumoto et al. 1983 , Pereira and Shieber 1987 .

S-25 It must be noted that the left-recursion elimination procedure can be applied to any DCG ,whether OP or not. S-26 Even in the case where the grammar is OP , however , it will not lead to a terminating parsing algorithm unless empty productions have been prealably eliminated from the grammar , a problem which is shared by the usual left-corner parser-interpreter .

S-27 Due to the space available , we do not give here correctness proofs for the algorithm presented , but expect to publish them in a fuller version of this paper . S-28 These algorithms have actually been implemented in a slightly extended version , where they are also used to decide whether the grammar proposed for transformation is in fact offline-parsable or not .

Figure 5.41: Unseen Document 9605023, Automatic Argumentative Zoning by Bigram (1 of 2)



Empty-production elimination

S-29 It can be proven that , if DCG0 is an OP DCG , the following transformation , which involves repeated partial evaluation of rules that rewrite into the empty string , terminates after a finite number of steps and produces a grammar DCG without empty-productions which is equivalent to the initial grammar on non-empty strings :

[IMAGE]

S-30 For instance the grammar consisting in the nine rules appearing above the separation in fig. <CREF/> is transformed into the grammar (see figure) :

[IMAGE]

Left-recursion elimination

S-31 The transformation can be logically divided into two steps: S-32 an encoding of DCG into a "generic " form DCG' , and S-33 a simple replacement of a certain group of left-recursive rules in DCG' by a certain equivalent non left-recursive group of rules , yielding a top-down interpretable DCG'' . S-34 An example of the transformation <EQN/> is given in fig. <CREF/> .

S-35 The encoding is performed by the following algorithm :

{IMAGE]

S-36 The procedure is very simple . S-37 It involves the creation of a generic nonterminal g(X) , of arity one , which performs a task equivalent to the original nonterminals <EQN/> . S-38 The goal <EQN/> , for instance , plays the same role for parsing a sentence as did the goal <EQN/> in the original grammar .

S-39 Two further generic nonterminals are introduced : t(X) accounts for rules whose right-hand side begins with a terminal , while d(Y,X) accounts for rules whose right-hand side begins with a non-terminal. S-40 The rationale behind the encoding is best understood from the following examples , where <EQN/> represents rule rewriting :

{IMAGE]

S-41 The second example illustrates the role played by d(Y, X) in the encoding. S-42 This nonterminal has the following interpretation : X is an " immediate " extension of Y using the given rule . S-43 In other words , Y corresponds to an " immediate left corner " of X .

S-44 The left-recursion elimination is now performed by the following " algorithm " :

[IMAGE]

S-45 In this transformation , the new nonterminal <EQN/> plays the role of a kind of transitive closure of d . S-46 It can be seen that , relative to DCG'' , for any string w and for any ground term z , the fact that g(z) rewrites into w -- or , equivalently , that there exists a ground term x such that <EQN/> rewrites into w -- is equivalent to the existence of a sequence of ground terms <EQN/> and a sequence of strings <EQN./> such that t(x1) rewrites to w1, d(x1, x2) rewrites into w2, ... , d(xk -1, xk) rewrites into wk , and such that w is the string concatenation <EQN/> . S-47 From our previous remark on the meaning of d(Y, X) , this can be interpreted as saying that " constituent x is a left-corner of constituent z " , relatively to string w .

S-48 The grammar DCG'' can now be compiled in the standard way -- via the adjunction of two " differential list " arguments -- into a Prolog program which can be executed directly. S-49 If we started from an offline-parsable grammar DCG0 , this program will enumerate all solutions to the parsing problem and terminate after a finite number of steps .

References

Marc Dymetman . A Generalized Greibach Normal Form for Definite Clause Grammars . In proceedings of the 15th International Conference on Computational Linguistics, volume 1, pages 366-372, Nantes, France, July 1992 .

Marc Dymetman. Transformations de grammaires logiques et reversibilites en Traduction Automatique . These d'Etat , 1992 . Universite Joseph Fourier (Grenoble 1) , Grenoble , France .

Marc Dymetman and Pierre Isabelle . Reversible logic grammars for machine translation . In Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Pittsburgh, PA, June 1988. Carnegie Mellon University.

Marc Dymetman, Pierre Isabelle, and Francois Perrault. A symmetrical approach to parsing and generation. In Proceedings of the 13th International Conference on Computational Linguistics, volume 3, pages 90-96, Helsinki , August 1990 .

Andrew Haas. A generalization of the offline-parsable grammars. In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, pages 237 - 42 , Vancover, June 1989 .

Mark Johnson. Attribute-Value Logic and the Theory of Grammar. CSLI Lecture Notes no. 16. Center fro the Study of Language and Information, Stanford, CA, 1988 .

Mark Johnson. A left-corner program transformation for natural language parsing. (forthcoming) .

R. Kaplan and J. Bresnan . Lexical functional grammar : a formal system for grammatical representation . In Bresnan , ed. The Mental Representation of Grammatical Relations , pages 173-281. MIT Presss, Cambridge, MA, 1982 .

Y. Matsumoto, H. Tanaka, H. Hirikawa, H. Miyoshi, and H. Yasukawa. BUP: A bottom-up parser embedded in Prolog. New Generation Computing 1(2):145-158, 1983 .

Fernando C. N. Pereira and Stuart M. Shieber. Prolog and Natural Language Analysis. CSLI Lecture Note No. 10. CSLI, Stanford , CA , 1987 .

Fernando, C . N. Pereira and David H. D. Warren . Parsing as deduction . In Proceedings of the 21th Annual Meeting of the Association for Computational Linguistics, pages 137-144, MIT Cambridge, MA, June 1983 .

D. J. Rosencrantz and P. M. Lewis . Deterministic left-corner parsing . In Eleventh Annual Symposium on Switching and Automata Theory , pages 139 - 153. IEEE , 1970 . Extended Abstract .

Stuart M. Shieber . Constraint-Based Grammar Formalisms . MIT Press , Cambridge, MA, 1992 .

Figure 5.42: Unseen Document 9605023, Automatic Argumentative Zoning by Bigram (2 of 2)

	Seen	Unseen
Cue Phrase Feature	60.9	54.9
All Features	71.6	65.3
Baseline	29.1	

Figure 5.43: Performance of Meta-Discourse Features; Unseen and Seen Data

were compiled from the other two parts. The advantage of this was that we now had gold standards for the "unseen" part, and we could compare the system's performance with both lists. Performance decreased significantly on unseen data, but not catastrophically, as can be seen from figure 5.43 (values refer to relevance-extraction, and are given in precision = recall values, in Kupiec et al. style). Even though the task is not the same, and the cue phrase method has been improved since to form our more recent meta-discourse features `Formu`, `Ag-1` and `Ag-2`, we still conclude from this experiment that meta-discourse features can be rather stable, even if only two thirds of the data is taken into account.

5.6. Conclusion

Annotator	Kappa	Raw Agr.	Random Agr.
System:			
Naive Bayes	.39	71%	54%
Naive Bayes + Bigram	.41	70%	49%
Humans:			
Task-trained	.71	87%	56%
Non task-trained (avg.)	.51	76%	49%
Baselines:			
Most frequent category	-.12	68%	71%
Random, uniform distribution	-.10	14%	22%
Random, observed distribution	0	48%	48%

Figure 5.44: Results of Human and Automatic Argumentative Zoning, I

Figures 5.44 and 5.45 summarize all evaluation results. If we compare humans and automatic results we see that there is still plenty of room for improvement for our systems. However, the automatic performance results are also a lot better than random, as the distance from the K=0 point (the most sensible baseline for our task) shows. Argumentative Zoning is a new task, so there are no direct numerical values to compare our prototype's performance with. When compared to Kupiec et al.'s result, both an earlier implementation (Teufel and Moens, 1997) and the current results compare favourably, if we consider our systems' success on AIM sentences. Additionally, if all one wants are extracted AIM and TEXTUAL sentences, our symbolic rules provide a good solution: both our implementations are much better at categorizing TEXTUAL and AIM sentences than they are at categorizing BASIS and CONTRAST sentences.

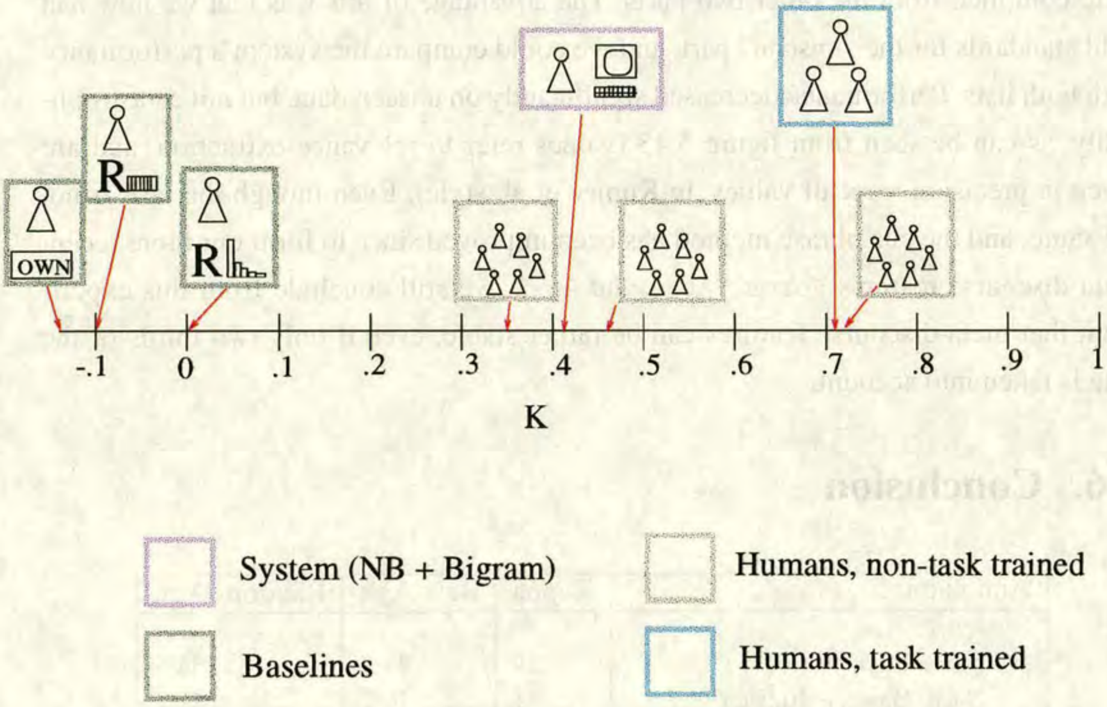


Figure 5.45: Results of Human and Automatic Argumentative Zoning, II

However, statistical classification is still rather noisy. We assume that the main reason for this is lack of training data: we were training on only 72 documents. However, as corpus collection and manual annotation with such a high level of document semantics is rather time consuming, it was not possible in the time frame of this thesis to expand the training data.

We believe that numerically high results are not absolutely required for a workable system. We see Argumentative Zoning as a forgiving task. Language is redundant, and the most important pieces of information will be repeated in the paper. Names of other peoples' solutions, for example, or references to based-on solutions, get repeated over and over—recognizing them *once* is enough to get the right kind of information into our RDP slot. We often found in the human annotation experiment that different versions of annotation on one paper still essentially contained the same information, i.e. would have resulted in similar RDPs. This effect would probably also apply to papers which are less than optimally zoned by an automatic process.

We see our results as an indication that we are on the right track for a difficult task, even though they are still modest at present. Some of the features known from text extraction have reconfirmed their usefulness for a new task. Our new features for argumentative sentence classification, which are based on agents and actions, have managed to increase our statistical results, and they have also provided useful input to the symbolic classification results.

Chapter 6

Conclusions

In this thesis, we have introduced a new task for document management, which we call *Argumentative Zoning*. Argumentative Zoning is the analysis of the argumentative status of sentences in scientific articles. Figure 6.1 shows how argumentative zones (and their derivatives, RDPs or Rhetorical Document Profiles) act as intermediaries between the reader and the writer. It also shows the setup of the experiments we performed to explore the task of Argumentative Zoning: a system for automatic Argumentative Zoning is evaluated intrinsically by comparison to human Argumentative Zoning. At the same time, the human annotation provides training material for the system.

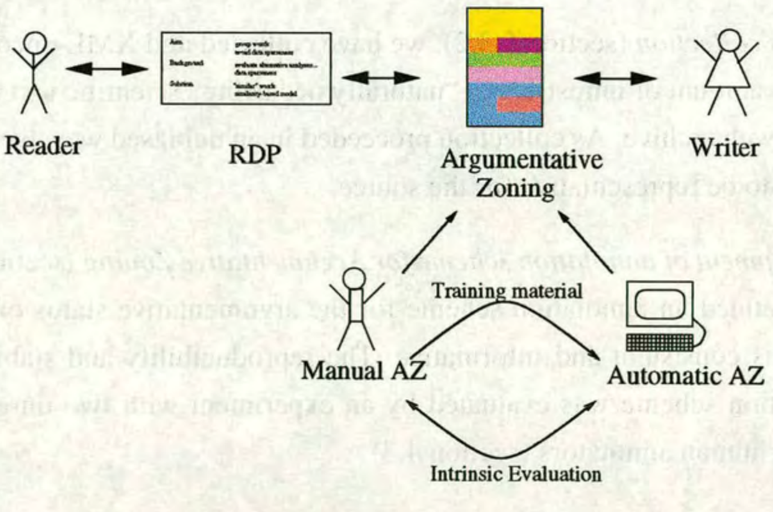


Figure 6.1: Overview of Argumentative Zoning Experiments

6.1. Contribution of the Thesis

The main theoretical claim of this thesis is that empirical discourse analysis can contribute towards the problem of document characterization in a document retrieval environment. We exemplify this by applying an analysis of prototypical scientific argumentation, Argumentative Zoning, to scientific articles. We claim that the type of document structure that argumentative zones capture is dominant in this text type, and also particularly useful for our task.

While Argumentative Zoning relies on rhetorical effects which are specific to the text type, it is independent of the subject matter treated. We have shown that the task of Argumentative Zoning is defined well enough for humans to be able to perform it consistently.

We have identified sentential features which correlate with the argumentative status of the given sentence. The existence of these correlates means that human annotation behaviour can in principle be simulated automatically. We have provided algorithms for the determination of these features. The more complicated features aim at modelling meta-discourse as an expression of prototypical scientific argumentation; we use linguistic heuristics and pattern matching to this end.

The practical contributions of this thesis are threefold:

- *Corpus collection* (section 5.3.2): we have collected and XML-encoded a substantial amount of unrestricted, "naturally occurring" scientific text from a scientific web archive. As collection proceeded in an unbiased way, we expect the corpus to be representative for the source.

- *Development of annotation scheme for Argumentative Zoning* (section 3.3): we have defined an annotation scheme for the argumentative status of sentences which is consistent and informative. The reproducibility and stability of the annotation scheme was evaluated by an experiment with two unrelated, task trained human annotators (section 4.3).

- *Implementation of a prototype system for automatic Argumentative Zoning*: we have provided evidence that this annotation scheme can be automatically applied (chapter 5). The prototype uses supervised learning on the basis of the previously hand-annotated corpus. The approach relies on corpus-based robust features, well-known from traditional text extraction work, but it is accompanied by a new, more linguistically motivated pattern matching to find prototypical agents and actions.

We have argued in chapter 2 that RDPs (Rhetorical Document Profiles) are document profiles which are specially useful for partially informed readers in a DR environment, and that they can be used for the production of tailored summaries and more informative citation information. Argumentative Zoning, as explored in this thesis, is a necessary and useful subtask for the generation of RDPs; however, this thesis does not accomplish the generation of RDPs. In the next section, we will sketch which tasks still need to be done in order to construct RDPs.

6.2. Future Work

6.2.1. RDP Generation

One avenue of future work is obvious: the algorithm for actually creating RDPs is not implemented yet. However, we have already given the outline of the two main parts of the algorithm:

- Determination of most appropriate slot fillers (in section 2.1.1);
- Association of identifiers of other approaches with the sentence expressing author's stance (in section 3.4). More advanced approaches for this subtask are discussed in the following.

Similarity matching between sentences could be used to determine the best filler for those slots which are filled by entire sentences (e.g. BACKGROUND). Different similarity measures are imaginable, from simple surface based algorithms like the Longest Common Substring as used by us in earlier work (cf. section 4.1.2.2), to more complicated ones like LIKEIT (Yianilos, 1997). Similarity as defined by vector space models is another option (Salton, 1971). One could, however, apply a deeper approach based on agent and action comparison, similar to Barzilay et al.'s (1999) work, and we would advocate this.

Given the stage of development reached in the thesis, extrinsic evaluation would be premature. Eventually, we envisage a task-based evaluation scenario, where the performance of subjects using RDPs for a certain task (e.g. question answering or relevance decision) is compared to a control group working with sentence extracts, and a group working with full documents. Such evaluation needs a clear definition of the task of information foraging for uninformed readers. The right task definition is not easy to find, particularly as user studies concentrating on this user group are rare (chapter 2). We are convinced at this point that simple relevance decision is underdefined and cannot be used as a task; we expect that a clearer picture of the best task for extrinsic evaluation will emerge during the actual generation of RDPs.

6.2.2. Improving the Prototype

We have shown in chapter 5 that it is possible to find patterns in the extracted sentential features with a relatively simple implementation and simple statistical techniques. As a result, our system can simulate human annotation behaviour to a certain degree. However, there are many aspects in which the existing prototype could be improved.

One could imagine a cascading system which performs an analysis of the agent-and-action structure of the text prior to the classification of the full annotation scheme. The first step, the attribution of intellectual ownership, could be learned from text annotated with the basic annotation scheme, by associating the patterns with agents (US_AGENT—THEM_AGENT—GENERAL_AGENT). In a second step, the finer distinctions could be applied.

In a cascading system, the high-precision rules described in section 5.3.5 could act as "sure-fire" rules: evidence of different levels of certainty could be collected before a statistically-based search, and "sure-fire" rules could provide the starting point, similar to the system presented by Mikheev et al. (1998).

In particular the actions are a topic which requires more research. We have created the action lexicon (figure 5.8; page 195) manually, based only on our intuitions after inspecting the corpus. But no clear methodology for creating the lexicon has emerged yet. We would like to perform tests varying the verbs included in the action lexicon and the classes assigned. Independent information sources like Levin's (1993) alternation classes, or WordNet (Klavans and Kan, 1998) could be used. And a more systematic way to create this lexicon would be to use learning in a bottom-up way.

We observed problems with verbal ambiguity: the same verbs are sometimes used in a meta-discourse interpretation and sometimes not. This is illustrated by the following examples:

CONTINUATION_ACTION:
For our analysis of gapping, we follow Sag (1976) in hypothesizing [...] (S-38, 9405010)

Not a CONTINUATION_ACTION:
From this or-node we follow an arc labelled Id [...] (S-73, 9405022)

CONTRAST_ACTION:
Hobbs' ordering of entities from a previous utterance varies from Brennan et al.'s [...] (S-104, 9410006)

Not a CONTRAST_ACTION:
The number of test contexts varies from word to word [...] (S-78, 9503025)

The examples seem to imply that an analysis of the syntactic context, in this case, the direct object, might help, but we fear the problem lies deeper. Given that we want to avoid the need for full text comprehension, traditional Word Sense Disambiguation (Schütze, 1998; Yarowsky, 1995) might help.

Apart from verbal polysemy, there are some other specific concepts which supposedly indicate meta-discourse, but which are problematic for our approach, e.g. "*goals*", "*topic*" and "*similarity*". These concepts are used at the object level (*science*) in some papers, e.g. in logic programming, discourse modelling and in statistical NLP:

The speaker attempts to achieve this goal by building a description of the object that she believes will give the hearer the ability to identify it when it is possible to do so. (S-6, 9405013)

The substructure check makes only sense if the semantics <EQN/> of the current goal is instantiated. (S-69, 9405004)

The sentential topic Hanako is the only possible antecedent of this zero subject in this example. (S-13S, 9405028)

In those models, the relationship between given words is modeled by analogy with other words that are in some sense similar to the given ones. (S-11, 9405001)

In experiments not reported here in detail, we have tried to ameliorate this problem by excluding those `Ag-1`, `Ag-2` and `Formu` patterns which contain "characteristic" words for this document, as determined by a *tf/idf* measure. The idea was that if a phrase which we intended to indicate meta-discourse occurred far more often than expected in a given document, then there was a chance that it is a concept at an object level. However, these experiments did not result in higher recognition results. We have to conclude that this is another problem which requires further enquiry.

Finding identifiers of other work is important for building RDPs (cf. above). Whereas this task is easy in the cases where a formal citation is present, it is much harder to identify well-known names of solutions in text, e.g. as in the following sentence:

> *I argue that Hidden Markov Models are unsuited to the task* [. . .]
> (S-9, 941002)

Only later in the text, *"Hidden Markov Models"* are associated with particular researchers:

> *Hidden Markov Models (HMMs) (Huang et al., 1990) offer a powerful statistical approach to this problem* [. . .] (S-24, 941002)

However, the identification of "*Hidden Markov Models*" as a solution name would have several advantages in this context:

- The names would be fillers of the RDP slots "SOLUTION ID" (parts of the complex slots BASIS/CONTINUATION and RIVAL/CONTRAST). Such a characterization of other work is more informative than formal citations in many cases, as names of solutions have more continuity than single papers and single researchers.

- A list of such names could help the uninformed reader acquire an overview of the field (cf. chapter 1). Names of commonly advocated solutions might help identify schools of thought, in this case, groups of researchers who have invented Hidden Markov Models or who work with them. Named problems, e.g. *"data sparseness"* also occur frequently in our texts, and their identification would be similarly useful to uninformed readers.

- Identifying names of solutions would help improve the agent feature, as researchers' names are often substituted with (named) approaches or solutions they are well-known for. At the moment, the sentence above would not be classified as part of prototypical argumentation, because the agent is not recognized as THEM_AGENT, but if the authors had used the expression "*Huang et al.'s (1990) approach*" it would. This lack of parallelism makes the method less robust towards writing style.

Recent advances in named entity recognition have made the association task technically feasible, cf. the results of the Named Entity Recognition Task in MUC-7, where F-measures are in the range of 93% for domain-specific text (MUC-7, 1998).

Note that there are typically contexts in the article where the association of "THEM" or "US" with a solution name is easier than in other contexts. Consider the following sentence:

> *LHIP provides a processing method which allows selected portions of the input to be ignored or handled differently.* (S-5, 9408006)

This sentence (and the role of *"LHIP"* in the argumentation) can only be understood in the context of a sentence several sentences earlier:

> *This paper describes LHIP (Left-Head Corner Island Parser), a parser designed for broad-coverage handling of unrestricted text.* (S-0, 9408006)

The sentence would have to be interpreted completely differently in the context of the following (imaginary) sentence:

> *Gold et al. (1989) introduced LHIP (Left-Head Corner Island Parser), a parser designed for broad-coverage handling of unrestricted text.*

Recognition of *"LHIP"* in close proximity with the phrase *"in this paper"* could add *"LHIP"* to a list of solutions associated with the authors, whereas in the other (fictional) case, it would have been added to a list of approaches associated with Gold et al. (THEM_AGENT).

There is one other possibility how agent recognition could be made more robust, and that is by anaphora resolution. As reported in section 5.2.2.2, not all agent classes are ambiguous. In fact, in many of them, interpretation is unambiguous (THEM_AGENT, US_AGENT); in others, we have found a strong tendency that the intended interpretation is almost always present (TEXTSTRUCTURE_AGENT, OUR_AIM_AGENT, US_PREVIOUS_AGENT, REF_US_AGENT, GAP_AGENT, SOLUTION_AGENT, PROBLEM_AGENT). However, a high level of ambiguity is associated with the classes REF_US_AGENT, THEM_PRONOUN_AGENT, AIM_REF_AGENT, REF_AGENT. Most of these ambiguities are between US_AGENT and THEM_AGENT, but the agent class THEM_PRONOUN_AGENT is actually ambiguous between THEM_AGENT and any plural objects in the scientific domain the paper is talking about, e.g. rules, arcs, probabilities. Examples for correct and incorrect interpretation of THEM_PRONOUN_AGENTs can be found in appendix B.7; p. 300. For example, agents no. 4 and 16 have the wrong interpretation.

We performed a simulation experiment to determine the distribution of US_AGENT, THEM_AGENT and GENERAL_AGENT for the most frequent of the ambiguous classes, REF_AGENT. There were 632 occurrences of REF_AGENT in the corpus (only 586 of which were used in the Naive Bayesian classification and the symbolic rules; the others were not the first agent in the sentence). We wanted to determine if anaphora resolution prior to classification would improve end results, so we manually simulated a perfect anaphora resolution algorithm by classifying the phrases by their referent: 436 (69%) of the 632 REF_AGENTs were classified as US_AGENT, 175 (28%) as THEM_AGENT, and 20 (3%) as GENERAL_AGENT.

As a result of this manual disambiguation, the performance of the `Ag-1` feature for the Naive Bayesian model increased dramatically from K=.07 to K=.14, making it the third best feature after `Cit-1` (K=.18) and `Loc` (K=.17); cf. figure 5.33 (p. 222). Classification results using the 14 successful features increased from K=.39 to K=.42. These results are surprisingly good, considering that we removed only one ambiguous class. Even though a practical anaphora resolution model would not achieve 100% correctness as we did in our simulation, our experiment still points to the fact that good anaphora resolution would make statistical classification less noisy by potentially removing the need for ambiguity classes, and that it could potentially be of great value for automatic Argumentative Zoning.

6.2.3. Learning Meta-discourse Expressions

The current experiments have shown that sentential features, particularly meta-discourse phrases, can help us perform Argumentative Zoning. It is a practical problem of how to arrive at good patterns other than manually generating them. There are some approaches which learn cue phrases automatically from text, either by ngram-techniques (Samuel et al., 1998, 1999) or by *tf/idf* style frequency techniques (Hovy and Lin, 1999; Hovy and Liu, 1998). Learning would be particularly useful for the clustering of values, which we have so far done manually. We performed some experiments with n-grams over words as approximations for indicator phrases (Teufel, 1998); these experiments showed over-fit and were thus not conclusive.

We take this as an indication that our corpus is still too small to automatically learn good patterns. The learning of agent and action patterns, however, is planned for the future, when our corpus of scientific articles will hopefully be expanded considerably.

6.2.4. Redefining the Annotation Task

The task of Argumentative Zoning could be refined by using a more fine-grained unit of annotation and classification. Currently, we use *sentences*; part of the reason for this decision was practical, as sentence boundary disambiguators like the one we use work very reliably. However, we came across many examples where a border between two argumentative zones cuts across a sentence:

> *However, this is not very satisfactory because one of the goals of our work is precisely to avoid the problems of data sparseness by grouping words into classes.* (S-41, 9408011)
>
> *While we know of previous work which associates scores with feature structures (Kim, 1994)* [sic] *are not aware of any previous treatment which makes explicit the link to classical probability theory.* (S-9, 9502022)

In the first case, there is a borderline between a CONTRAST and an AIM zone which cuts across the sentence, in the second between an OTHER and CONTRAST zone. Cases like this confuse both symbolic and stochastic accounts of Argumentative Zoning, as correlates of both zones can be found in the sentence, but only one target outcome is annotated.

Our experience with the heuristics for action and agent detection in sections 5.3.3.7 have shown that it is theoretically possible to dissect the sentence into clause-like units—though we have so far used this information only for feature determination. These heuristics rely only on the most likely finite verbs in the sentence as determined by a POS-Tagger. Even though a definition of a clause as centered around a finite verb is simplistic (cf. also the discussions in section 3.5 in the context of RST), and even though such heuristics are not correct in all cases, we nevertheless argue that a clause-based approach would have advantages for Argumentative Zoning. The finer unit of annotation is intuitively more appealing, as clauses map more directly to propositions. A move towards the clause would thus be a move towards a slightly deeper representation.

Another way to improve the task of Argumentative Zoning would be to ask the subjects to indicate a relevance-level (or confidence-level) for the annotation of each sentence. This would indicate how well suited the sentence is to serve as an RDP slot. Of course, such instructions would result in a higher training effort, but would also provide us with a more valuable gold standard for the task.

6.2.5. Application to a Different Domain

Finally, we take a look at the kinds of texts treated. We have assumed that argumentative moves and zones are to be expected in *all* scientific research articles, as they are based on the function associated with the text type, i.e. the goal of justifying the validity of the research presented. We have concluded from this that our annotation scheme should in principle apply to all kinds of scientific research articles. One of the reasons for choosing computational linguistics articles was the interdisciplinary nature of the field, which would make the corpus a difficult test bed. Nevertheless, our claim would find a more rigorous verification if we could successfully apply the analysis to texts of a different domain.

It is plausible that some of the meta-discourse we found is specific to our corpus. Research by Hyland (1998) confirms that there are differences in meta-discourse between domains. In that case, an approach which learns new cue phrases from text, as mentioned above, would be particularly useful for porting our implementation to a new domain.

It might also be the case that our young, interdisciplinary domain contains particularly many argumentative moves of explicit comparison. In such domains, contrast with other researchers and intellectual ancestry is very important, as there are many methodologies, which are often identified by similarities to and contrast with existing ones. It might thus be the case that other domains do not express comparisons to other work as overtly as our texts do.

We have used *conference* articles in this thesis. Practical reasons have kept us from using journal articles as data so far: the difficulty of corpus collection due to copy right problems, and due to the increased length and subsequent time effort of human experiments. In principle, however, we are particularly interested in journal articles, for several reasons. On the one hand, they can be expected to be of higher textual quality, as they are more rigorously edited. On the other hand, as journal articles are much longer, they pose a particularly difficult problem for current summarization approaches, as these do not take large-scale discourse structure into account. As the scientific argumentation in journal articles is basically the same as in conference articles, we are confident that our scheme should be applicable to journal articles at least as consistently as to conference articles.

Bibliography

Abney, Steven. 1990. Rapid incremental parsing with repair. In *Proceedings of the 6th New OED Conference*, 1–9.

Abracos, Jose, and Gabriel Pereira Lopes. 1997. Statistical methods for retrieving most significant paragraphs in newspaper articles. In Mani and Maybury 1997, 51–57.

ACP online. 1997. Annals Extracts. `http://www.acponline.org/journals/annals/01apr97/extracts/extractintro.%htm`.

Adhoc. 1987. Ad Hoc Working Group For Critical Appraisal Of The Medical Literature. A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine* 106: 508–604.

Adler, Annette, Anuj Gujar, Beverly L. Harrison, Kenton O'Hara, and Abigail Sellen. 1998. A diary study of work-related reading: Design implications for digital reading devices. In *Proceedings of CHI-98, ACM*, 241–248.

Alexandersson, Jan, Elisabeth Maier, and Norbert Reithinger. 1995. A robust and efficient three-layered dialogue component for a speech-to-speech translation system. In *Proceedings of the Seventh Meeting of the European Chapter of the Association for Computational Linguistics*, 188–193.

Alley, Michael. 1996. *The Craft of Scientific Writing*. Englewood Cliffs, NJ: Prentice-Hall.

Alterman, Richard. 1985. A dictionary based on concept coherence. *Artificial Intelligence* 25(2): 153–186.

ANSI. 1979. American National Standard for Writing Abstracts. Technical report, American National Standards Institute, Inc., New York, NY. ANSI Z39.14.1979.

Arndt, Kenneth A. 1992. The informative abstract. *Archives of Dermatology* 128(1): 101.

Baldwin, Breck, and Tom Morton. 1998. Dynamic coreference-based summarization. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3).*

Baldwin, Breck, Tom Morton, Amit Bagga, Jason Baldridge, Raman Chandrasekar, Alexis Dimitriadis, Kieran Snyder, and Magdalena Wolska. 1998. Description of the UPenn CAMP system as used for coreference. In MUC-7 1998.

Barzilay, Regina, and Michael Elhadad. 1999. Using lexical chains for text summarization. In Mani and Maybury 1999, 111–121.

Barzilay, Regina, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, 550–557.

Bates, Marcia J. 1998. Indexing and access for digital libraries and the internet: Human, database and domain factors. *Journal of the American Society for Information Science* 49: 1185–1205.

Baxendale, Phyllis B. 1958. Man-made index for technical literature—an experiment. *IBM Journal of Research and Development* 2(4): 354–361.

Bazerman, Charles. 1985. Physicists reading physics, schema-laden purposes and purpose-laden schema. *Written Communication* 2(1): 3–23.

Bazerman, Charles. 1988. *Shaping Writing Knowledge*. Madison, WI: University of Wisconsin Press.

Belkin, N. 1980. Anomalous states of knowledge as a basis for information retrieval. *The Canadian Journal of Information Science* 5: 133–143.

Biber, Douglas, and E. Finegan. 1994. Intra-textual variation within medical research articles. In Oostdijk and de Haan, eds., *Corpus-Based Research into Language*, chapter 13, 201–221. Amsterdam: Rodoph.

Blicq, Ron. 1983. *Technically-write!: Communicating in a Technological Era*. Scarborough, Ont.: Prentice-Hall Canada.

Boguraev, Branimir, and Christopher Kennedy. 1999. Salience-based content characterization of text documents. In Mani and Maybury 1999, 99–110.

Bonzi, Susan. 1982. Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science* 33(4): 208–216.

Borgman, Christine L. 1996. Why are online catalogs still hard to use? *Journal of the American Society for Information Science* 47: 493–503.

Borko, Harold, and C. L. Bernier. 1975. *Abstracting Concepts and Methods.* San Diego, CA: Academic Press.

Borko, Harold, and Seymour Chatman. 1963. Criteria for acceptable abstracts: A survey of abstractors' instructions. *American Documentation* 14(2): 149–160.

Brandow, Ronald, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management* 31(5): 675–685.

British Telecom. 1998. `http://transend.labs.bt.com/cgi-bin/prosum/prosum`.

Broer, J. W. 1971. Abstracts in block diagram form. *IEEE Transactions on Engineering Writing and Speech* 14(2): 64–67.

Brooks, Terrence A. 1986. Evidence of complex citer motivations. *Journal of the American Society for Information Science* 37: 34–36.

Brouwer, M., C. C. Clark, G. Gerbner, and K. Krippendorff. 1969. The television world of violence. In *Mass Media and Violence: A Report to the National Commission on the Causes and Prevention of Violence*, 311–339 and 519–591. Washington, D.C.: Government Printing Office. Cited after Krippendorff:80.

Brown, Ann L., and Jeanne D. Day. 1983. Macrorules for summarizing text: The developments of expertise. *Journal of Verbal Learning and Verbal Behaviour* 22: 1–14.

Brown, Penelope, and Levinson Stephen C. 1987. *Politeness: Some Universals in Language Usage.* Cambridge, England: Cambridge University Press.

Busch-Lauer, Ines A. 1995. Abstracts in German medical journals: A linguistic analysis. *Information Processing and Management* 31(5): 769–776.

Buxton, A. B., and A. J. Meadows. 1978. Categorization of the information in experimental papers and their author abstracts. *Journal of Research in Communication Studies* 1: 161–182.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22(2): 249–254.

Carletta, Jean, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics* 23(1): 13–31.

Chalmers, Matthew, and Paul Chitson. 1992. Bead: Explorations in information visualization. In *Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval (SIGIR-92)*, 330–337.

Charney, Davida. 1993. A study in rhetorical reading—How evolutionists read "The Spandrels of San Marco". In Jack Selzer, ed., *Understanding Scientific Prose*. Madison, WI: The University of Wisconsin Press.

Chinchor, Nancy A., and Elaine Marsh. 1998. *MUC-7 Information Extraction Task Definition*. DARPA. `www.muc.saic.com/proceedings/muc_7_toc.html`.

Chubin, Daryl E., and S. D. Moitra. 1975. Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science* 5(4): 423–441.

Cleverdon, Cyril W. 1984. Optimizing convenient online access to bibliographic databases. *Information Services and Use* 4: 37–47.

Clove, J. F., and B. C. Walsh. 1988. Online text retrieval via browsing. *Information Processing and Management* 24(1): 31–37.

Clyne, Michael. 1987. Cultural differences in the organization of academic texts. *Journal of Pragmatics* 11: 211–247.

CMP_LG. 1994. The Computation and Language E-Print Archive, `http://xxx.lanl.gov/cmp-lg`.

Cohen, Robin. 1987. Analyzing the structure of argumentative discourse. *Computational Linguistics* 13: 11–24.

Cohen, William W. 1995. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, 115–123.

Cohen, William W. 1996. Learning trees and rules with set-valued features. In *Proceedings of AAAI-96*.

Conway, William D. 1987. *Essentials of Technical Writing*. Rexburg, Idaho: TechWrite Press. 4th ed.

Cremmins, Edward T. 1996. *The Art of Abstracting*. Arlington, VA: Information Resources Press, 2nd edn.

Crookes, Graham. 1986. Towards a validated analysis of scientific text structure. *Applied Linguistics* 7(1): 57–70.

Day, Robert A. 1995. *How to Write and Publish a Scientific Paper*. Cambridge, England: Cambridge University Press, 4th edn.

DeJong, Gerald F. 1982. An Overview of the FRUMP system. In Wendy G. Lehner and Ringle, eds., *Strategies for Natural Language Processing*, chapter 5. Hillsdale NJ: Lawrence Erlbaum.

Dillon, Andrew. 1992. Reading from paper versus from screens: A critical review of the empirical literature. *Ergonomics* 35(10): 1297–1326.

Dillon, Andrew, John Richardson, and Cliff McKnight. 1989. Human factors of journal usage and the design of electronic text. *Interacting with Computers* 1(2): 183–189.

Drakos, Nikos. 1994. From Text to Hypertext: A Post-Hoc Rationalisation of LaTeX2HTML. In *The Proceedings of the First WorldWide Web Conference*.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61–74.

Duszak, Anna. 1994. Academic discourse and intellectual styles. *Journal of Pragmatics* 21: 291–313.

Earl, Lois L. 1970. Experiments in automatic extracting and indexing. *Information Storage and Retrieval* 6(6): 313–334.

Edmundson, H. P. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery* 16(2): 264–285.

Edmundson, H. P. et al. 1961. *Final Report on the Study for Automatic Abstracting*. Canoga Park, CA: Thompson Ramo Wooldridge.

Elhadad, Michael. 1993. Using Argumentation to Control Lexical Choice: A Unification-Based Implementation. Ph.D. thesis, Computer Science Department, Columbia University, New York, NY.

Ellis, D. 1989a. A behavioural approach to information system design. *Journal of Documentation* 45(3): 171–212.

Ellis, D. 1989b. A behavioural model for information system design. *Journal of*

Information Science 15(4): 237–247.

Ellis, David. 1992. The physical and cognitive paradigms in information retrieval research. *Journal of Documentation* 48: 45–64.

Endres-Niggemeyer, Brigitte, Elisabeth Maier, and Alexander Sigel. 1995. How to implement a naturalistic model of abstracting: Four core working steps of an expert abstractor. *Information Processing and Management* 31(5): 631–674.

Farr, A. D. 1985. *Science Writing for Beginners*. Oxford: Blackwell Scientific Publications.

Fidel, R. 1985. Moves in online searching. *Online Review* 9(1): 61–74.

Fidel, R. 1991. Searchers' selection of search keys. *Journal of the American Society for Information Science* 42(7): 490–527.

Finch, Steven, and Andrei Mikheev. 1995. Towards a workbench for acquisition of domain knowledge from natural language. In *Proceedings of the 7th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-95)*, 194–201.

Francis, W. Nelson, and Henry Kucera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston, MA: Houghton Mifflin.

Froom, P., and J. Froom. 1993. Deficiencies in structured medical abstracts. *Journal of Clinical Epidemiology* 46: 591–594.

Frost, Carolyn O. 1979. The use of citations in Literary Research: A preliminary Classification of Citation Functions. *Library Quarterly* 49: 405.

Garfield, Eugene. 1979. *Citation Indexing: Its Theory and Application in Science, Technology and Humanities*. New York, NY: J. Wiley.

Garfield, Eugene. 1996. The significant scientific literature appears in a amall group of journals. *The Scientist* 10(17): 13–16. See also `http://165.123.34.41/yr1996/sept/research_960902.html`.

Georgantopoulos, Byron. 1996. Automatic Summarising Based on Sentence Extraction: A Statistical Approach. Master's thesis, Dept. of Linguistics, University of Edinburgh, Edinburgh, UK.

Giles, C. Lee, Kurt D. Bollacker, and Steve Lawrence. 1998. CiteSeer: An automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital*

Libraries, 89–98.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Boston, MA: Kluwer Academic Press.

Grishman, Ralph, and Beth Sundheim. 1995. Design of the MUC-6 evaluation. In *Proceedings of the Sixth Message Understanding Conference*, 1–11. DARPA, San Francisco, CA: Morgan Kaufmann Publishers.

Grosz, Barbara J., and Candace L. Sidner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics* 12(3): 175–204.

Grover, Claire, Andrei Mikheev, and Colin Matheson. 1999. LT TTT Version 1.0: Text Tokenisation Software. Technical report, Human Communication Research Centre, University of Edinburgh. `http://www.ltg.ed.ac.uk/software/ttt/`.

Hartley, James. 1997. Is is appropriate to use structured abstracts in social science journals? *Learned Publication* 10(4): 313–317.

Hartley, James, and Matthew Sydes. 1997. Are structured abstracts easier to read than traditional ones? *Journal of Research in Reading* 20(2): 122–136.

Hartley, James, Matthew Sydes, and Antony Blurton. 1996. Obtaining information accurately and quickly: are structured abstracts more efficient? *Journal of Information Science* 22(5): 349–356.

Haynes, R. B. 1990. More informative abstracts revisited. *Annals of Internal Medicine* 113: 69–76.

Hearst, Marti A. 1995. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 59–66.

Hearst, Marti A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1): 33–64.

Hearst, Marti A., and Jan O. Pedersen. 1996. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR-96)*, 76–84.

Herner, Saul. 1959. Subject slanting in scientific abstracting publications. In *Proceedings on the International Conference on Scientific Information*, vol. 1, 407–427.

Hoey, Michael. 1979. *Signalling in Discourse.* No. 6 in Discourse Analysis Monograph. Birmingham, UK: University of Birmingham.

Horsella, Maria, and Gerda Sindermann. 1992. Aspects of scientific discourse: Conditional argumentation. *English for Specific Purposes* 11: 129–139.

Houp, Kenneth W., and T. E. Pearsall. 1988. *Reporting Technical Information.* New York, NY: Maxwell Macmillan International, 6th edn.

Hovy, Eduard H. 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence* 63: 341–385.

Hovy, Eduard H., and Chin-Yew Lin. 1999. Automated text summarization in SUMMARIST. In Mani and Maybury 1999, 81–94.

Hovy, Eduard H., and Hao Liu. 1998. Personal Communication.

Hwang, Chung Hee, and Lenhart K. Schubert. 1992. Tense trees as the "fine structure" of discourse. In *Proceedings of 30th Annual Meeting of the Association for Computational Linguistics (ACL-92)*, 232–240.

Hyland, Ken. 1998. Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics* 30(4): 437–455.

Ingwersen, Peter. 1996. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation* 52: 3–50.

InXight. 1999. `http://www.inxight.com/Products/Enterprise/SummServ.html`.

ISI. 1999. Institute for Scientific Information, `http://www.isinet.com/products/citation/citssci.html`.

ISO. 1976. Documentation—Abstracts for Publication and Documentation. ISO 214-1976. Technical report, International Organisation for Standardisation.

Iwanska, L. 1985. Discourse Structure in Factual Reports. Technical report, GE Artificial Intelligence Laboratory, NY. Unpublished.

Johnson, Frances C., Chris D. Paice, William J. Black, and A. P. Neal. 1993. The application of linguistic processing to automatic abstract generation. *Journal of Document and Text Management* 1(3): 215–241.

Jordan, M. P. 1984. *Rhetoric of Everyday English Texts.* London, UK: George Allen and Unwin.

Jurafsky, Daniel, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. TR–97-02. Technical report, Institute of Cognitive Science, University of Colorado at Boulder, Boulder, CO.

Kan, Min-Yen, Judith L. Klavans, and Kathleen R. McKeown. 1998. Linear Segmentation and Segment Significance. In *Proceedings of the Sixth Workshop on Very Large Corpora (COLIN G/ACL-98)*, 197–205.

Kando, Noriko. 1997. Text-level structure of research papers: Implications for text-based information processing systems. In *Proceedings of BCS-IRSG Colloquium*, 68–81. Also available from `http://www.rd.nacsis.ac.jp/~kando/kando.ps`.

Kessler, Myer Mike. 1963. Bibliographic coupling between scientific papers. *American Documentation* 14(1): 10–25.

Kilgarriff, Adam. 1999. 95% replicability for manual word sense tagging. In *Proceedings of the 8th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99), Poster Session*, 277.

Kintsch, Walter, and Teun A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review* 85(5): 363–394.

Kircz, Joost G. 1991. The rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of Documentation* 47(4): 354–372.

Kircz, Joost G. 1998. Modularity: The next form of scientific information presentation? *Journal of Documentation* 54: 210–235.

Klavans, Judith L., and Min-Yen Kan. 1998. Role of verbs in document analysis. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (ACL/COLING-98)*, 680–686.

Klavans, Judith L., Kathleen R. McKeown, and Susan Lee. 1998. Resources for evaluation of summarization techniques. In *Proceedings of First International Conference on Language Resources and Evaluation.*

Kleinberg, Jon. 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*. Also available from `http://www.cs.cornell.edu/home/kleinber/`.

Knott, Alistair. 1996. A Data-Driven Methodology for Motivating a Set of Discourse Relations. Ph.D. thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh, UK.

Kozima, Hideki. 1993. Text segmentation based on similarity between words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93)*, 286–288.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to its Methodology*. Beverly Hills, CA: Sage Publications.

Krohn, Uwe. 1995. Visualization of navigational retrieval in virtual information spaces. In *Proceedings of the Workshop on New Paradigms in Information Visualization and Manipulation*, 26–32.

Kupiec, Julian, Jan O. Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-95)*, 68–73.

Lancaster, Frederick Wilfrid. 1998. *Indexing and Abstracting in Theory and Practice*. London, UK: Library Association.

Lannon, John M. 1993. *Technical Writing*. New York, NY: HarperCollins Publishers, 6th edn.

Latex2Html. 1999. `http://cbl.leeds.ac.uk/nikos/tex2html/doc/latex2html/latex2html.html`.

Latour, Bruno, and Steven Woolgar. 1986. *Laboratory Life: The Social Construction of Scientific Facts*. Beverley Hills, CA: Sage Publications.

Lawrence, Steve, C. Lee Giles, and Kurt Bollacker. 1999. Digital libraries and autonomous citation indexing. *IEEE Computer* 32(6): 67–71.

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In J. Svartvik, ed., *Directions in Corpus Linguistics*, 105–122. Berlin: Mouton de Gruyter.

Levin, Beth. 1993. *English Verb Classes and Alternations*. Chicago, IL: University of Chicago Press.

Levy, D. M. 1997. I read the news today, oh boy: Reading and attention in digital libraries. In *Proceedings of Digital Libraries T97, ACM*, 228–235.

Liddy, Elizabeth DuRoss. 1991. The discourse-level structure of empirical abstracts:

An exploratory study. *Information Processing and Management* 27(1): 55–81.

Litman, Diane J. 1996. Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research* 5: 53–94.

Longacre, Robert E. 1979. The paragraph as a grammatical unit. In Talmy Givon, ed., *Syntax and Semantics: Discourse and Syntax*, vol. 12, 115–134. New York NY: Academic Press.

Luhn, Hans Peter. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2): 159–165.

Luukkonen, Terttu. 1992. Is scientists' publishing behaviour reward-seeking? *Scientometrics* 24: 297–319.

MacRoberts, Michael H., and Barbara R. MacRoberts. 1984. The negational reference: Or the art of dissembling. *Social Studies of Science* 14: 91–94.

Maier, Elisabeth, and Eduard H. Hovy. 1993. Organizing discourse structure relations using metafunctions. In H. Horacek and M. Zock, eds., *New Concepts in Natural Language Generation: Planning, Realization, and Systems*, 69–86. London, UK: Pinter.

Maizell, R. E., J. F. Smith, and T. E. R. Singer. 1971. *Abstracting Scientific and Technical Literature: An Introductory Guide and Texts for Scientists, Abstractors and Management*. New York, NY: Wiley-Interscience.

Malcolm, L. 1987. What rules govern tense usage in scientific articles? *English for Specific Purposes* 6: 31–43.

Mani, Inderjeet, and Eric Bloedorn. 1998. Machine learning of generic and user-focused summarization. In *Proceedings of the Fifteenth National Conference on AI (AAAI-98)*, 821–826.

Mani, Inderjeet, and Mark T. Maybury, eds. 1997. *Proceedings of the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization.*

Mani, Inderjeet, and Mark T. Maybury, eds. 1999. *Advances in Automatic Text Summarization.* Cambridge, MA: MIT Press.

Mann, William C., and Sandra A. Thompson. 1987. Rhetorical Structure Theory: A Theory of Text Organisation. ISI/RS-87-190. Technical report, Information Sciences Institute, University of Southern California, Marina del Rey, CA.

Mann, William C., and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organisation. *Text* 8(3): 243–281.

Manning, Alan D. 1990. Abstracts in relation to larger and smaller discourse structures. *Journal of Technical Writing and Communication* 20(4): 369–390.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcu, Daniel. 1997a. From discourse structures to text summaries. In Mani and Maybury 1997, 82–88.

Marcu, Daniel. 1997b. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. Ph.D. thesis, University of Toronto, Ont., Canada.

Marcu, Daniel. 1999a. A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, 365–372.

Marcu, Daniel. 1999b. Discourse structures are good indicators of importance in text. In Mani and Maybury 1999, 123–136.

Maron, M. E., and J. L. Kuhns. 1960. On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery* 7: 216–244.

Mathes, John C., and Dwight W. Stevenson. 1976. *Designing Technical Reports—Writing for Audiences in Organizations*. Indianapolis, IN: Bobbs-Merrill Educational Publishing.

Mauldin, Michael L. 1991. Retrieval performance in FERRET: A conceptual information retrieval system. In *Proceedings of the 14th Annual International Conference on Research and Development in Information Retrieval (SIGIR-91)*, 347–355.

McGirr, Clinton J. 1973. Guidelines for abstracting. *Technical Communication* 25(2): 2–5.

McKeown, Kathleen R., Karen Kukich, and James Shaw. 1994. Practical issues in automatic document generation. In *Proceedings of ANLP-94 (Applied Natural Language Processing)*, 7–14.

McKeown, Kathleen R., and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-95)*, 74–82.

Michaelson, Herbert B. 1980. *How to Write and Publish Engineering Papers and Reports*. Phoenix, AZ: Oryx Press.

Microsoft. 1997. Office-97. `http://www.microsoft.com/Office/`.

Miike, Seijii, Etsuo Itoh, Kenji Ono, and Kazuo Sumita. 1994. A full-text retrieval system with a dynamic abstract generation function. In *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval (SIGIR-94)*, 152–163.

Mikheev, Andrei. To Appear. Feature Lattices and Maximum Entropy Models. *Journal of Machine Learning* Available from `http://www.ltg.ed.ac.uk/~mikheev/papers.html`.

Mikheev, Andrei, Claire Grover, and Marc Moens. 1998. Description of the LTG system used for MUC-7. In MUC-7 1998.

Milas-Bracovic, Milica. 1987. The structure of scientific papers and their author abstracts. *Informatologia Yugoslavica* 19(1–2): 51–67.

Minel, Jean-Luc, Sylvaine Nugier, and Gerald Piat. 1997. How to appreciate the quality of automatic text summarization. In Mani and Maybury 1997, 25–30.

Minsky, M. 1975. A framework for representing knowledge. In P. Winston, ed., *The Psychology of Computer Vision*. New York, NY: McGraw-Hill.

Mitchell, John Howard. 1968. *Writing for Professional and Technical Journals*. New York, NY: J. Wiley.

Moore, Johanna D., and Cecile L. Paris. 1993. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics* 19: 651–694.

Moravcsik, Michael J., and Poovanalingan Murugesan. 1975. Some results on the function and quality of citations. *Social Studies of Science* 5: 88–91.

Morris, Andrew H., George M. Kasper, and Dennis A. Adams. 1992. The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research* 3(1): 17–35.

Morris, Jane, and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17: 21–48.

Moser, Megan G., and Johanna D. Moore. 1996. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics* 22(3): 409–420.

MUC-7. 1998. *Proceedings of the Seventh Message Understanding Conference.* DARPA. `www.muc.saic.com/proceedings/muc_7_toc.html`.

Mullins, Nicholas C., William E. Snizek, and Kay Oehler. 1988. The structural analysis of a scientific paper. In A. F. J. van Raan, ed., *Handbook of Quantitative Studies of Science and Technology*, 81–106. Amsterdam, NL: North-Holland.

Myers, Greg. 1992. In this paper we report...—speech acts and scientific facts. *Journal of Pragmatics* 17(4): 295–313.

Nanba, Hidetsugu, and Manabu Okumura. 1999. Towards multi-paper summarization using reference information. In *Proceedings of IJCAI-99*, 926–931. `http://galaga.jaist.ac.jp:8000/~nanba/study/papers.html`.

Nowell, Lucy Terry, Robert K. France, Deborah Hix, Lenwood S. Heath, and Edward A. Fox. 1996. Visualizing search result: Some alternatives to query-document similarity. In *Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR-96)*, 67–75.

Oakes, Michael, and Chris Paice. 1999. The automatic generation of templates for automatic abstracting. In *Proceedings of the 21st BCS IRSG Colloquium on IR.*

O'Connor, John. 1982. Citing statements: Computer recognition and use to improve retrieval. *Information Processing and Management* 18(3): 125–131.

Oddy, Robert Norman, Elizabeth DuRoss Liddy, B. Balakrichnan, A. Bishop, J. Elewononi, and E. Martin. 1992. Towards the use of situational information in information retrieval. *Journal of Documentation* 48: 123–171.

O'Hara, Kenton, and Abigail Sellen. 1997. A comparison of reading paper and on-line documents. In *Proceedings of CHI-97, ACM*, 335–342. Available from `http://www1.acm.org/sigchi/chi97/proceedings/paper/koh.htm`.

O'Hara, Kenton, F. Smith, W. Newman, and Abigail Sellen. 1998. Student reader's use of library documents: implications for library technologies. In *Proceedings of CHI-98, ACM*, 233–240.

Olsen, Kai A., Robert R. Korfhage, Kenneth M. Sochats, Michael B. Spring, and James G. Williams. 1993. Visualizing of a document collection: The VIBE system. *Information Processing and Management* 29(1): 69–81.

Ono, Kenji, Kazuo Sumita, and Seijii Miike. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, 344–348.

Oppenheim, Charles, and Susan P. Renn. 1978. Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science* 29: 226–230.

Oracle. 1993. Introduction to Oracle ConText. Technical report, Oracle Corporation, Redwood Shores, CA.

Paice, Chris D. 1981. The automatic generation of literary abstracts: an approach based on the identification of self-indicating phrases. In Robert Norman Oddy, Stephen E. Robertson, Cornelis Joost van Rijsbergen, and P. W. Williams, eds., *Information Retrieval Research*, 172–191. London, UK: Butterworth.

Paice, Chris D. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management* 26: 171–186.

Paice, Chris D., and A. Paul Jones. 1993. The identification of important concepts in highly structured technical papers. In *Proceedings of the 16th Annual International Conference on Research and Development in Information Retrieval (SIGIR-93)*, 69–78.

Paris, Cecile L. 1988. Tailoring Object Descriptions to a User's Level of Expertise. *Computational Linguistics* 14(3): 64–78. Special Issue on User Modelling.

Paris, Cecile L. 1993. Dagstuhl Seminar on Summarization webpage. `http://www.ik.fh-hannover.de/ik/projekte/Dagstuhl/Abstract/Answers/Pari%s/paris.html`.

Paris, Cecile L. 1994. User Modeling in Text Generation. *Computational Linguistics* 20(2): 318–321.

Perelman, Chaim, and Lucie Olbrechts-Tyteca. 1969. *The New Rhetoric, a Tractise on Argumentation*. Notre Dame, IN: University of Notre Dame Press.

Pinelli, Thomas E., Virginia M. Cordle, and Raymond F. Vondran. 1984. The function of report components in the screening and reading of technical reports. *Journal of Technical Writing and Communication* 14(2): 87–94.

Polanyi, Livia. 1988. A formal model of the structure of discourse. *Journal of Pragmatics* 12: 601–638.

Pollack, Martha E. 1986. A model of plan inference that distinguishes between the beliefs of actors and observers. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics (ACL-86)*, 207–214.

Pollock, Joseph J., and Antonio Zamora. 1975. Automatic abstracting research at the Chemical Abstracts service. *Journal of Chemical Information and Computer Sciences* 15(4): 226–232.

Radev, Dragomir R., and Eduard H. Hovy, eds. 1998. *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization.*

Radev, Dragomir R., and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics* 24(3): 469–500.

Rath, G.J, A. Resnick, and T. R. Savage. 1961. The formation of abstracts by the selection of sentences. *American Documentation* 12(2): 139–143.

Raynar, Jeffrey C. 1999. Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, 357–364.

Reed, Chris. 1999. The role of saliency in generating natural language arguments. In *Proceedings of IJCAI-99*, 876–881.

Reed, Chris, and Derek Long. 1998. Generating the structure of an argument. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (ACL/COLING-98)*, 1091–1097.

Rees, Alan M. 1966. The relevance of relevance to the testing and evaluation of document retrieval systems. *Aslib Proceedings* 18: 316–324.

Rennie, D., and R. M. Glass. 1991. Structuring abstracts to make them more informative. *Journal of the American Medical Association* 266(1): 116–117.

Richmond, Korin, Andrew Smith, and Einat Amitay. 1997. Detecting subject boundaries within text: A language independent statistical approach. In *The Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*.

Riley, Kathryn. 1991. Passive voice and rhetorical role in scientific writing. *Journal of Technical Writing and Communication* 21(3): 239–257.

Robertson, Stephen E., Steve Walker, Micheline M. Hancock-Beaulieu, Aaron Gull, and Marianna Lau. 1993. Okapi at TREC. In D. K. Harman, ed., *The first Text REtrieval Conference (TREC-1)*, 21–30.

Robin, Jacques. 1994. Revision-Based Generation of Natural Language Summaries Providing Historical Background. Ph.D. thesis, Computer Science Department, Columbia University, New York, NY.

Robin, Jacques, and Kathleen R. McKeown. 1996. Empirically designing and evaluating a new revision-based model for summary generation. *Artificial Intelligence* 85: 135–179.

Rowley, Jennifer. 1982. *Abstracting and Indexing.* London, UK: Bingley.

Salager-Meyer, Francoise. 1990. Discoursal flaws in medical English abstracts: A genre analysis per research- and text type. *Text* 10(4): 365–384.

Salager-Meyer, Francoise. 1991. Medical English abstracts: How well structured are they? *Journal of the American Society for Information Science* 42: 528–532.

Salager-Meyer, Francoise. 1992. A text-type and move analysis study of verb tense and modality distributions in medical English abstracts. *English for Specific Purposes* 11: 93–113.

Salager-Meyer, Francoise. 1994. Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes* 13(2): 149–170.

Salton, Gerard. 1971. Cluster search strategies and the optimization of retrieval effectiveness. In Gerard Salton, ed., *The SMART Retrieval System; Experiments in Automatic Document Processing*, 223–242. Englewood Cliffs, NJ: Prentice Hill.

Salton, Gerard, James Allan, Chris Buckley, and Amit Singhal. 1994a. Automatic analysis, theme generation, and summarisation of machine readable texts. *Science* 264: 1421–1426.

Salton, Gerard, and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval.* Tokyo: McGraw-Hill.

Salton, Gerard, Amit Singhal, Chris Buckley, and Mandar Mitra. 1994b. Automatic Text Decomposition Using Text Segments and Text Themes. Technical report, Cornell University.

Samuel, Ken, Sandra Carberry, and K. Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL/COLING-98)*, 1150–1156.

Samuel, Ken, Sandra Carberry, and K. Vijay-Shanker. 1999. Automatically selecting useful phrases for dialogue act tagging. In *Proceedings of the Pacific Association for Computational Linguistics (PACLING-99)*.

Samuels, S. J., R. Tennyson, L. Sax, P. Mulcahy, N. Schermer, and H. Hajovy. 1987. Adults' use of text structure in the recall of a scientific journal article. *Journal of Education Research* 81: 171–174.

Saracevic, Tefko. 1975. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science* 26(6): 321–343.

Saracevic, Tefko, Paul B. Kantor, A. Y. Chamis, and D. Trivison. 1988. A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science* 39(3): 161–176.

Schamber, Linda, Michael B. Eisenberg, and Michael S. Nilan. 1990. A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management* 26: 755–776.

Schank, Roger C., and Robert P. Abelson. 1977. *Scripts, Goals, Plans and Understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1): 97–124.

Sherrard, Carol. 1985. The psychology of summary writing. *Journal of Technical Writing and Communication* 15(3): 247–258.

Shum, Simon Buckingham. 1998. Evolving the web for scientific knowledge: First steps towards an "HCI knowledge web". *Interfaces, British HCI Group Magazine* 39: 16–21. Also available from `http://kmi.open.ac.uk/sbs/hciweb/Interfaces98.html`.

Shum, Simon Buckingham, Enrico Motta, and John Domingue. 1999. Representing scholarly claims in internet digital libraries: A knowledge modelling approach. In *Proceedings of ECDL'99: Third European Conference on Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science. Heidelberg, Germany: Springer Verlag.

Siegel, Sidney, and N. John Jr. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. Berkeley, CA: McGraw-Hill, 2nd edn.

SIGMOD. 1999. `http://www.acm.org/sigs/sigmod/sigmod99`.

Sillince, John Anthony Arthur. 1992. Literature searching with unclear objectives: A new approach using argumentation. *Online Review* 16(6): 391–410.

Skorochod'ko, E. F. 1972. Adaptive method of automatic abstracting and indexing. In *Information Processing 71*, vol. 2, 1179–1182. North-Holland.

Small, Henry G. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science* 24: 265–269.

Solov'ev, V. I. 1981. Functional characteristics of the author's abstract of a dissertation and the specifics of writing it. *Scientific and Technical Information Processing* 3: 80–88. English translation of *Nauchno-Tekhnicheskaya Informatsiya*, Seriya 1, Number 6, 1981, 20–24.

Spärck Jones, Karen. 1988. Tailoring Output to the User: What does User Modelling in Generation Mean? TR–158. Technical report, Computer Laboratory, University of Cambridge, Cambridge, UK.

Spärck Jones, Karen. 1990. What sort of thing is an AI experiment? In D. Partridge and Yorick Wilks, eds., *The Foundations of Artificial Intelligence: A Sourcebook*. Cambridge, UK: Cambridge University Press.

Spärck Jones, Karen. 1994. Discourse Modelling for Automatic Summarising, TR-290. Technical report, Computer Laboratory, University of Cambridge.

Spärck Jones, Karen. 1999. Automatic summarising: Factors and directions. In Mani and Maybury 1999, 1–12.

Spiegel-Rüsing, Ina. 1977. Bibliometric and content analysis. *Social Studies of Science* 7: 97–113.

Starck, Heather A. 1988. What do paragraph markings do? *Discourse Processes* 11(3): 275–304.

Strzalkowski, Tomek, Gees Stein, Jin Wang, and Bowden Wise. 1999. A robust practical text summarizer. In Mani and Maybury 1999, 137–154.

Sumita, Kazuo, Kenji Ono, Tetsuro Chino, Teruhiko Ukita, and Shin'ya Amaro. 1992. A discourse structure analyzer for Japanese text. In *Proceedings of the International Conference on Fifth Generation Computer Systems*.

Sumner, Tamara, and Simon Buckingham Shum. 1998. From documents to discourse: Shifting conceptions of scholarly publishing. In ACM Press, ed., *Proceedings of the CHI-98 ACM*, 95–102. New York, NY.

Suppe, Frederick. 1998. The structure of a scientific paper. *Philosophy of Science* 65: 381–405.

Swales, John. 1981. Aspects of Article Introductions. Aston ESP Research Project No. 1. Technical report, The University of Aston, Birmingham, U.K.

Swales, John. 1986. Citation analysis and discourse analysis. *Applied Linguistics* 7(1): 39–56.

Swales, John. 1990. *Genre Analysis: English in Academic and Research Settings. Chapter 7: Research articles in English*, 110–176. Cambridge, UK: Cambridge University Press.

Taddio, A., T. Pain, F. F. Fassos, H. Boon, A. L. Ilersich, and Elnarson T. R. 1994. Quality of nonstructured and structured abstracts of original research articles in the *British Medical Journal*, the *Canadian Medical Association Journal* and the *Journal of the American Medical Association*. *Canadian Medical Association Journal* 150(10): 1611–1615.

Taylor, Paul, Richard Caley, Alan W. Black, and Simon King. 1999. The Edinburgh Speech Tools Library (Centre for Speech Technology Research). `http://www.cstr.ed.ac.uk/projects/speech_tools/`.

Teufel, Simone. 1998. Meta-discourse markers and problem-structuring in scientific articles. In *Proceedings of the ACL-98 Workshop on Discourse Structure and Discourse Markers*, 43–49.

Teufel, Simone, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the 8th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99)*, 110–117.

Teufel, Simone, and Marc Moens. 1997. Sentence extraction as a classification task. In Mani and Maybury 1997, 58–65.

Teufel, Simone, and Marc Moens. 1998. Sentence extraction and rhetorical classification for flexible abstracts. In Radev and Hovy 1998, 16–25.

Teufel, Simone, and Marc Moens. 1999a. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In Mani and Maybury 1999, 155–171.

Teufel, Simone, and Marc Moens. 1999b. Discourse-level argumentation in scientific articles: human and automatic annotation. In *Proceedings of ACL-99 Workshop "Towards Standards and Tools for Discourse Tagging"*, 84–93.

Teufel, Simone, and Marc Moens. In Prep. Argumentative Zoning of Scientific Text.

Thomas, Sarah, and Thomas Hawes. 1994. Reporting verbs in medical journal articles. *English for Specific Purposes* 13(4): 129–148.

Thompson, Geoff, and Ye Yiyun. 1991. Evaluation in the reporting verbs used in academic papers. *Applied Linguistics* 12(4): 365–382.

Tibbo, Helen R. 1992. Abstracting across the disciplines: A content analysis of abstracts from the natural sciences, and the humanities with implications for abstracting standards and online information retrieval. *Library and Information Science Research* 14(1): 31–56.

Tipster SUMMAC. 1999. `http://www.itl.nist.gov/div894/894.02/related_projects/tipster_summac/i%ndex/cmp_lg.html`.

Toulmin, Stephen, ed. 1972. *Human Understanding: The Collective Use and Evolution of Concepts*. Princeton, NJ: Princeton University Press.

Trawinski, Bogdan. 1989. A methodology for writing problem-structured abstracts. *Information Processing and Management* 25(6): 693–702.

van Dijk, Teun A. 1980. *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction and Cognition*. Hillsdale, NJ: Lawrence Erlbaum.

van Eemeren, F. H., R. Grootendorst, and F. Snoeck-Henkemans. 1996. *Fundamentals of Argumentation Theory*. Lawrence Erlbaum.

van Emden, Joan, and Jennifer Easteal. 1996. *Technical Writing and Speaking: An Introduction*. London, UK: McGraw-Hill.

van Rijsbergen, Cornelis Joost. 1979. *Information Retrieval*. London, UK: Butterworth, 2nd edn.

Weil, B. H., H. Owen, and I. Zarember. 1963. Technical abstracting fundamentals. II. Writing principles and practices. *Journal of Chemical Documentation* 3(2): 125–132.

Weinstock, Melvin. 1971. Citation indexes. In *Encyclopedia of Library and Information Science*, vol. 5, 16–40. New York, NY: Dekker.

Wellons, M. E., and G. P. Purcell. 1999. Task-specific extracts for using the medical literature. In *Proceedings of the American Medical Informatics Symposium*, 1004–1008.

West, Gregory K. 1980. That-nominal constructions in traditional rhetorical divisions of scientific research papers. *TESOL Quarterly* 14(4): 483–488.

Wiebe, Janyce. 1994. Tracking point of view in narrative. *Computational Linguistics* 20(2): 223–287.

Wiebe, Janyce, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, 246–253.

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, 189–196.

Yianilos, Peter. 1997. The LikeIt Intelligent String Comparison Facility. TR-97-093. Technical report, NEC Research Institute, Princeton, NJ. Also available from `http://www.neci.nj.nec.com/homepages/pny/papers/likeit/main.html`.

Zappen, James P. 1983. A rhetoric for research in sciences and technologies. In Paul V. Anderson, R. John Brockman, and Carolyn R. Miller, eds., *New Essays in Technical and Scientific Communication Research Theory Practice*, 123–138. Farmingdale, NY: Baywood Publishing Company, Inc.

Zechner, Klaus. 1995. Automatic Text Abstracting by Selecting Relevant Passages. Master's thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh, UK.

Ziman, John M. 1968. *Public Knowledge: An Essay Concerning the Social Dimensions of Science*. Cambridge, UK: Cambridge University Press.

Ziman, John M. 1969. Information, Communication, Knowledge. *Nature* 224: 318–324.

Zuckerman, Harriet, and Robert K. Merton. 1973. Institutionalized patterns of evaluation in science. In Robert K. Merton, ed., *The Sociology of Science: Theoretical and Empirical Investigations*, 460–496. Chicago, IL: University of Chicago Press.

Appendix A

The Corpus

A.1. Format of Article Encoding

```
<!ELEMENT PAPER      (TITLE,REFLABEL,AUTHORS,FILENO,APPEARED,ANNOTATOR?,DATE?,ABSTRACT,
                      BODY,REFERENCES?)>
<!ELEMENT TITLE      (#PCDATA)>
<!ELEMENT AUTHORS    (AUTHOR+)>
<!ELEMENT AUTHOR     (#PCDATA)>
<!ELEMENT FILENO     (#PCDATA)>
<!ELEMENT ANNOTATOR  (#PCDATA)>
<!ELEMENT DATE       (#PCDATA)>
<!ELEMENT YEAR       (#PCDATA)>
<!ELEMENT APPEARED   (#PCDATA)>
<!ELEMENT EQN         EMPTY>
<!ATTLIST EQN
          C           CDATA       'NP'>
<!ELEMENT CREF        EMPTY>
<!ATTLIST CREF
          C           CDATA       'NP'>
<!ELEMENT REFERENCES (P|REFERENCE)*>
<!ELEMENT REFERENCE  (#PCDATA|REFLABEL|W|EQN|NAME|SURNAME|DATE|ETAL|REFAUTHOR|YEAR)*>
<!ELEMENT NAME       (#PCDATA|SURNAME|INVERTED)* >
<!ELEMENT SURNAME    (#PCDATA)>
<!ELEMENT REF        (#PCDATA)*>
<!ATTLIST REF
          SELF       (YES|NO)  "NO"
          C           CDATA       'NNP'>
<!ELEMENT REFAUTHOR  (#PCDATA|SURNAME)*>
<!ATTLIST REFAUTHOR
          C           CDATA       'NNP'>
<!ELEMENT ETAL       (#PCDATA)>
<!ELEMENT BODY       (DIV)+>
<!ELEMENT DIV        (HEADER?, (DIV|P|IMAGE|EXAMPLE)*)>
<!ATTLIST DIV
          DEPTH       CDATA  #REQUIRED >
<!ELEMENT HEADER     (#PCDATA|EQN|REF|REFAUTHOR|CREF|W)*>
<!ATTLIST HEADER      ID  ID  #REQUIRED >
<!ELEMENT P          (S|IMAGE|EXAMPLE)*>
<!ATTLIST P
          TYPE       (ITEM|TXT) "TXT">
<!ELEMENT IMAGE       EMPTY>
<!ATTLIST IMAGE
          ID          ID #REQUIRED
          CATEGORY   (AIM|CONTRAST|TEXTUAL|OWN|BACKGROUND|BASIS|OTHER)  #IMPLIED>
```

```
<!ELEMENT S           (#PCDATA|EQN|REF|REFAUTHOR|CREF|FORMULAIC|AGENT|FINITE|W)*>
<!ATTLIST S
         TYPE         (ITEM|TXT) "TXT"
         ID            ID     #REQUIRED
         ABSTRACTC     CDATA  #IMPLIED
         CATEGORY     (AIM|CONTRAST|TEXTUAL|OWN|BACKGROUND|BASIS|OTHER)  #IMPLIED>
<!ELEMENT ABSTRACT    (A-S)*>
<!ELEMENT A-S         (#PCDATA|EQN|REF|REFAUTHOR|CREF|FORMULAIC|AGENT|FINITE|W)*>
<!ATTLIST A-S
         ID            ID          #REQUIRED
         TYPE         (ITEM|TXT)   "TXT"
         DOCUMENTC     CDATA       #IMPLIED
         CATEGORY     (AIM|CONTRAST|TEXTUAL|OWN|BACKGROUND|BASIS|OTHER)  #IMPLIED>
<!ELEMENT EXAMPLE     (EX-S)+>
<!ATTLIST EXAMPLE
         ID            ID #REQUIRED
         CATEGORY     (AIM|CONTRAST|TEXTUAL|OWN|BACKGROUND|BASIS|OTHER)  #IMPLIED>
<!ELEMENT EX-S        (#PCDATA|EQN|W)*>
<!ELEMENT W           (#PCDATA)>
<!ATTLIST W
         C            CDATA  #IMPLIED>
<!ELEMENT FINITE_VERB (#PCDATA)>
<!ATTLIST FINITE_VERB
ACTION
(AFFECT_ACTION|ARGUMENTATION_ACTION|AWARE_ACTION|BETTER_SOLUTION_ACTION|CHANGE_ACTION|
COMPARISON_ACTION|CONTINUE_ACTION|CONTRAST_ACTION|FUTURE_INTEREST_ACTION|INTEREST_ACTION|
NEED_ACTION|PRESENTATION_ACTION|PROBLEM_ACTION|RESEARCH_ACTION|SIMILAR_ACTION|
SOLUTION_ACTION|TEXTSTRUCTURE_ACTION|USE_ACTION|POSSESSION|COPULA|0)
"0">

<!ELEMENT FORMULAIC (#PCDATA|EQN|CREF|REF|REFAUTHOR)*>
<!ATTLIST FORMULAIC TYPE
(US_AGENT|REF_US_AGENT|REF_AGENT|OUR_AIM_AGENT|US_PREVIOUS_AGENT|THEM_PRONOUN_AGENT|THEM_AGENT|
GENERAL_AGENT|PROBLEM_AGENT|SOLUTION_AGENT|THEM_FORMULAIC|US_PREVIOUS_FORMULAIC|
TEXTSTRUCTURE_AGENT|NO_TEXTSTRUCTURE_FORMULAIC|IN_ORDER_TO_FORMULAIC|AIM_FORMULAIC|
TEXTSTRUCTURE_FORMULAIC|METHOD_FORMULAIC|HERE_FORMULAIC|CONTINUE_FORMULAIC|SIMILARITY_FORMULAIC|
COMPARISON_FORMULAIC|CONTRAST_FORMULAIC|GAP_FORMULAIC|FUTURE_FORMULAIC|AFFECT_FORMULAIC|
GOOD_FORMULAIC|BAD_FORMULAIC|0)
"0">

<!ELEMENT AGENT (#PCDATA|EQN|REF|CREF|REFAUTHOR)*>
<!ATTLIST AGENT
  TYPE
  (US_AGENT|THEM_AGENT|THEM_PRONOUN_AGENT|US_PREVIOUS_AGENT|REF_US_AGENT|REF_AGENT|
GENERAL_AGENT|PROBLEM_AGENT|SOLUTION_AGENT|0) "0">
```

A.2. List of Scientific Articles

No.	CMP-LG	Conference	Title	Authors	Words	Sent.	Abstr. sent.
0	9405001	ACL94	Similarity-Based Estimation of Word Cooccurrence Probabilities	I.Dagan, F.Pereira, L.Lee	4343	160	7
1	9405002	ACL94 Student	Temporal Relations: Reference or Discourse Coherence?	A.Kehler	2320	79	5
2	9405004	COLING94	Syntactic-Head-Driven Generation	E.Koenig	3438	116	4
3	9405010	ACL94	Common Topics and Coherent Situations: Interpreting Ellipsis in the Context of Discourse Inference	A.Kehler	5326	156	5
4	9405013	COLING94	Collaboration on Reference to Objects that are not Mutually Known	P.Edmonds	3994	135	5
5	9405022	ACL94	Grammar Specialization through Entropy Thresholds	C.Samuelsson	4639	170	4
6	9405023	ACL94 Student	An Integrated Heuristic Scheme for Partial Parse Evaluation	A.Lavie	2454	102	5
7	9405028	COLING94	Semantics of Complex Sentences in Japanese	H.Nakagawa S.Nishizawa	4700	200	5
8	9405033	ACL94	Relating Complexity to Practical Performance in Parsing with Wide-Coverage Unification Grammars	J.Carroll	5353	121	2
9	9405035	ACL94 Student	Dual-Coding Theory and Connectionist Lexical Selection	Y.Wang	1889	90	2
10	9407011	ACL94	Discourse Obligations in Dialogue Processing	D.Traum, J.Allen	6498	233	2
11	9408003	COLING94 Reserve	Typed Feature Structures as Descriptions	P.King	2490	167	2
12	9408004	ACL94 Workshop	Parsing with Principles and Probabilities	A.Fordham, M.Crocker	3645	97	3
13	9408006	COLING94	LHIP: Extended DCGs for Configurable Robust Parsing	A.Ballim, G.Russell	4468	184	2
14	9408011	ACL93	Distributional Clustering of English Words	F.Pereira, N.Tishby, L.Lee	4778	170	4
15	9408014	ACL94 Workshop	Qualitative and Quantitative Models of Speech Translation	H.Alshawi	7635	296	4
16	9409004	COLING94	An Experiment on Learning Appropriate Selectional Restrictions from a Parsed Corpus	F.Ribas	4060	179	3
17	9410001	ANLP94	Improving Language Models by Clustering Training Sentences	D.Carter	5372	150	6
18	9410005	ACL87	A Centering Approach to Pronouns	S.Brennan, M.Friedman, C.Pollard	2494	98	4
19	9410006	ACL89	Evaluating Discourse Processing Algorithms	M.Walker	7281	258	8
20	9410008	COLING94	Recognizing Text Genres with Simple Metrics Using Discriminant Analysis	J.Karlgren, D.Cutting	1952	66	3
21	9410009	COLING94	Reserve Lexical Functions and Machine Translation	D.Heylen, K.Maxwell, M.Verhagen	3766	135	2
22	9410012	ANLP94	Does Baum-Welch Re-estimation Help Taggers?	D.Elworthy	4167	1411	0
23	9410022	ACL94 SIG	Automated Tone Transcription	S.Bird	7139	322	8
24	9410032	COLING94	Planning Argumentative Texts	X.Huang	3824	183	4
25	9410033	COLING94	Default Handling in Incremental Generation	K.Harbusch, G.Kikui, A.Kilger	4224	176	5
26	9411019	COLING94	Focus on "only" and "not"	A.Ramsay	2815	99	2
27	9411021	COLING94	Free-ordered CUG on Chemical Abstract Machine	S.Tojo	2060	86	5
28	9411023	COLING94	Abstract Generation Based on Rhetorical Structure Extraction	K.Ono, K.Sumita, S.Miike	2824	112	4

No.	CMP-LG	Conference	Title	Authors	Words	Sent.	Abstr. sent.
29	9412005	ACL94 SIG	Segmenting Speech without a Lexicon: the Roles of Phonotactics and Speech Source	T.Cartwright, M.Brent	5481	166	6
30	9412008	COLING94	Analysis of Japanese Compound Nouns using Collocational Information	Y.Kobayasi, T.Tokunaga, H.Tanaka	3459	172	4
31	9502004	COLING94	Bottom-Up Earley Deduction	G.Erbach	3591	126	3
32	9502005	EACL95	Off-line Optimization for Earley-style HPSG Processing	G.Minnen, D.Gerdemann, T.Goetz	4134	129	3
33	9502006	EACL95	Rapid Development of Morphological Descriptions for Full Language Processing Systems	D.Carter	5292	162	4
34	9502009	EACL95	On Learning More Appropriate Selectional Restrictions	F.Ribas	3759	166	4
35	9502014	EACL95	Ellipsis and Quantification: A Substitutional Approach	R.Crouch	5324	230	2
36	9502015	EACL95	The Semantics of Resource Sharing in Lexical-Functional Grammar	A.Kehler, M.Dalrymple, J.Lamping, V.Saraswat	4259	155	3
37	9502018	EACL95	Algorithms for Analysing the Temporal Structure of Discourse	J.Hitzeman, M.Moens, C.Grover	3980	137	4
38	9502021	EACL95	A Tractable Extension of Linear Indexed Grammars	B.Keller, D.Weir	3963	140	3
39	9502022	EACL95	Stochastic HPSG	C.Brew	3390	129	3
40	9502023	EACL95	Splitting the Reference Time: Temporal Anaphora and Quantification in DRT	R.Nelken, N.Francez	4283	149	5
41	9502024	EACL95	A Robust Parser Based on Syntactic Information	K.Lee, C.Kweon, J.Seo, G.Kim	3308	159	7
42	9502031	EACL95 Student	Cooperative Error Handling and Shallow Processing	T.Bowden	2443	88	6
43	9502033	EACL95 Student	An Algorithm to Co-Ordinate Anaphora Resolution and PPS Disambiguation Process	S.Azzam	1301	45	3
44	9502035	EACL95 Student	Incorporating " Unconscious Reanalysis " into an Incremental, Monotonic Parser	P.Sturt	4352	126	4
45	9502037	EACL95 Student	A State-Transition Grammar for Data-Oriented Parsing	D.Tugwell	3305	116	2
46	9502038	EACL95 Workshop	Implementation and evaluation of a German HMM for POS disambiguation	H.Feldweg	3625	129	5
47	9502039	EACL95 Workshop	Multilingual Sentence Categorization according to Language	E.Giguet	2142	93	13
48	9503002	EACL95	Computational Dialectology in Irish Gaelic	B.Kessler	4576	165	5
49	9503004	EACL95 Workshop	Creating a Tagset, Lexicon and Guesser for a French tagger	J.Chanod, P.Tapanainen	4690	170	3
50	9503005	EACL95	A Specification Language for Lexical Functional Grammars	P.Blackburn, C.Gardent	4968	218	4
51	9503007	EACL95	The Semantics of Motion	P.Sablayrolles	2361	85	3
52	9503009	EACL95	Distributional Part-of-Speech Tagging	H.Schuetze	5014	184	3
53	9503013	COLING95	Incremental Interpretation: Applications, Theory, and Relationship to Dynamic Semantics	D.Milward, R.Cooper	5676	186	6
54	9503014	COLING94	Non-Constituent Coordination: Theory and Practice	D.Milward	5278	192	3
55	9503015	EACL95	Incremental Interpretation of Categorial Grammar	D.Milward	4903	165	4

No.	CMP-LG	Conference	Title	Authors	Words	Sent.	Abstr. sent.
56	9503017	COLING92	Redundancy in Collaborative Dialogue	M.Walker	5255	212	9
57	9503018	COLING94	Discourse and Deliberation: Testing a Collaborative Strategy	M.Walker	5331	182	4
58	9503023	EACL95	A Fast Partial Parse of Natural Language Sentences Using a Connectionist Method	C.Lyon, B.Dickerson	5027	230	4
59	9503025	COLING94	Occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries	Y.Niwa, Y.Nitta	2749	110	3
60	9504002	EACL95 Workshop	Tagset Design and Inflected Languages	D.Elworthy	3467	130	3
61	9504006	ACL88	Cues and Control in Expert-Client Dialogues	S.Whittaker, P.Stenton	3925	152	4
62	9504007	ACL90	Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation	M.Walker, S.Whittaker	5019	190	9
63	9504017	ACL95	A Uniform Treatment of Pragmatic Inferences in Simple and Complex Utterances and Sequences of Utterances	D.Marcu, G.Hirst	3911	132	4
64	9504024	ACL95	A Morphographemic Model for Error Correction in Nonconcatenative Strings	T.Bowden, G.Kiraz	3171	143	4
65	9504026	ACL95	The Intersection of Finite State Automata and Definite Clause Grammars	G.vanNoord	3614	151	8
66	9504027	ACL95	An Efficient Generation Algorithm for Lexicalist MT	V.Poznanski, J.Beaven, P.Whitelock	4236	175	3
67	9504030	ACL95	Statistical Decision-Tree Models for Parsing	D.Magerman	4555	188	8
68	9504033	ACL95	Corpus Statistics Meet the Noun Compound: Some Empirical Results	M.Lauer	4384	191	4
79	9504034	ACL95	Bayesian Grammar Induction for Language Modeling	S.Chen	4581	175	5
70	9505001	ACL95	Response Generation in Collaborative Negotiation	J.Chu-Carroll, S.Carberry	5962	154	5
71	9506004	ACL95	Using Higher-Order Logic Programming for Semantic Interpretation of Coordinate Constructs	S.Kulick	3362	130	4
72	9511001	COLING94	Countability and Number in Japanese-to-English Machine Translation	F.Bond, K.Ogura, S.Ikehara	3439	136	2
73	9511006	ACL95 Workshop	Disambiguating Noun Groupings with Respect to WordNet Senses	P.Resnik	5970	159	5
74	9601004	EACL93	Similarity between Words Computed by Spreading Activation on an English Dictionary	H.Kozima, T.Furugori	4384	212	4
75	9604019	ACL96	Magic for Filter Optimization in Dynamic Bottom-up Processing	G.Minnen	3964	157	3
76	9604022	ACL96	Unsupervised Learning of Word-Category Guessing Rules	A.Mikheev	6138	236	4
77	9605013	COLING96	Learning Dependencies between Case Frame Slots	H.Li, N.Abe	4858	170	8
78	9605014	COLING96	Clustering Words with the MDL Principle	H.Li, N.Abe	4467	167	5
79	9605016	ACL96	Parsing for Semidirectional Lambek Grammar is NP-Complete	J.Doerre	3060	126	4

Appendix B

Example Paper cmp_lg-9408011

B.1. XML Format

```
<?xml version='1.0'?>
<!DOCTYPE STRUCT-PAPER SYSTEM "/projects/ltg/users/simone/src/dtd/structure.dtd" [
<!ENTITY S "9408011.p">
]>
<STRUCT-PAPER>
<TITLE> Distributional Clustering of English Words </TITLE>
<AUTHORS>
<AUTHOR>Fernando Pereira</AUTHOR>
<AUTHOR>Naftali Tishby</AUTHOR>
<AUTHOR>Lillian Lee</AUTHOR>
</AUTHORS>
<FILENO>9408011</FILENO>
<APPEARED>ACL93</APPEARED>
<ABSTRACT>
<A-S ID='A-0' DOCUMENTC=S-0;S-164> We describe and experimentally evaluate a method for automatically clustering words according
to their distribution in particular syntactic contexts . </A-S>
<A-S ID='A-1'> Deterministic annealing is used to find lowest distortion sets of clusters . </A-S>
<A-S ID='A-2'> As the annealing parameter increases , existing clusters become unstable and subdivide , yielding a hierarchical ''
soft '' clustering of the data . </A-S>
<A-S ID='A-3'> Clusters are used as the basis for class models of word coocurrence , and the models evaluated with respect to
held-out test data . </A-S>
</ABSTRACT>
<BODY>
<DIV DEPTH='1'>
<HEADER ID='H-0'> Introduction </HEADER>
<P>
<S ID='S-0' ABSTRACTC=A-0> Methods for automatically classifying words according to their contexts of use have both scientific and
practical interest . </S>
<S ID='S-1'> The scientific questions arise in connection to distributional views of linguistic ( particularly lexical ) structure
and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives . </S>
<S ID='S-2'> From the practical point of view , word classification addresses questions of data sparseness and generalization in
statistical language models , particularly models for deciding among alternative analyses proposed by a grammar . </S>
</P>
<P>
<S ID='S-3'> It is well known that a simple tabulation of frequencies of certain words participating in certain configurations ,
for example of frequencies of pairs of a transitive main verb and the head noun of its direct object , cannot be reliably used for
comparing the likelihoods of different alternative configurations . </S>
<S ID='S-4'> The problem is that for large enough corpora the number of possible joint events is much larger than the number of
event occurrences in the corpus , so many events are seen rarely or never , making their frequency counts unreliable estimates of
their probabilities . </S>
</P>
<P>
<S ID='S-5'> <REF>Hindle 1990</REF> proposed dealing with the sparseness problem by estimating the likelihood of unseen events
from that of '' similar '' events that have been seen . </S>
<S ID='S-6'> For instance , one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that
direct object for similar verbs . </S>
<S ID='S-7'> This requires a reasonable definition of verb similarity and a similarity estimation method . </S>
<S ID='S-8'> In <REFAUTHOR>Hindle</REFAUTHOR> 's proposal , words are similar if we have strong statistical evidence that they
tend to participate in the same events . </S>
<S ID='S-9'> His notion of similarity seems to agree with our intuitions in many cases , but it is not clear how it can be used
directly to construct word classes and corresponding models of association . </S>
</P>
```

```
<P>
<S ID='S-10'> Our research addresses some of the same questions and uses similar raw data , but we investigate how to factor word
association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves .
</S>
<S ID='S-11'> While it may be worthwhile to base such a model on preexisting sense classes <REF>Resnik 1992</REF> , in the work
described here we look at how to derive the classes directly from distributional data . </S>
<S ID='S-12'> More specifically , we model senses as probabilistic concepts or clusters c with corresponding cluster membership
probabilities <EQN/> for each word w . </S>
<S ID='S-13'> Most other class-based modeling techniques for natural language rely instead on '' hard '' Boolean classes
<REF>Brown et al. 1990</REF> . </S>
<S ID='S-14'> Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving
particular words , a potentially unreliable source of information as we noted above . </S>
<S ID='S-15'> Our approach avoids both problems . </S>
</P>
<DIV DEPTH='2'>
<HEADER ID='H-1'> Problem Setting </HEADER>
<P>
<S ID='S-16'> In what follows , we will consider two major word classes , <EQN/> and <EQN/> , for the verbs and nouns in our
experiments , and a single relation between them , in our experiments relation between a transitive main verb and the head noun of
its direct object . </S>
<S ID='S-17'> Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs (v,n) in
the required configuration in a training corpus . </S>
<S ID='S-18'> Some form of text analysis is required to collect such a collection of pairs . </S>
<S ID='S-19'> The corpus used in our first experiment was derived from newswire text automatically parsed by
<REFAUTHOR>Hindle</REFAUTHOR> 's parser Fidditch <REF>Hindle 1993</REF> . </S>
<S ID='S-20'> More recently , we have constructed similar tables with the help of a statistical part-of-speech tagger <REF>Church
1988</REF> and of tools for regular expression pattern matching on tagged corpora <REF>Yarowsky 1992</REF> . </S>
<S ID='S-21'> We have not yet compared the accuracy and coverage of the two methods , or what systematic biases they might
introduce , although we took care to filter out certain systematic errors , for instance the misparsing of the subject of a
complement clause as the direct object of a main verb for report verbs like '' say '' . </S>
</P>
<P>
<S ID='S-22'> We will consider here only the problem of classifying nouns according to their distribution as direct objects of
verbs ; the converse problem is formally similar . </S>
<S ID='S-23'> More generally , the theoretical basis for our method supports the use of clustering to build models for any n-ary
relation in terms of associations between elements in each coordinate and appropriate hidden units ( cluster centroids ) and
associations between those hidden units . </S>
</P>
<P>
<S ID='S-24'> For the noun classification problem , the empirical distribution of a noun n is then given by the conditional
density <EQN/> . </S>
<S ID='S-25'> The problem we study is how to use the <EQN/> to classify the <EQN/> . </S>
<S ID='S-26'> Our classification method will construct a set <EQN/> of clusters and cluster membership probabilities <EQN/> .
</S>
<S ID='S-27'> Each cluster c is associated to a cluster centroid <EQN/> , which is discrete density over <EQN/> obtained by
averaging appropriately the <EQN/> . </S>
</P>
</DIV>
<DIV DEPTH='2'>
<HEADER ID='H-2'> Distributional Similarity </HEADER>
<P>
<S ID='S-28'> To cluster nouns n according to their conditional verb distributions <EQN/> , we need a measure of similarity
between distributions . </S>
<S ID='S-29'> We use for this purpose the relative entropy or Kullback-Leibler ( KL ) distance between two distributions . </S>
</P>
<IMAGE ID='I-0'/>
<P>
<S ID='S-30'> This is a natural choice for a variety of reasons , which we will just sketch here . </S>
</P>
<P>
<S ID='S-31'> First of all , <EQN/> is zero just in case p = q , and it increases as the probability decreases that p is the
relative frequency distribution of a random sample drawn according to p . </S>
<S ID='S-32'> More formally , the probability mass given by q to the set of all samples of length n with relative frequency
distribution p is bounded by <EQN/> <REF>Cover and Thomas 1991</REF> . </S>
<S ID='S-33'> Therefore , if we are trying to distinguish among hypotheses <EQN/> when p is the relative frequency distribution
of observations , <EQN/> gives the relative weight of evidence in favor of <EQN/> . </S>
<S ID='S-34'> Furthermore , a similar relation holds between <EQN/> for two empirical distributions p and p ' and the probability
that p and p ' are drawn from the same distribution q . </S>
<S ID='S-35'> We can thus use the relative entropy between the context distributions for two words to measure how likely they are
to be instances of the same cluster centroid . </S>
</P>
<P>
<S ID='S-36'> From an information theoretic perspective <EQN/> measures how inefficient on average it would be to use a code
based on q to encode a variable distributed according to p . </S>
<S ID='S-37'> With respect to our problem , <EQN/> thus gives us the loss of information in using cluster centroid <EQN/>
instead of the actual distribution for word <EQN/> when modeling the distributional properties of n . </S>
</P>
<P>
<S ID='S-38'> Finally , relative entropy is a natural measure of similarity between distributions for clustering because its
minimization leads to cluster centroids that are a simple weighted average of member distributions . </S>
</P>
<P>
<S ID='S-39'> One technical difficulty is that <EQN/> is not defined when p'(x) = 0 but <EQN/> . </S>
<S ID='S-40'> We could sidestep this problem ( as we did initially ) by smoothing zero frequencies appropriately <REF>Church and
Gale 1991</REF> . </S>
<S ID='S-41'> However , this is not very satisfactory because one of the goals of our work is precisely to avoid the problems of
data sparseness by grouping words into classes . </S>
<S ID='S-42'> It turns out that the problem is avoided by our clustering technique , since it does not need to compute the KL
distance between individual word distributions , but only between a word distribution and average distributions , the current
cluster centroids , which are guaranteed to be nonzero whenever the word distributions are . </S>
```

```
<S ID='S-43'> This is a useful advantage of our method compared with agglomerative clustering techniques that need to compare
individual objects being considered for grouping . </S>
</P>
</DIV>
</DIV>
<DIV DEPTH='1'>
<HEADER ID='H-3'> Theoretical Basis </HEADER>
<P>
<S ID='S-44'> In general , we are interested on how to organize a set of linguistic objects such as words according to the
contexts in which they occur , for instance grammatical constructions or n-grams . </S>
<S ID='S-45'> We will show elsewhere that the theoretical analysis outlined here applies to that more general problem , but for
now we will only address the more specific problem in which the objects are nouns and the contexts are verbs that take the nouns as
direct objects . </S>
</P>
<P>
<S ID='S-46'> Our problem can be seen as that of learning a joint distribution of pairs from a large sample of pairs . </S>
<S ID='S-47'> The pair coordinates come from two large sets <EQN/> and <EQN/> , with no preexisting topological or metric
structure , and the training data is a sequence S of N independently drawn pairs . </S>
</P>
<IMAGE ID='I-1'/>
<P>
<S ID='S-48'> From a learning perspective , this problem falls somewhere in between unsupervised and supervised learning . </S>
<S ID='S-49'> As in unsupervised learning , the goal is to learn the underlying distribution of the data . </S>
<S ID='S-50'> But in contrast to most unsupervised learning settings , the objects involved have no internal structure or
attributes allowing them to be compared with each other . </S>
<S ID='S-51'> Instead , the only information about the objects is the statistics of their joint appearance . </S>
<S ID='S-52'> These statistics can thus be seem as a weak form of object labelling analogous to supervision . </S>
</P>
<DIV DEPTH='2'>
<HEADER ID='H-4'> Distributional Clustering </HEADER>
<P>
<S ID='S-53'> While clusters based on distributional similarity are interesting on their own , they can also be profitably seen as
a means of summarizing a joint distribution . </S>
<S ID='S-54'> In particular , we would like to find a set of clusters <EQN/> such that each conditional distribution <EQN/> can
be approximately decomposed as </S>
</P>
<IMAGE ID='I-2'/>
<P>
<S ID='S-55'> where <EQN/> is the membership probability of n in c and <EQN/> is v 's conditional probability given by the
centroid distribution for cluster c . </S>
</P>
<P>
<S ID='S-56'> The above decomposition can be written in a more symmetric form as </S>
</P>
<IMAGE ID='I-3'/>
<P>
<S ID='S-57'> assuming that <EQN/> and <EQN/> coincide . </S>
<S ID='S-58'> We will take <CREF/> as our basic clustering model . </S>
</P>
<P>
<S ID='S-59'> To determine this decomposition we need to solve the two connected problems of finding find suitable forms for
the cluster membership and centroid distributions <EQN/> , and of maximizing the goodness of fit between the model distribution
<EQN/> and the observed data . </S>
</P>
<P>
<S ID='S-60'> Goodness of fit is determined by the model 's likelihood of the observations . </S>
<S ID='S-61'> The maximum likelihood ( ML ) estimation principle is thus the natural tool to determine the centroid distributions
<EQN/> . </S>
</P>
<P>
<S ID='S-62'> As for the membership probabilities , they must be determined solely by the relevant measure of object-to-cluster
similarity , which in the present work is the relative entropy between object and cluster centroid distributions . </S>
<S ID='S-63'> Since no other information is available , the membership is determined by maximizing the configuration entropy
subject for a fixed average distortion . </S>
<S ID='S-64'> With the maximum entropy ( ME ) membership distribution , ML estimation is equivalent to the minimization of the
average distortion of the data . </S>
<S ID='S-65'> The combined entropy maximization entropy and distortion minimization is carried out by a two-stage iterative
process similar to the EM method <REF>Dempster et al. 1977</REF> . </S>
<S ID='S-66'> The first stage of an iteration is a maximum likelihood , or minimum distortion , estimation of the cluster
centroids given fixed membership probabilities . </S>
<S ID='S-67'> In the second iteration stage , the entropy of the membership distribution is maximized with a fixed average
distortion . </S>
<S ID='S-68'> This joint optimization searches for a saddle point in the distortion-entropy parameters , which is equivalent to
minimizing a linear combination of the two known as free energy in statistical mechanics . </S>
<S ID='S-69'> This analogy with statistical mechanics is not coincidental , and provide us with a better understanding of the
clustering procedure . </S>
</P>
<DIV DEPTH='3'>
<HEADER ID='H-5'> Maximum Likelihood Cluster Centroids </HEADER>
<P>
<S ID='S-70'> For the maximum likelihood argument , we start by estimating the likelihood of the sequence S of N independent
observations of pairs <EQN/> . </S>
<S ID='S-71'> Using <CREF/> , the sequence 's model log likelihood is </S>
</P>
<IMAGE ID='I-4'/>
<P>
<S ID='S-72'> Fixing the number of clusters ( model size ) <EQN/> , we want to maximize <EQN/> with respect to the distributions
<EQN/> and <EQN/> . </S>
```

```
<S ID='S-73'> The variation of <EQN/> with respect to these distributions is </S>
</P>
<IMAGE ID='I-5'/>
<P>
<S ID='S-74'> with <EQN/> and <EQN/> kept normalized . </S>
<S ID='S-75'> Using Bayes 's formula , we have </S>
</P>
<IMAGE ID='I-6'/>
<P>
<S ID='S-76'> or </S>
</P>
<IMAGE ID='I-7'/>
<P>
<S ID='S-77'> for any c , which we substitute into <CREF/> to obtain </S>
</P>
<IMAGE ID='I-8'/>
<P>
<S ID='S-78'> since <EQN/> . </S>
<S ID='S-79'> This expression is particularly useful when the cluster distributions <EQN/> and <EQN/> are of exponential form ,
precisely what will be provided by the ME step described below . </S>
</P>
<P>
<S ID='S-80'> At this point we need to specify the clustering model in more detail . </S>
<S ID='S-81'> In the derivation so far we have treated <EQN/> and <EQN/> symmetrically , corresponding to clusters not of verbs
or nouns but of verb-noun associations . </S>
<S ID='S-82'> In principle such a symmetric model may be more accurate , but in this paper we will concentrate on asymmetric
models in which cluster memberships are associated to just one of the components of the joint distribution and the cluster centroids
are specified only by the other component . </S>
<S ID='S-83'> In particular , the model we use in our experiments has noun clusters with cluster memberships determined by <EQN/>
and centroid distributions determined by <EQN/> . </S>
</P>
<P>
<S ID='S-84'> The asymmetric model simplifies the estimation significantly by dealing with a single component , but it has
the disadvantage that the joint distribution , <EQN/> has two different and not necessarily consistent expressions in terms of
asymmetric models for the two coordinates . </S>
</P>
</DIV>
<DIV DEPTH='3'>
<HEADER ID='H-6'> Maximum Entropy Cluster Membership </HEADER>
<P>
<S ID='S-85'> While variations of <EQN/> and <EQN/> in equation <CREF/> are not independent , we can treat them separately .
</S>
<S ID='S-86'> First , for fixed average distortion between the cluster centroid distributions <EQN/> and the data <EQN/> ,
we find the cluster membership probabilities , which are the Bayes 's inverses of the <EQN/> , that maximize the entropy of the
cluster distributions . </S>
<S ID='S-87'> With the membership distributions thus obtained , we then look for the <EQN/> that maximize the log likelihood l (
S ) . </S>
<S ID='S-88'> It turns out that this will also be the values of <EQN/> that minimize the average distortion between the
asymmetric cluster model and the data . </S>
</P>
<P>
<S ID='S-89'> Given any similarity measure <EQN/> between nouns and cluster centroids , the average cluster distortion is </S>
</P>
<IMAGE ID='I-9'/>
<P>
<S ID='S-90'> If we maximize the cluster membership entropy </S>
</P>
<IMAGE ID='I-10'/>
<P>
<S ID='S-91'> subject to normalization of <EQN/> and fixed <CREF/> , we obtain the following standard exponential forms for the
class and membership distributions </S>
</P>
<IMAGE ID='I-11'/>
<IMAGE ID='I-12'/>
<P>
<S ID='S-92'> where the normalization sums ( partition functions ) are <EQN/> and <EQN/> . </S>
<S ID='S-93'> Notice that <EQN/> does not need to be symmetric for this derivation , as the two distributions are simply related
by Bayes 's rule . </S>
</P>
<P>
<S ID='S-94'> Returning to the log-likelihood variation <CREF/> , we can now use <CREF/> for <EQN/> and the assumption for the
asymmetric model that the cluster membership stays fixed as we adjust the centroids , to obtain </S>
</P>
<IMAGE ID='I-13'/>
<P>
<S ID='S-95'> where the variation of [EQn] is now included in the variation of <EQN/> . </S>
</P>
<P>
<S ID='S-96'> For a large enough sample , we may replace the sum over observations in <CREF/> by the average over <EQN/> . </S>
</P>
<IMAGE ID='I-14'/>
<P>
<S ID='S-97'> which , applying Bayes 's rule , becomes </S>
</P>
<IMAGE ID='I-15'/>
<P>
<S ID='S-98'> At the log-likelihood maximum , the variation <CREF/> must vanish . </S>
<S ID='S-99'> We will see below that the use of relative entropy for similarity measure makes <EQN/> vanish at the maximum as
well , so the log likelihood can be maximized by minimizing the average distortion with respect to the class centroids while class
membership is kept fixed </S>
</P>
```

```
<IMAGE ID='I-16'/>
<P>
<S ID='S-100'> or , sufficiently , if each of the inner sums vanish </S>
</P>
<IMAGE ID='I-17'/>
</DIV>
<DIV DEPTH='3'>
<HEADER ID='H-7'> Minimizing the Average KL Distortion </HEADER>
<P>
<S ID='S-101'> We first show that the minimization of the relative entropy yields the natural expression for cluster centroids
</S>
</P>
<IMAGE ID='I-18'/>
<P>
<S ID='S-102'> To minimize the average distortion <CREF/> , we observe that the variation of the KL distance between noun and
centroid distributions with respect to the centroid distribution <EQN/> , with each centroid distribution normalized by the
Lagrange multiplier <EQN/> , is given by </S>
</P>
<IMAGE ID='I-19'/>
<P>
<S ID='S-103'> Substituting this expression into <CREF/> , we obtain </S>
</P>
<IMAGE ID='I-20'/>
<P>
<S ID='S-104'> Since the <EQN/> are now independent , we obtain immediately the desired centroid expression <CREF/> , which is
the desired weighted average of noun distributions . </S>
</P>
<P>
<S ID='S-105'> We can now see that the variation <EQN/> vanishes for centroid distributions given by <CREF/> , since it follows
from <CREF/> that </S>
</P>
<IMAGE ID='I-21'/>
</DIV>
<DIV DEPTH='3'>
<HEADER ID='H-8'> The Free Energy Function </HEADER>
<P>
<S ID='S-106'> The combined minimum distortion and maximum entropy optimization is equivalent to the minimization of a single
function , the free energy </S>
</P>
<IMAGE ID='I-22'/>
<P>
<S ID='S-107'> where <EQN/> is the average distortion <CREF/> and H is the cluster membership entropy <CREF/> . </S>
</P>
<P>
<S ID='S-108'> The free energy determines both the distortion and the membership entropy through </S>
</P>
<IMAGE ID='I-23'/>
<P>
<S ID='S-109'> with temperature <EQN/> . </S>
</P>
<P>
<S ID='S-110'> The most important property of the free energy is that its minimum determines the balance between the ''
disordering '' maximum entropy and '' ordering '' distortion minimization in which the system is most likely to be found . </S>
<S ID='S-111'> In fact the probability to find the system at a given configuration is exponential in F </S>
</P>
<IMAGE ID='I-24'/>
<P>
<S ID='S-112'> so a system is most likely to be found in its minimal free energy configuration . </S>
</P>
</DIV>
</DIV>
<DIV DEPTH='2'>
<HEADER ID='H-9'> Hierarchical Clustering </HEADER>
<P>
<S ID='S-113'> The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering <REF>Rose et
al. 1990</REF> , in which the number of clusters is determined through a sequence of phase transitions by continuously increasing
the parameter <EQN/> following an annealing schedule . </S>
</P>
<P>
<S ID='S-114'> The higher <EQN/> , the more local is the influence of each noun on the definition of centroids . </S>
<S ID='S-115'> The dissimilarity plays here the role of distortion . </S>
<S ID='S-116'> When the scale parameter <EQN/> is close to zero , the dissimilarities are almost irrelevant , all words
contribute about equally to each centroid , and so the lowest average distortion solution involves just one cluster which is the
average of all word densities . </S>
<S ID='S-117'> As <EQN/> is slowly increased , a point ( phase transition ) is eventually reached which the natural solution
involves two distinct centroids . </S>
<S ID='S-118'> We say then that the original cluster has split into the two new clusters . </S>
</P>
<P>
<S ID='S-119'> In general , if we take any cluster c and a twin c ' of c such that the centroid <EQN/> is a small random
pertubation of <EQN/> , below the critical <EQN/> at which c splits the membership and centroid reestimation procedure given by
equations <CREF/> and <CREF/> will make <EQN/> and <EQN/> converge , that is , c and c ' are really the same cluster . </S>
<S ID='S-120'> But with <EQN/> above the critical value for c , the two centroids will diverge , giving rise to two daughters of
c . </S>
</P>
<P>
<S ID='S-121'> Our clustering procedure is thus as follows . </S>
<S ID='S-122'> We start with very low <EQN/> and a single cluster whose centroid is the average of all noun distributions . </S>
<S ID='S-123'> For any given <EQN/> , we have a current set of leaf clusters corresponding to the current free energy ( local )
minimum . </S>
```

```
<S ID='S-124'> To refine such a solution , we search for the lowest <EQN/> which is the critical value for some current leaf
cluster splits . </S>
<S ID='S-125'> Ideally , there is just one split at that critical value , but for practical performance and numerical accuracy
reasons we may have several splits at the new critical point . </S>
<S ID='S-126'> The splitting procedure can then be repeated to achieve the desired number of clusters or model cross-entropy .
</S>
</P>
<IMAGE ID='I-25'/>
</DIV>
</DIV>
<DIV DEPTH='1'>
<HEADER ID='H-10'> Clustering Examples </HEADER>
<P>
<S ID='S-127'> All our experiments involve the asymmetric model described in the previous section . </S>
<S ID='S-128'> As explained there , our clustering procedure yields for each value of <EQN/> a set <EQN/> of clusters minimizing
the free energy F , and the asymmetric model for <EQN/> estimates the conditional verb distribution for a noun n by </S>
</P>
<IMAGE ID='I-26'/>
<P>
<S ID='S-129'> where <EQN/> also depends on <EQN/> . </S>
</P>
<P>
<S ID='S-130'> As a first experiment , we used our method to classify the 64 nouns appearing most frequently as heads of direct
objects of the verb '' fire '' in one year ( 1988 ) of Associated Press newswire . </S>
<S ID='S-131'> In this corpus , the chosen nouns appear as direct object heads of a total of 2147 distinct verbs , so each noun is
represented by a density over the 2147 verbs . </S>
</P>
<P>
<S ID='S-132'> Figure <CREF/> shows the five words most similar to the each cluster centroid for the four clusters resulting from
the first two cluster splits . </S>
<S ID='S-133'> It can be seen that first split separates the objects corresponding to the weaponry sense of '' fire '' ( cluster 1
) from the ones corresponding to the personnel action ( cluster 2 ) . </S>
<S ID='S-134'> The second split then further refines the weaponry sense into a projectile sense ( cluster 3 ) and a gun sense (
cluster 4 ) . </S>
<S ID='S-135'> That split is somewhat less sharp , possibly because not enough distinguishing contexts occur in the corpus . </S>
</P>
<IMAGE ID='I-27'/>
<P>
<S ID='S-136'> Figure <CREF/> shows the four closest nouns to the centroid of each of a set of hierarchical clusters derived from
verb-object pairs involving the 1000 most frequent nouns in the June 1991 electronic version of Grolier 's Encyclopedia ( 10 million
words ) . </S>
</P>
</DIV>
<DIV DEPTH='1'>
<HEADER ID='H-11'> Model Evaluation </HEADER>
<P>
<S ID='S-137'> The preceding qualitative discussion provides some indication of what aspects of distributional relationships may
be discovered by clustering . </S>
<S ID='S-138'> However , we also need to evaluate clustering more rigorously as a basis for models of distributional relationships
. </S>
<S ID='S-139'> So , far , we have looked at two kinds of measurements of model quality : </S>
<S ID='S-140' TYPE='ITEM'> relative entropy between held-out data and the asymmetric model , and </S>
<S ID='S-141' TYPE='ITEM'> performance on the task of deciding which of two verbs is more likely to take a given noun as direct
object when the data relating one of the verbs to the noun has been witheld from the training data . </S>
</P>
<P>
<S ID='S-142'> The evaluation described below was performed on the largest data set we have worked with so far , extracted from 44
million words of 1988 Associated Press newswire with the pattern matching techniques mentioned earlier . </S>
<S ID='S-143'> This collection process yielded 1112041 verb-object pairs . </S>
<S ID='S-144'> We selected then the subset involving the 1000 most frequent nouns in the corpus for clustering , and randomly
divided it into a training set of 756721 pairs and a test set of 81240 pairs . </S>
</P>
<DIV DEPTH='2'>
<HEADER ID='H-12'> Relative Entropy </HEADER>
<IMAGE ID='I-28'/>
<P>
<S ID='S-145'> Figure <CREF/> plots the average relative entropy of several data sets to asymmetric clustered models of different
sizes , given by </S>
</P>
<IMAGE ID='I-29'/>
<P>
<S ID='S-146'> where <EQN/> is the relative frequency distribution of verbs taking n as direct object in the test set . </S>
<S ID='S-147'> For each critical value of <EQN/> , we show the relative entropy with respect to the asymmetric model based
on <EQN/> of the training set ( set train ) , of randomly selected held-out test set ( set test ) , and of held-out data for a
further 1000 nouns that were not clustered ( set new ) . </S>
<S ID='S-148'> Unsurprisingly , the training set relative entropy decreases monotonically . </S>
<S ID='S-149'> The test set relative entropy decreases to a minimum at 206 clusters , and then starts increasing , suggesting that
larger models are overtrained . </S>
</P>
<P>
<S ID='S-150'> The new noun test set is intended to test whether clusters based on the 1000 most frequent nouns are useful
classifiers for the selectional properties of nouns in general . </S>
<S ID='S-151'> As the figure shows , the cluster model provides over one bit of information about the selectional properties of
the new nouns , but the overtraining effect is even sharper than for the held-out data involving the 1000 clustered nouns . </S>
</P>
</DIV>
<DIV DEPTH='2'>
<HEADER ID='H-13'> Decision Task </HEADER>
<IMAGE ID='I-30'/>
<P>
<S ID='S-152'> We also evaluated asymmetric cluster models on a verb decision task closer to possible applications to
disambiguation in language analysis . </S>
```

```
<S ID='S-153'> The task consists judging which of two verbs v and v ' is more likely to take a given noun n as object , when all
occurrences of ( v , n ) in the training set were deliberately deleted . </S>
<S ID='S-154'> Thus this test evaluates how well the models reconstruct missing data in the verb distribution for n from the
cluster centroids close to n . </S>
</P>
<P>
<S ID='S-155'> The data for this test was built from the training data for the previous one in the following way , based on a
suggestion by <REF>Dagan et al. 1993</REF> . </S>
<S ID='S-156'> A small number ( 104 ) of ( v , n ) pairs with a fairly frequent verb ( between 500 and 5000 occurrences ) was
randomly picked , and all occurrences of each pair in the training set were deleted . </S>
<S ID='S-157'> The resulting training set was used to build a sequence of cluster models as before . </S>
<S ID='S-158'> Each model was used to decide which of two verbs v and v ' are more likely to appear with a noun n where the ( v ,
n ) data was deleted from the training set , and the decisions compared with the corresponding ones derived from the original event
frequencies in the initial data set . </S>
<S ID='S-159'> More specifically , for each deleted pair ( v , n ) and each verb v ' that occurred with n in the initial data
either at least twice as frequently or at most half as frequently as v , we compared the sign of <EQN/> with that of <EQN/> for
the initial data set . </S>
<S ID='S-160'> The error rate for each model is simply the proportion of sign disagreements in the selected ( v , n , v ' )
triples . </S>
<S ID='S-161'> Figure <CREF/> shows the error rates for each model for all the selected ( v , n , v ' ) ( all ) and for just
those exceptional triples in which the log frequency ratio of ( n , v ) and ( n , v ' ) differs from the log marginal frequency
ratio of v and v ' . </S>
<S ID='S-162'> In other words , the exceptional cases are those in which predictions based just on the marginal frequencies ,
which the initial one-cluster model represents , would be consistently wrong . </S>
</P>
<P>
<S ID='S-163'> Here too we see some overtraining for the largest models considered , although not for the exceptional verbs .
</S>
</P>
</DIV>
</DIV>
<DIV DEPTH='1'>
<HEADER ID='H-14'> Conclusions </HEADER>
<P>
<S ID='S-164' ABSTRACTC=A-0> We have demonstrated that a general divisive clustering procedure for probability distributions can
be used to group words according to their participation in particular grammatical relations with other words . </S>
<S ID='S-165'> The resulting clusters are intuitively informative , and can be used to construct class-based word coocurrence
models with substantial predictive power . </S>
</P>
<P>
<S ID='S-166'> While the clusters derived by the proposed method seem in many cases semantically significant , this intuition
needs to be grounded in a more rigorous assessment . </S>
<S ID='S-167'> In addition to predictive power evaluations of the kind we have already carried out , it might be worth comparing
automatically-derived clusters with human judgements in a suitable experimental setting . </S>
</P>
<P>
<S ID='S-168'> Moving further in the direction of class-based language models , we plan to consider additional distributional
relations ( for instance , adjective-noun ) and apply the results of clustering to the grouping of lexical associations in
lexicalized grammar frameworks such as stochastic lexicalized tree-adjoining grammars <REF>Schabes 1992</REF> . </S>
</P>
</DIV>
<DIV DEPTH='1'>
<HEADER ID='H-15'> Acknowledgments </HEADER>
<P>
<S ID='S-169'> We would like to thank Don Hindle for making available the 1988 Associated Press verb-object data set , the
Fidditch parser and a verb-object structure filter , Mats Rooth for selecting the objects of '' fire '' data set and many
discussions , David Yarowsky for help with his stemming and concordancing tools , and Ido Dagan for suggesting ways of testing
cluster models . </S>
</P>
</DIV>
</BODY>
<REFERENCES>
<REFERENCE>
<REFLABEL>Brown et al 1999</REFLABEL>
Peter F. <SURNAME>Brown</SURNAME>, Vincent J. <SURNAME>Della</SURNAME> <SURNAME>Pietra</SURNAME>, Peter V.
<SURNAME>deSouza</SURNAME>, Jenifer C. <SURNAME>Lai</SURNAME>, and Robert L. <SURNAME>Mercer</SURNAME>. <DATE>1990</DATE>.
Class-based n-gram models of natural language. In Proceedings of the IBM Natural Language ITL, pages 283-298, Paris, France, March.
</REFERENCE>
<REFERENCE>
<REFLABEL>Church and Gale 1991</REFLABEL>
Kenneth W. <SURNAME>Church</SURNAME> and William A. <SURNAME>Gale</SURNAME>. <DATE>1991</DATE>. A comparison of the enhanced
Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. Computer Speech and Language, 5:19-54.
</REFERENCE>
<REFERENCE>
<REFLABEL>Church 1988</REFLABEL>
Kenneth W. <SURNAME>Church</SURNAME>. <DATE>1988</DATE>. A stochastic parts program and noun phrase parser for unrestricted
text. In Proceedings of the Second Conference on Applied Natural Language Processing, pages 136-143, Austin, Texas. Association for
Computational Linguistics, Morristown, New Jersey.
</REFERENCE>
<REFERENCE>
<REFLABEL>Cover 1991</REFLABEL>
Thomas M. <SURNAME>Cover</SURNAME> and Joy A. <SURNAME>Thomas</SURNAME>. <DATE>1991</DATE>. Elements of Information Theory.
Wiley-Interscience, New York, New York.
</REFERENCE>
<REFERENCE>
<REFLABEL>Dagan et al. 1992</REFLABEL>
Ido <SURNAME>Dagan</SURNAME>, Shaul <SURNAME>Markus</SURNAME>, and Shaul <SURNAME>Markovitch</SURNAME>. <DATE>1992</DATE>.
Contextual word similarity and the estimation of sparse lexical relations. Submitted for publication.
</REFERENCE>
```

```
<REFERENCE>
<REFLABEL>Dempster and Rubin 1977</REFLABEL>
A. P. <SURNAME>Dempster</SURNAME>, N. M. <SURNAME>Laird</SURNAME>, and D. B. <SURNAME>Rubin</SURNAME>. <DATE>1977</DATE>.
Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1):1-38.
</REFERENCE>
<REFERENCE>
<REFLABEL>Duda and Hart 1973</REFLABEL>
Richard O. <SURNAME>Duda</SURNAME> and Peter E. <SURNAME>Hart</SURNAME>. <DATE>1973</DATE>. Pattern Classification and Scene
Analysis. Wiley-Interscience, New York, New York.
</REFERENCE>
<REFERENCE>
<REFLABEL>Hindle 1990</REFLABEL>
Donald <SURNAME>Hindle</SURNAME>. <DATE>1990</DATE>. Noun classification from predicate-argument structures. In 28th Annual
Meeting of the Association for Computational Linguistics, pages 268-275, Pittsburgh, Pennsylvania. Association for Computational
Linguistics, Morristown, New Jersey.
</REFERENCE>
<REFERENCE>
<REFLABEL>Hindle 1993</REFLABEL>
Donald <SURNAME>Hindle</SURNAME>. <DATE>1993</DATE>. A parser for text corpora. In B.T.S. Atkins and A. Zampoli, editors,
Computational Approaches to the Lexicon. Oxford University Press, Oxford, England. <DATE>To appear</DATE>.
</REFERENCE>
<REFERENCE>
<REFLABEL>Resnik 1992</REFLABEL>
Philip <SURNAME>Resnik</SURNAME>. <DATE>1992</DATE>. WordNet and distributional analysis: A class-based approach to lexical
discovery. In AAAI Workshop on Statistically-Based Natural-Language-Processing Techniques, San Jose, California, July.
</REFERENCE>
<REFERENCE>
<REFLABEL>Rose et al. 1990</REFLABEL>
Kenneth <SURNAME>Rose</SURNAME>, Eitan <SURNAME>Gurewitz</SURNAME>, and Geoffrey C. <SURNAME>Fox</SURNAME>.
<DATE>1990</DATE>. Statistical mechanics and phase transitions in clustering. Physical Review Letters, 65(8):945-948.
</REFERENCE>
<REFERENCE>
<REFLABEL>Schabes 1992</REFLABEL>
Yves <SURNAME>Schabes</SURNAME>. <DATE>1992</DATE>. Stochastic lexicalized tree-adjoining grammars. In Proceeedings of the 14th
International Conference on Computational Linguistics, Nantes, France.
</REFERENCE>
<REFERENCE>
<REFLABEL>Yarowsky 1992</REFLABEL>
David <SURNAME>Yarowsky</SURNAME>. <DATE>1992</DATE>. Personal communication.
</REFERENCE>
</REFERENCES>
</STRUCT-PAPER>
```

B.2. As Published

DISTRIBUTIONAL CLUSTERING OF ENGLISH WORDS

Fernando Pereira
AT&T Bell Laboratories
600 Mountain Ave.
Murray Hill, NJ 07974
pereira@research.att.com

Naftali Tishby
Dept. of Computer Science
Hebrew University
Jerusalem 91904, Israel
tishby@cs.huji.ac.il

Lillian Lee
Dept. of Computer Science
Cornell University
Ithaca, NY
llee@cs.cornell.edu

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word coocurrence, and the models evaluated with respect to held-out test data.

INTRODUCTION

Methods for automatically classifying words according to their contexts of use have both scientific and practical interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example of frequencies of pairs of a transitive main verb and the head noun of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct word classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden *senses classes* and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or *clusters* c with corresponding cluster membership probabilities $p(c|w)$ for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, $\mathcal{V}$ and $\mathcal{N}$, for the verbs and nouns in our experiments, and a single relation between them, in our experiments relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies f_{vn} of occurrence of particular pairs (v, n) in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, 1992). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster centroids) and associations between those hidden units.

For the noun classification problem, the empirical distribution of a noun n is then given by the conditional density $p_n(v) = f_{vn} / \sum_v f_{vn}$. The problem we study is how to use the p_n to classify the $n \in \mathcal{N}$. Our classification method will construct a set C of clusters and cluster membership probabilities $p(c|n)$. Each cluster c is associated to a cluster *centroid* p_c, which is discrete density over $\mathcal{V}$ obtained by averaging appropriately the p_n.

Distributional Similarity

To cluster nouns n according to their conditional verb distributions p_n, we need a measure of similarity between distributions. We use for this purpose the *relative entropy* or *Kullback-Leibler (KL) distance* between two distributions

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \quad .$$

This is a natural choice for a variety of reasons, which we will just sketch here.[1]

First of all, $D(p \parallel q)$ is zero just in case $p = q$, and it increases as the probability decreases that p is the relative frequency distribution of a random sample drawn according to p. More formally, the probability mass given by q to the set of all samples of length n with relative frequency distribution p is bounded by $2^{-nD(p\|q)}$ (Cover and Thomas, 1991). Therefore, if we are trying to distinguish among hypotheses q_i when p is the relative frequency distribution of observations, $D(p \parallel q_i)$ gives the relative weight of evidence in favor of q_i. Furthermore, a similar relation holds between $D(p \parallel p')$ for two empirical distributions p and p' and the probability that p and p' are drawn from the same distribution q. We can thus use the relative entropy between the context distributions for two words to measure how likely they are to be instances of the same cluster centroid.

From an information theoretic perspective $D(p \parallel q)$ measures how inefficient on average it would be to use a code based on q to encode a variable distributed according to p. With respect to our problem, $D(p_n \parallel p_c)$ thus gives us the loss of information in using cluster centroid p_c instead of the actual distribution for word p_n when modeling the distributional properties of n.

Finally, relative entropy is a natural measure of similarity between distributions for clustering because its minimization leads to cluster centroids that are a simple weighted average of member distributions.

One technical difficulty is that $D(p \parallel p')$ is not defined when $p'(x) = 0$ but $p(x) > 0$. We could sidestep this problem (as we did initially) by smoothing zero frequencies appropriately (Church and Gale, 1991). However, this is not very satisfactory because one of the goals of our work is precisely to avoid the problems of data sparseness by grouping words into classes. It turns out that the problem is avoided by our clustering technique, since it does not need to compute the KL distance between individual word distributions, but only between a word distribution and average distributions, the current cluster centroids, which are guaranteed to be nonzero whenever the word distributions are. This is a useful advantage of our method compared with agglomerative clustering techniques that need to compare individual objects being considered for grouping.

THEORETICAL BASIS

In general, we are interested on how to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams. We will show elsewhere that the theoretical analysis outlined here applies to that more general problem, but for now we will only address the more specific problem in which the objects are nouns and the contexts are verbs that take the nouns as direct objects.

Our problem can be seen as that of learning a joint distribution of pairs from a large sample of pairs. The pair coordinates come from two large sets $\mathcal{N}$ and $\mathcal{V}$, with no preexisting topological or metric structure, and the training data is a sequence S of N independently drawn pairs

$$S_i = (n_i, v_i) \qquad 1 \leq i \leq N \ .$$

From a learning perspective, this problem falls somewhere in between unsupervised and supervised learn-

[1] A more formal discussion will appear in our paper *Distributional Clustering*, in preparation.

ing. As in unsupervised learning, the goal is to learn the underlying distribution of the data. But in contrast to most unsupervised learning settings, the objects involved have no internal structure or attributes allowing them to be compared with each other. Instead, the only information about the objects is the statistics of their joint appearance. These statistics can thus be seem as a weak form of object labelling analogous to supervision.

Distributional Clustering

While clusters based on distributional similarity are interesting on their own, they can also be profitably seen as a means of summarizing a joint distribution. In particular, we would like to find a set of clusters $\mathcal{C}$ such that each conditional distribution $p_n(v)$ can be approximately decomposed as

$$\hat{p}_n(v) = \sum_{c\in\mathcal{C}} p(c|n)p_c(v) \quad ,$$

where $p(c|n)$ is the membership probability of n in c and $p_c(v) = p(v|c)$ is v's conditional probability given by the centroid distribution for cluster c.

The above decomposition can be written in a more symmetric form as

$$\begin{aligned} \hat{p}(n,v) &= \sum_{c\in\mathcal{C}} p(c,n)p(v|c) \\ &= \sum_{c\in\mathcal{C}} p(c)p(n|c)p(v|c) \qquad (1) \end{aligned}$$

assuming that $p(n)$ and $\hat{p}(n)$ coincide. We will take (1) as our basic clustering model.

To determine this decomposition we need to solve the two connected problems of finding find suitable forms for the cluster membership and centroid distributions $p(v|c)$, and of maximizing the goodness of fit between the model distribution $\hat{p}(n,v)$ and the observed data

Goodness of fit is determined by the model's likelihood of the observations. The maximum likelihood (ML) estimation principle is thus the natural tool to determine the centroid distributions $p_c(v)$.

As for the membership probabilities, they must be determined solely by the relevant measure of object-to-cluster similarity, which in the present work is the relative entropy between object and cluster centroid distributions. Since no other information is available, the membership is determined by maximizing the configuration entropy subject for a fixed average distortion. With the maximum entropy (ME) membership distribution, ML estimation is equivalent to the minimization of the average distortion of the data. The combined entropy maximization entropy and distortion minimization is carried out by a two-stage iterative process similar to the EM method (Dempster et al., 1977). The first stage of an iteration is a maximum likelihood, or minimum distortion, estimation of the cluster centroids given fixed membership probabilities. In the second iteration stage, the entropy of the membership distribution is maximized with a fixed average distortion. This joint optimization searches for a *saddle point* in the distortion-entropy parameters, which is equivalent to minimizing a linear combination of the two known as *free energy* in statistical mechanics. This analogy with statistical mechanics is not coincidental, and provide us with a better understanding of the clustering procedure.

Maximum Likelihood Cluster Centroids For the maximum likelihood argument, we start by estimating the likelihood of the sequence S of N independent observations of pairs (n_i, v_i). Using (1), the sequence's model log likelihood is

$$l(S) = \log \hat{p}(S) = \sum_{i=1}^{N} \log \sum_{c\in\mathcal{C}} p(c)p(n_i|c)p(v_i|c) \quad .$$

Fixing the number of clusters (model size) $|\mathcal{C}|$, we want to maximize $l(S)$ with respect to the distributions $p(n|c)$ and $p(v|c)$. The variation of $l(S)$ with respect to these distributions is

$$\delta l(S) = \sum_{i=1}^{N} \frac{1}{\hat{p}(n_i,v_i)} \sum_{c\in\mathcal{C}} p(c) \begin{pmatrix} p(v_i|c)\delta p(n_i|c) \\ + \\ p(n_i|c)\delta p(v_i|c) \end{pmatrix} \qquad (2)$$

with $p(n|c)$ and $p(v|c)$ kept normalized. Using Bayes's formula, we have [2]

$$p(n_i|c)p(v_i|c) = \frac{p(c|n_i,v_i)}{p(c)}\hat{p}(n_i,v_i) \ ,$$

or

$$\frac{1}{\hat{p}(n_i,v_i)} = \frac{p(c|n_i,v_i)}{p(c)p(n_i|c)p(v_i|c)}$$

for any c, which we substitute into (2) to obtain

$$\delta l(S) = \sum_{i=1}^{N} \sum_{c\in\mathcal{C}} p(c|n_i,v_i) \begin{pmatrix} \delta \log p(n_i|c) \\ + \\ \delta \log p(v_i|c) \end{pmatrix} \qquad (3)$$

since $\delta \log p = \delta p/p$. This expression is particularly useful when the cluster distributions $p(n|c)$ and $p(v|c)$

[2] As usual in clustering models (Duda and Hart, 1973), we assume that the model distribution and the empirical distribution are interchangeable at the solution of the parameter estimation equations, since the model is assumed to be able to represent correctly the data at that solution point. In practice, the data may not come exactly from the chosen model class, but the model obtained by solving the estimation equations may still be the closest one to the data.

are of exponential form, precisely what will be provided by the ME step described below.

At this point we need to specify the clustering model in more detail. In the derivation so far we have treated $p(n|c)$ and $p(v|c)$ symmetrically, corresponding to clusters not of verbs or nouns but of verb-noun associations. In principle such a symmetric model may be more accurate, but in this paper we will concentrate on *asymmetric models* in which cluster memberships are associated to just one of the components of the joint distribution and the cluster centroids are specified only by the other component. In particular, the model we use in our experiments has noun clusters with cluster memberships determined by $p(n|c)$ and centroid distributions determined by $p(v|c)$.

The asymmetric model simplifies the estimation significantly by dealing with a single component, but it has the disadvantage that the joint distribution, $p(n, v)$ has two different and not necessarily consistent expressions in terms of asymmetric models for the two coordinates.

Maximum Entropy Cluster Membership While variations of $p(n|c)$ and $p(v|c)$ in equation (3 are not independent, we can treat them separately. First, for fixed average distortion between the cluster centroid distributions $p(v|c)$ and the data $p(v|n)$, we find the cluster membership probabilities, which are the Bayes's inverses of the $p(n|c)$, that maximize the entropy of the cluster distributions. With the membership distributions thus obtained, we then look for the $p(v|c)$ that maximize the log likelihood $l(S)$. It turns out that this will also be the values of $p(v|c)$ that minimize the average distortion between the asymmetric cluster model and the data.

Given any similarity measure $d(n, c)$ between nouns and cluster centroids, the average cluster distortion is

$$\langle D\rangle = \sum_{n\in\mathcal{N}}\sum_{c\in C} p(c|n)d(n,c) \qquad (4)$$

If we maximize the cluster membership entropy

$$H = -\sum_{n\in\mathcal{N}}\sum_{c\in C} p(c|n)\log p(n|c) \qquad (5)$$

subject to normalization of $p(n|c)$ and fixed (4), we obtain the following standard exponential forms for the class and membership distributions

$$p(n|c) = \frac{1}{Z_c}\exp -\beta d(n,c) \qquad (6)$$

$$p(c|n) = \frac{1}{Z_n}\exp -\beta d(n,c) \qquad (7)$$

where the normalization sums (partition functions) are $Z_c = \sum_n \exp -\beta d(n, c)$ and $Z_n = \sum_c \exp -\beta d(n, c)$. Notice that $d(n, c)$ does not need to be symmetric for this derivation, as the two distributions are simply related by Bayes's rule.

Returning to the log-likelihood variation (3), we can now use (6) for $p(n|c)$ and the assumption for the asymmetric model that the cluster membership stays fixed as we adjust the centroids, to obtain

$$\delta l(S) = -\sum_{i=1}^{N}\sum_{c\in C} p(c|n_i)\delta\beta d(n_i,c) + \delta\log Z_c \qquad (8)$$

where the variation of $p(v|c)$ is now included in the variation of $d(n, c)$.

For a large enough sample, we may replace the sum over observations in (8) by the average over $\mathcal{N}$

$$\delta l(S) = -\sum_{n\in N} p(n)\sum_{c\in C} p(c|n)\delta\beta d(n,c) + \delta\log Z_c$$

which, applying Bayes's rule, becomes

$$\delta l(S) = -\sum_{c\in C}\frac{1}{p(c)}\sum_{n\in N} p(n|c)\delta\beta d(n,c) + \delta\log Z_c \quad (9)$$

At the log-likelihood maximum, the variation (9) must vanish. We will see below that the use of relative entropy for similarity measure makes $\delta \log Z_c$ vanish at the maximum as well, so the log likelihood can be maximized by minimizing the average distortion with respect to the class centroids while class membership is kept fixed

$$\sum_{c\in C}\frac{1}{p(c)}\sum_{n\in\mathcal{N}} p(n|c)\delta d(n,c) = 0 \quad ,$$

or, sufficiently, if each of the inner sums vanish

$$\sum_{c\in C}\sum_{n\in\mathcal{N}} p(n|c)\delta d(n,c) = 0 \qquad (10)$$

Minimizing the Average KL Distortion We first show that the minimization of the relative entropy yields the natural expression for cluster centroids

$$p(v|c) = \sum_{n\in\mathcal{N}} p(n|c)p(v|n) \qquad (11)$$

To minimize the average distortion (10), we observe that the variation of the KL distance between noun and centroid distributions with respect to the centroid distribution $p(v|c)$, with each centroid distribution normalized by the Lagrange multiplier λ_c, is given by

$$\begin{aligned}\delta d(n,c) &= \delta\left(\begin{array}{c} -\sum_{v\in\mathcal{V}} p(v|n)\log p(v|c) \\ + \\ \lambda_c(\sum_{v\in\mathcal{V}} p(v|c) - 1) \end{array}\right) \\ &= \sum_{v\in\mathcal{V}}\left(-\frac{p(v|n)}{p(v|c)} + \lambda_c\right)\delta p(v|c) \quad .\end{aligned}$$

Substituting this expression into (10), we obtain

$$\sum_c \sum_n \sum_v \left(-\frac{p(v|n)p(n|c)}{p(v|c)} + \lambda_c \right) \delta p(v|c) = 0 \quad .$$

Since the $\delta p(v|c)$ are now independent, we obtain immediately the desired centroid expression (11), which is the desired weighted average of noun distributions.

We can now see that the variation $\delta \log Z_c$ vanishes for centroid distributions given by (11), since it follows from (10) that

$$\begin{aligned} \delta \log Z_c &= -\frac{\beta}{Z_c} \sum_n \exp - \beta d(n,c) \delta d(n,c) \\ &= -\beta \sum_n p(n|c) \delta d(x,c) = 0. \end{aligned}$$

The Free Energy Function The combined minimum distortion and maximum entropy optimization is equivalent to the minimization of a single function, the *free energy*

$$\begin{aligned} F &= -\frac{1}{\beta} \sum_n \log Z_n \\ &= \langle D \rangle - H/\beta \end{aligned}$$

where $\langle D \rangle$ is the average distortion (4) and H is the cluster membership entropy (5).

The free energy determines both the distortion and the membership entropy through

$$\begin{aligned} \langle D \rangle &= \frac{\partial \beta F}{\partial \beta} \\ H &= -\frac{\partial F}{\partial T} \quad , \end{aligned}$$

with *temperature* $T = \beta^{-1}$.

The most important property of the free energy is that its minimum determines the balance between the "disordering" maximum entropy and "ordering" distortion minimization in which the system is most likely to be found. In fact the probability to find the system at a given configuration is exponential in F

$$P \propto \exp - \beta F \quad ,$$

so a system is most likely to be found in its minimal free energy configuration.

Hierarchical Clustering

The analogy with statistical mechanics suggests a *deterministic annealing* procedure for clustering (Rose et al., 1990), in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter β following an *annealing schedule*.

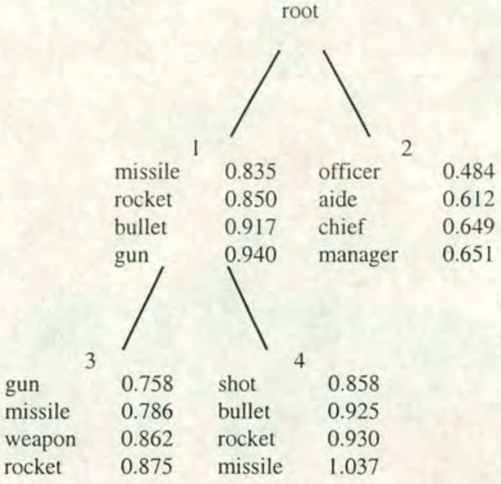


Figure 1: Direct object clusters for *fire*

The higher β, the more local is the influence of each noun on the definition of centroids. The dissimilarity plays here the role of distortion. When the scale parameter β is close to zero, the dissimilarities are almost irrelevant, all words contribute about equally to each centroid, and so the lowest average distortion solution involves just one cluster which is the average of all word densities. As β is slowly increased, a point (phase transition) is eventually reached which the natural solution involves two distinct centroids. We say then that the original cluster has *split* into the two new clusters.

In general, if we take any cluster c and a *twin* c' of c such that the centroid $p_{c'}$ is a small random pertubation of p_c, below the critical β at which c splits the membership and centroid reestimation procedure given by equations (7) and (11) will make p_c and $p_{c'}$ converge, that is, c and c' are really the same cluster. But with β above the critical value for c, the two centroids will diverge, giving rise to two daughters of c.

Our clustering procedure is thus as follows. We start with very low β and a single cluster whose centroid is the average of all noun distributions. For any given β, we have a current set of *leaf* clusters corresponding to the current free energy (local) minimum. To refine such a solution, we search for the lowest β which is the critical value for some current leaf cluster splits. Ideally, there is just one split at that critical value, but for practical performance and numerical accuracy reasons we may have several splits at the new critical point. The splitting procedure can then be repeated to achieve the desired number of clusters or model cross-entropy.

CLUSTERING EXAMPLES

All our experiments involve the asymmetric model described in the previous section. As explained there, our clustering procedure yields for each value of β a set C_β of clusters minimizing the free energy F, and the asymmetric model for β estimates the conditional verb distribution for a noun n by

$$\hat{p}_n = \sum_{c \in C_\beta} p(c|n) p_c$$

where $p(c|n)$ also depends on β.

As a first experiment, we used our method to classify the 64 nouns appearing most frequently as heads of direct objects of the verb "fire" in one year (1988) of Associated Press newswire. In this corpus, the chosen nouns appear as direct object heads of a total of 2147 distinct verbs, so each noun is represented by a density over the 2147 verbs.

Figure 1 shows the five words most similar to the each cluster centroid for the four clusters resulting from the first two cluster splits. It can be seen that first split separates the objects corresponding to the weaponry sense of "fire" (cluster 1) from the ones corresponding to the personnel action (cluster 2). The second split then further refines the weaponry sense into a projectile sense (cluster 3) and a gun sense (cluster 4). That split is somewhat less sharp, possibly because not enough distinguishing contexts occur in the corpus.

Figure 2 shows the four closest nouns to the centroid of each of a set of hierarchical clusters derived from verb-object pairs involving the 1000 most frequent nouns in the June 1991 electronic version of Grolier's Encyclopedia (10 million words).

MODEL EVALUATION

The preceding qualitative discussion provides some indication of what aspects of distributional relationships may be discovered by clustering. However, we also need to evaluate clustering more rigorously as a basis for models of distributional relationships. So, far, we have looked at two kinds of measurements of model quality: (i) relative entropy between held-out data and the asymmetric model, and (ii) performance on the task of deciding which of two verbs is more likely to take a given noun as direct object when the data relating one of the verbs to the noun has been witheld from the training data.

The evaluation described below was performed on the largest data set we have worked with so far, extracted from 44 million words of 1988 Associated Press newswire with the pattern matching techniques mentioned earlier. This collection process yielded 1112041 verb-object pairs. We selected then the subset involving the 1000 most frequent nouns in the corpus for clustering, and randomly divided it into a training set of 756721 pairs and a test set of 81240 pairs.

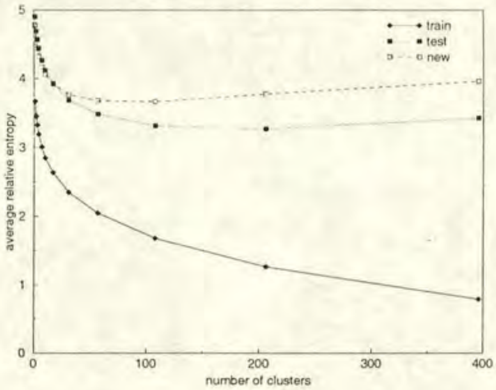


Figure 3: Asymmetric Model Evaluation, AP88 Verb-Direct Object Pairs

Relative Entropy

Figure 3 plots the average relative entropy of several data sets to asymmetric clustered models of different sizes, given by

$$\sum_n D(t_n \| \hat{p}_n)$$

where t_n is the relative frequency distribution of verbs taking n as direct object in the test set. For each critical value of β, we show the relative entropy with respect to the asymmetric model based on C_β of the training set (set *train*), of randomly selected held-out test set (set *test*), and of held-out data for a further 1000 nouns that were not clustered (set *new*). Unsurprisingly, the training set relative entropy decreases monotonically. The test set relative entropy decreases to a minimum at 206 clusters, and then starts increasing, suggesting that larger models are overtrained.

The new noun test set is intended to test whether clusters based on the 1000 most frequent nouns are useful classifiers for the selectional properties of nouns in general. As the figure shows, the cluster model provides over one bit of information about the selectional properties of the new nouns, but the overtraining effect is even sharper than for the held-out data involving the 1000 clustered nouns.

Decision Task

We also evaluated asymmetric cluster models on a verb decision task closer to possible applications to disambiguation in language analysis. The task consists judging which of two verbs v and v' is more likely to take a

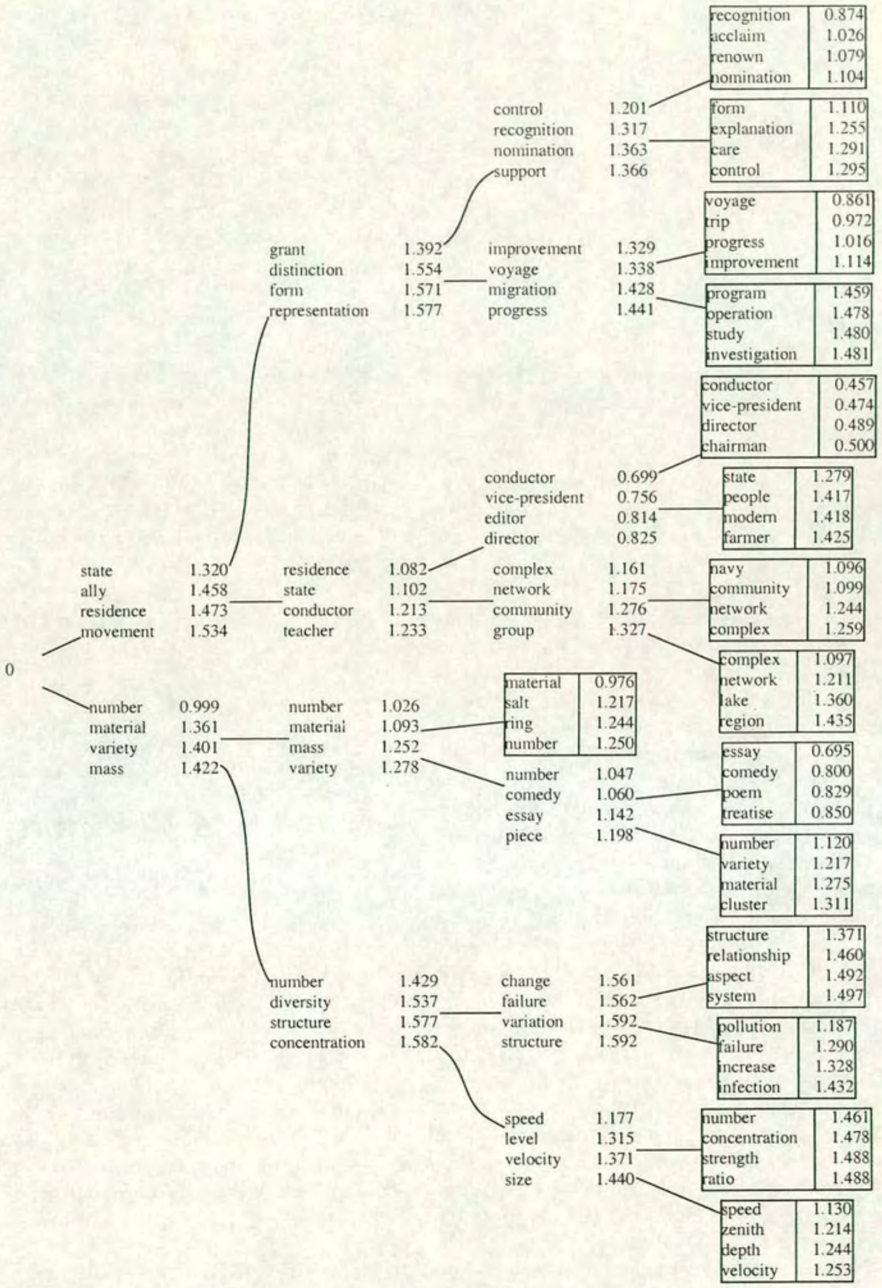


Figure 2: Noun Clusters for Grolier's Encyclopedia

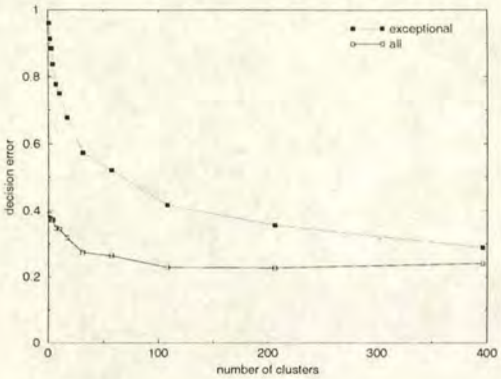


Figure 4: Pairwise Verb Comparisons, AP88 Verb-Direct Object Pairs

given noun n as object, when all occurrences of (v, n) in the training set were deliberately deleted. Thus this test evaluates how well the models reconstruct missing data in the verb distribution for n from the cluster centroids close to n.

The data for this test was built from the training data for the previous one in the following way, based on a suggestion by Dagan *et al.* (1992). A small number (104) of (v, n) pairs with a fairly frequent verb (between 500 and 5000 occurrences) was randomly picked, and all occurrences of each pair in the training set were deleted. The resulting training set was used to build a sequence of cluster models as before. Each model was used to decide which of two verbs v and v' are more likely to appear with a noun n where the (v, n) data was deleted from the training set, and the decisions compared with the corresponding ones derived from the original event frequencies in the initial data set. More specifically, for each deleted pair (v, n) and each verb v' that occurred with n in the initial data either at least twice as frequently or at most half as frequently as v, we compared the sign of $\log \hat{p}_n(v)/\hat{p}_n(v')$ with that of $\log p_n(v)/p_n(v')$ for the initial data set. The error rate for each model is simply the proportion of sign disagreements in the selected (v, n, v') triples. Figure 4 shows the error rates for each model for all the selected (v, n, v') (*all*) and for just those *exceptional* triples in which the log frequency ratio of (n, v) and (n, v') differs from the log marginal frequency ratio of v and v'. In other words, the exceptional cases are those in which predictions based just on the marginal frequencies, which the initial one-cluster model represents, would be consistently wrong.

Here too we see some overtraining for the largest models considered, although not for the exceptional verbs.

CONCLUSIONS

We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words. The resulting clusters are intuitively informative, and can be used to construct class-based word coocurrence models with substantial predictive power.

While the clusters derived by the proposed method seem in many cases semantically significant, this intuition needs to be grounded in a more rigorous assessment. In addition to predictive power evaluations of the kind we have already carried out, it might be worth comparing automatically-derived clusters with human judgements in a suitable experimental setting.

Moving further in the direction of class-based language models, we plan to consider additional distributional relations (for instance, adjective-noun) and apply the results of clustering to the grouping of lexical associations in lexicalized grammar frameworks such as stochastic lexicalized tree-adjoining grammars (Schabes, 1992).

ACKNOWLEDGMENTS

We would like to thank Don Hindle for making available the 1988 Associated Press verb-object data set, the Fidditch parser and a verb-object structure filter, Mats Rooth for selecting the objects of "fire" data set and many discussions, David Yarowsky for help with his stemming and concordancing tools, and Ido Dagan for suggesting ways of testing cluster models.

REFERENCES

[Brown et al.1990]
Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1990. Class-based n-gram models of natural language. In *Proceedings of the IBM Natural Language ITL*, pages 283–298, Paris, France, March.

[Church and Gale1991] Kenneth W. Church and William A. Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5:19–54.

[Church1988] Kenneth W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin, Texas. Association for Computational Linguistics, Morristown, New Jersey.

[Cover and Thomas1991] Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory.* Wiley-Interscience, New York, New York.

[Dagan et al.1992] Ido Dagan, Shaul Markus, and Shaul Markovitch. 1992. Contextual word similarity and the estimation of sparse lexical relations. Submitted for publication.

[Dempster et al.1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

[Duda and Hart1973] Richard O. Duda and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis.* Wiley-Interscience, New York, New York.

[Hindle1990] Donald Hindle. 1990. Noun classification from predicate-argument structures. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, Pennsylvania. Association for Computational Linguistics, Morristown, New Jersey.

[Hindle1993] Donald Hindle. 1993. A parser for text corpora. In B.T.S. Atkins and A. Zampoli, editors, *Computational Approaches to the Lexicon.* Oxford University Press, Oxford, England. To appear.

[Resnik1992] Philip Resnik. 1992. WordNet and distributional analysis: A class-based approach to lexical discovery. In *AAAI Workshop on Statistically-Based Natural-Language-Processing Techniques*, San Jose, California, July.

[Rose et al.1990] Kenneth Rose, Eitan Gurewitz, and Geoffrey C. Fox. 1990. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65(8):945–948.

[Schabes1992] Yves Schabes. 1992. Stochastic lexicalized tree-adjoining grammars. In *Proceeedings of the 14th International Conference on Computational Linguistics*, Nantes, France.

[Yarowsky1992] David Yarowsky. 1992. Personal communication.

This paper appeared in the *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 183–190. Ohio State University, Columbus, Ohio, 1993. Reprinted by permission of the Association for Computational Linguistics and the individual authors.

B.3. RDP

1. Solution Identifier —

2. Specific Aim/Scope

164 to group words according to their participation in particular grammatical relations with other words

10 how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves

44 how to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.

11 how to derive the classes directly from distributional data

46 learning a joint distribution of pairs from a large sample of pairs.

22 we will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs

45 we will only address the more specific problem in which the objects are nouns and the contexts are verbs that take the nouns as direct objects.

3. Background

Aim	Problem/Phenomenon
1 automatically classifying words	**4** The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.

4. Solution/Inventive Step

164 a general divisive clustering procedure for probability distributions can be used...

12 we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN> for each word w.

5. Claim/Conclusion

165 The resulting clusters are intuitively informative, and can be used to construct class-based word coocurrence models with substantial predictive power.

6. Rival/Contrast

Reference	Solution ID	Type of Contrast
• **5** [Hindle 1990]		**9** it is not clear how it can be used directly to construct word classes and corresponding models of association.
• **13** [Brown et al. 1992]	**13** other class-based modeling techniques	**13** Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information.

6. RIVAL/CONTRAST (CT'D)

REFERENCE	SOLUTION ID	TYPE OF CONTRAST
• **11** [Resnik 1992]		**11** preexisting sense classes (Resnik) vs. we derive the classes directly from distributional data.
•	**43** agglomerative clustering techniques	**43** need to compare individual objects being considered for grouping. (advantage of our method)
• **40** [Church and Gale 1991]	**40** smoothing zero frequencies appropriately	**41** However, this is not very satisfactory as our goal is to avoid the problems of data sparseness by clustering words together

7. BASIS/CONTINUATION

REFERENCE	SOLUTION ID	TYPE OF CONTINUATION
• **113** [Rose et al. 1990]	**113** deterministic annealing	**113** The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering [Rose et al. 1990] . . .
• **155** [Dagan et al. 1993]		**155** based on a suggestion by
•	**29** Kullback-Leibler (KL) distance	**29** used
• **19** [Hindle 1993]		**19** automatically parsed by Hindle's parser
• **20** [Church 1988]		**20** with the help of a statistical part-of-speech tagger
• **20** [Yarowsky 1992]		**20** [with the help of] tools for regular expression pattern matching on tagged corpora

EXTERNAL STRUCTURE

HEADLINES		8. TEXTUAL STRUCTURE
1.	Introduction	
1.1	Problem Setting	
1.2	Distributional Similarity	
2.	Theoretical Basis	
2.1	Distributional Clustering	
2.1.1.	Maximum Likelihood Cluster Centroids	
2.1.2.	Maximum Entropy Cluster Membership	
2.1.3.	Minimizing the Average KL Distortion	
2.1.4.	The Free Energy Function	
2.2.	Hierarchical Clustering	
3.	Clustering Examples	**127** All our experiments involve the asymmetric model described in the previous section.
4.	Model Evaluation	
4.1.	Relative Entropy	
4.2.	Decision Task	
5.	Conclusions	

B.4. RDP Sentence Material

SPECIFIC AIM/SCOPE

10 Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves.

11 While it may be worthwhile to base such a model on preexisting sense classes [Resnik 1992], in the work described here we look at how to derive the classes directly from distributional data.

22 We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar.

44 In general, we are interested on how to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.

45 We will show elsewhere that the theoretical analysis outlined here applies to that more general problem, but for now we will only address the more specific problem in which the objects are nouns and the contexts are verbs that take the nouns as direct objects.

46 Our problem can be seen as that of learning a joint distribution of pairs from a large sample of pairs.

164 We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words.

BACKGROUND (AIM)

1 Methods for automatically classifying words according to their contexts of use have both scientific and practical interest.

BACKGROUND (PROBLEM/PHENOMENON)

4 The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.

SOLUTION/INVENTIVE STEP

12 More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN> for each word w.

164 We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words.

CLAIM/CONCLUSION

165 The resulting clusters are intuitively informative, and can be used to construct class-based word coocurrence models with substantial predictive power.

RIVAL/CONTRAST

5 [Hindle 1990] proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen.

9 His notion of similarity seems to agree with our intuitions in many cases, but is not clear how it can be used directly to construct word classes and corresponding models of association.

11 While it may be worthwhile to base such a model on preexisting sense classes [Resnik 1992], in the work described here we look at how to derive the classes directly from distributional data.

13 Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes [Brown et al. 1990].

14 Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information as we noted above.

40 We could sidestep this problem (as we did initially) by smoothing zero frequencies appropriately [Church and Gale 1991].

41 However, this is not very satisfactory as our goal is to avoid the problems of data sparseness by clustering words together.

43 This is a useful advantage of our method compared with agglomerative clustering techniques that need to compare individual objects being considered for grouping.

BASIS/CONTINUATION

19 The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch [Hindle 1993].

20 More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger [Church 1988] and of tools for regular expression pattern matching on tagged corpora [Yarowsky 1992].

29 We use for this purpose the relative entropy or Kullback-Leibler (KL) distance between two distributions.

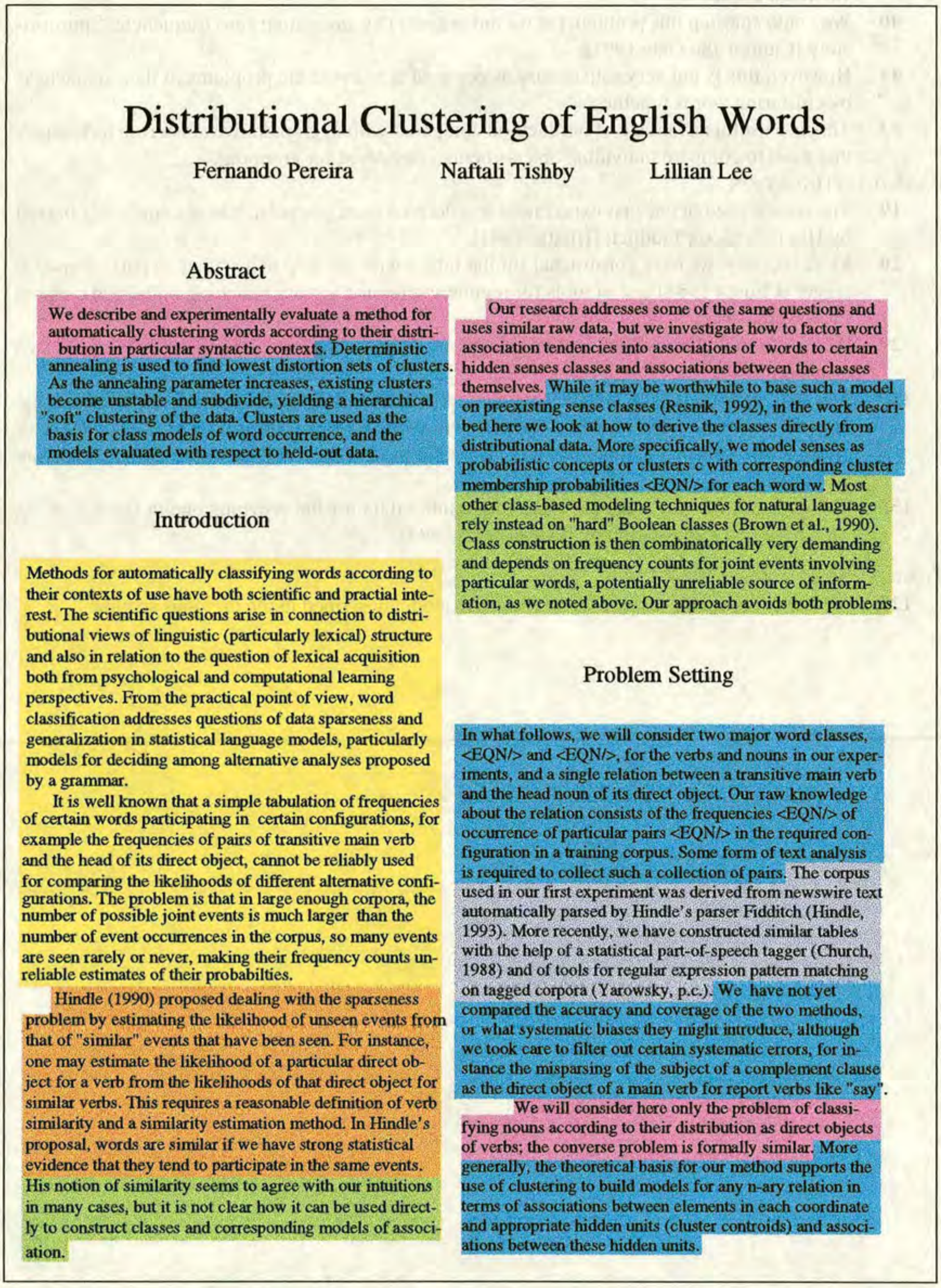
113 The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering [Rose et al. 1990], in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter <EQN> following an annealing schedule.

155 The data for this test was built from the training data for the previous one in the following way, based on a suggestion by [Dagan et al. 1993].

TEXTUAL STRUCTURE

127 All our experiments involve the asymmetric model described in the previous section.

B.5. Human Annotation (Annotator A)



Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

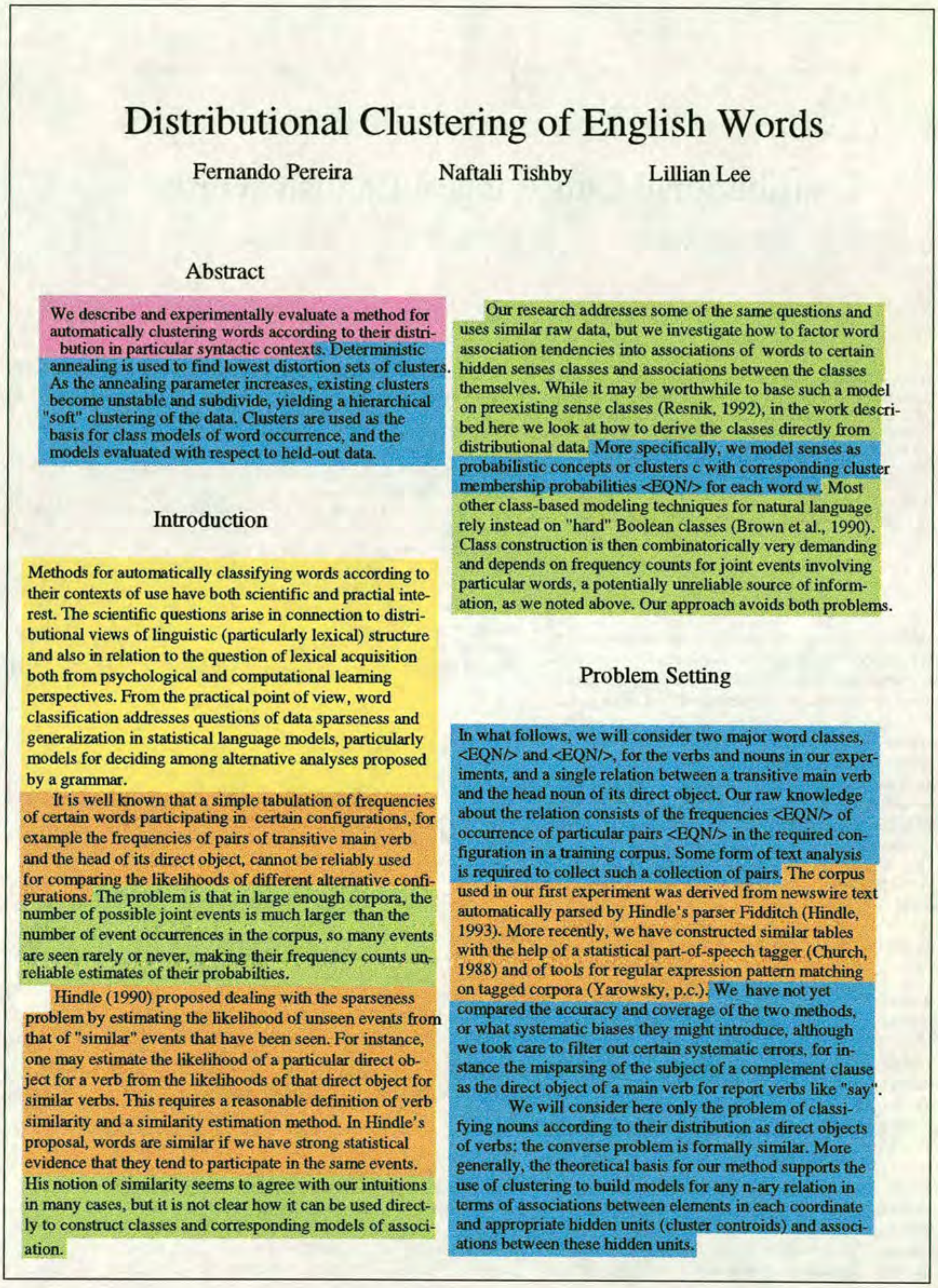
Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster controids) and associations between these hidden units.

B.6. Human Annotation (Annotator B)



Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

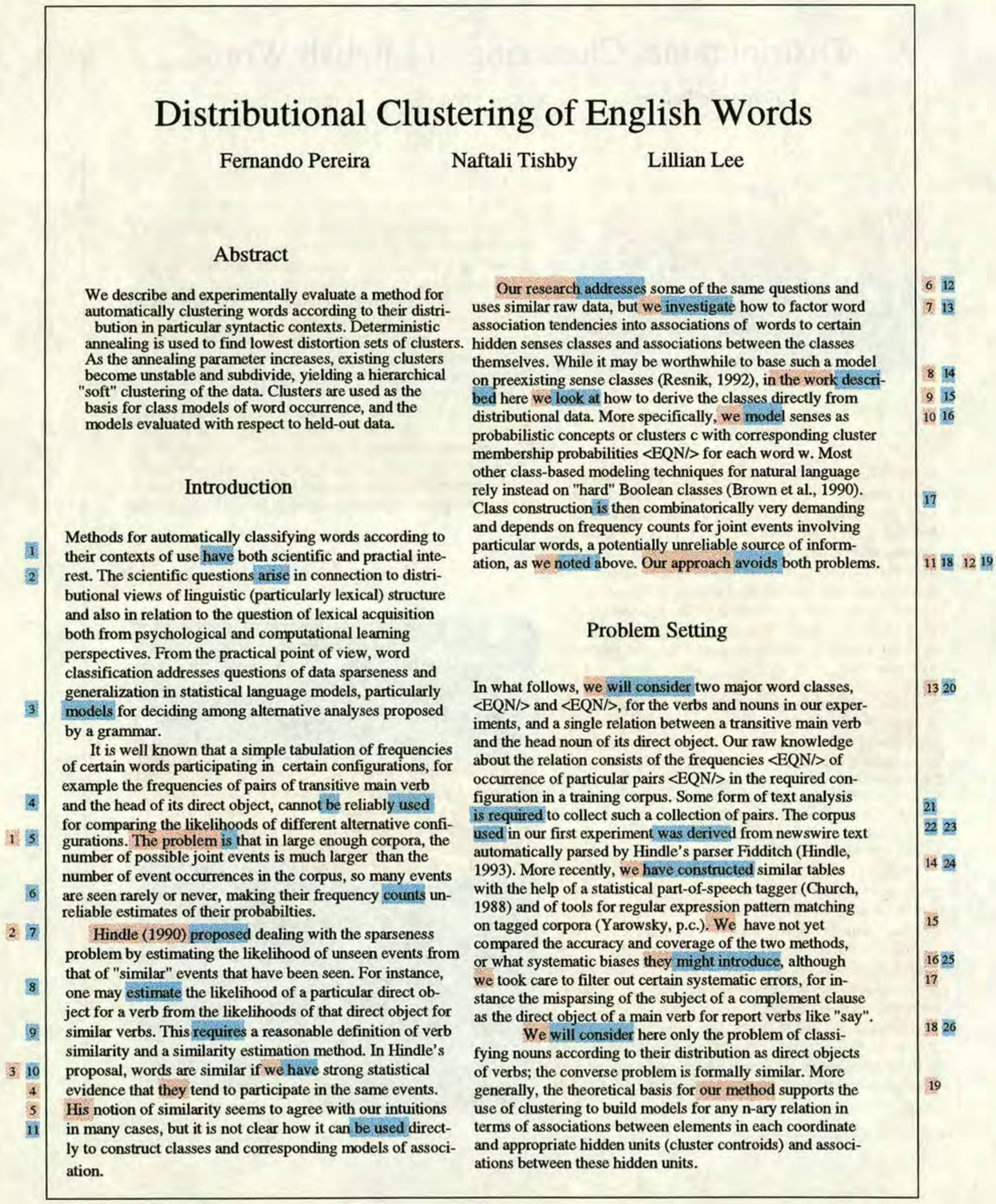
Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster centroids) and associations between these hidden units.

B.7. Agent and Action Recognition



Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to
1 their contexts of use have both scientific and practial interest.
2 The scientific questions arise in connection to distributional
views of linguistic (particularly lexical) structure
and also in relation to the question of lexical acquisition
both from psychological and computational learning
perspectives. From the practical point of view, word
classification addresses questions of data sparseness and
generalization in statistical language models, particularly
3 models for deciding among alternative analyses proposed
by a grammar.

It is well known that a simple tabulation of frequencies
of certain words participating in certain configurations, for
example the frequencies of pairs of transitive main verb
4 and the head of its direct object, cannot be reliably used
for comparing the likelihoods of different alternative configurations.
1 5 The problem is that in large enough corpora, the
number of possible joint events is much larger than the
number of event occurrences in the corpus, so many events
6 are seen rarely or never, making their frequency counts unreliable
estimates of their probabilties.

2 7 Hindle (1990) proposed dealing with the sparseness
problem by estimating the likelihood of unseen events from
that of "similar" events that have been seen. For instance,
8 one may estimate the likelihood of a particular direct object
for a verb from the likelihoods of that direct object for
9 similar verbs. This requires a reasonable definition of verb
similarity and a similarity estimation method. In Hindle's
3 10 proposal, words are similar if we have strong statistical
4 evidence that they tend to participate in the same events.
5 His notion of similarity seems to agree with our intuitions
11 in many cases, but it is not clear how it can be used directly
to construct classes and corresponding models of association.

Our research addresses some of the same questions and 6 12
uses similar raw data, but we investigate how to factor word 7 13
association tendencies into associations of words to certain
hidden senses classes and associations between the classes
themselves. While it may be worthwhile to base such a model
on preexisting sense classes (Resnik, 1992), in the work described 8 14
here we look at how to derive the classes directly from 9 15
distributional data. More specifically, we model senses as 10 16
probabilistic concepts or clusters c with corresponding cluster
membership probabilities <EQN/> for each word w. Most
other class-based modeling techniques for natural language
rely instead on "hard" Boolean classes (Brown et al., 1990).
Class construction is then combinatorically very demanding 17
and depends on frequency counts for joint events involving
particular words, a potentially unreliable source of information,
as we noted above. Our approach avoids both problems. 11 18 12 19

Problem Setting

In what follows, we will consider two major word classes, 13 20
<EQN/> and <EQN/>, for the verbs and nouns in our experiments,
and a single relation between a transitive main verb
and the head noun of its direct object. Our raw knowledge
about the relation consists of the frequencies <EQN/> of
occurrence of particular pairs <EQN/> in the required configuration
in a training corpus. Some form of text analysis
is required to collect such a collection of pairs. The corpus 21
used in our first experiment was derived from newswire text 22 23
automatically parsed by Hindle's parser Fidditch (Hindle,
1993). More recently, we have constructed similar tables 14 24
with the help of a statistical part-of-speech tagger (Church,
1988) and of tools for regular expression pattern matching
on tagged corpora (Yarowsky, p.c.). We have not yet 15
compared the accuracy and coverage of the two methods,
or what systematic biases they might introduce, although 16 25
we took care to filter out certain systematic errors, for instance 17
the misparsing of the subject of a complement clause
as the direct object of a main verb for report verbs like "say".

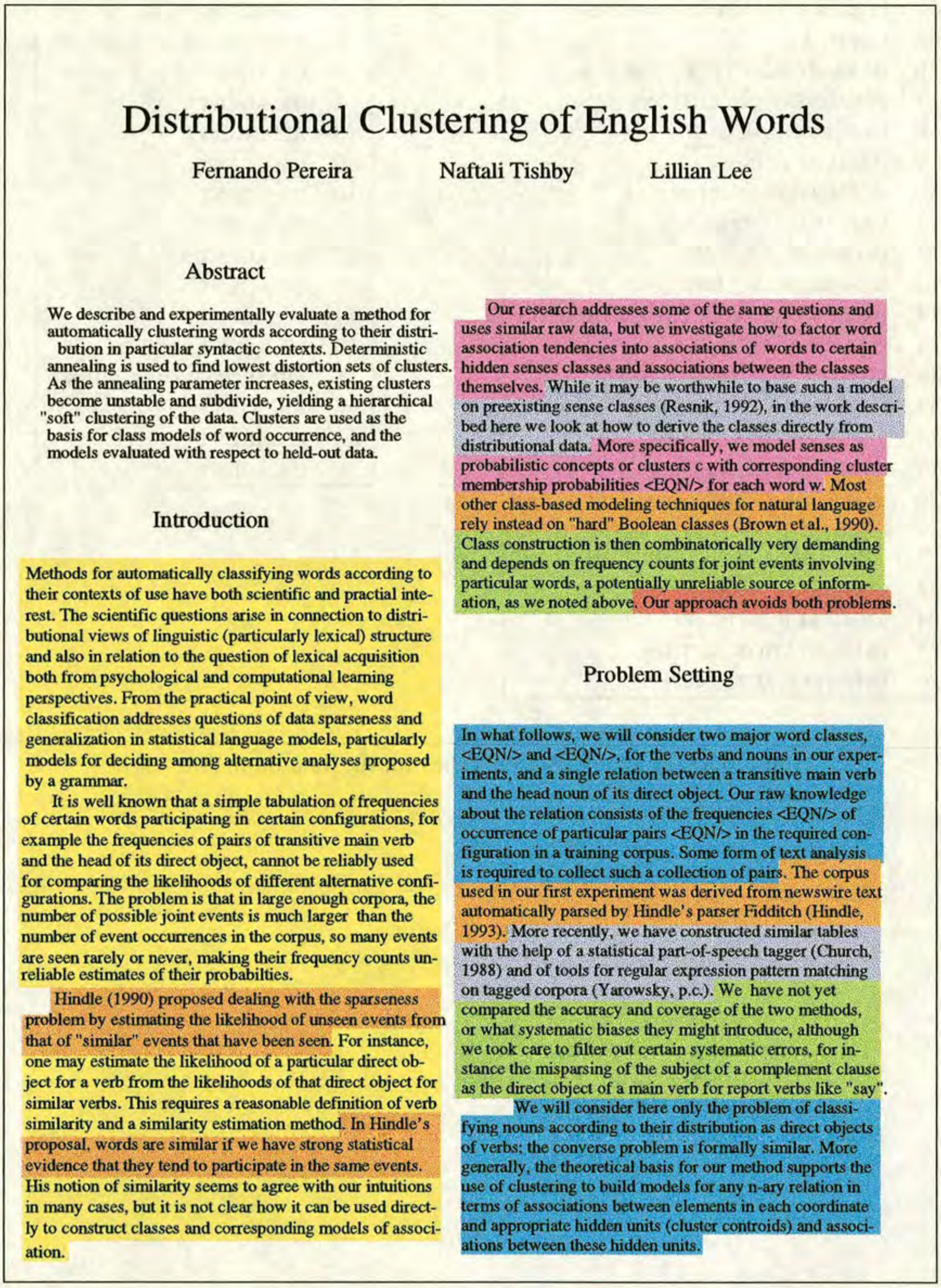
We will consider here only the problem of classifying 18 26
nouns according to their distribution as direct objects
of verbs; the converse problem is formally similar. More
generally, the theoretical basis for our method supports the 19
use of clustering to build models for any n-ary relation in
terms of associations between elements in each coordinate
and appropriate hidden units (cluster controids) and associations
between these hidden units.

	Actions (blue)
1	POSSESSION_ACTION
2	PROBLEM_ACTION
3	SOLUTION_ACTION (POS-error)
4	negated USE_ACTION (passive)
5	COPULA
6	RESEARCH_ACTION (POS-error)
7	PRESENTATION_ACTION
8	RESEARCH_ACTION
9	NEED_ACTION
10	POSSESSION_ACTION
11	USE_ACTION (passive)
12	INTEREST_ACTION
13	RESEARCH_ACTION
14	PRESENTATION_ACTION (POS-error)
15	INTEREST_ACTION
16	SOLUTION_ACTION
17	COPULA
18	PRESENTATION_ACTION
19	SOLUTION_ACTION
20	INTEREST_ACTION
21	NEED_ACTION
22	USE_ACTION (POS-error)
23	CONTINUE_ACTION
24	RESEARCH_ACTION
25	PRESENTATION_ACTION
26	INTEREST_ACTION

	Agents (pink)
1	PROBLEM_AGENT
2	THEM_AGENT
3	US_AGENT
4	THEM_PRONOUN_AGENT
5	THEM_PRONOUN_AGENT
6	US_AGENT
7	US_AGENT
8	REF_AGENT
9	US_AGENT
10	US_AGENT
11	US_AGENT
12	US_AGENT
13	US_AGENT
14	US_AGENT
15	US_AGENT
16	THEM_PRONOUN_AGENT
17	US_AGENT
18	US_AGENT
19	US_AGENT

Figure B.1: Agent and Action Types for the Text on p. 300

B.8. Automatic Annotation (Naive Bayes)



Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

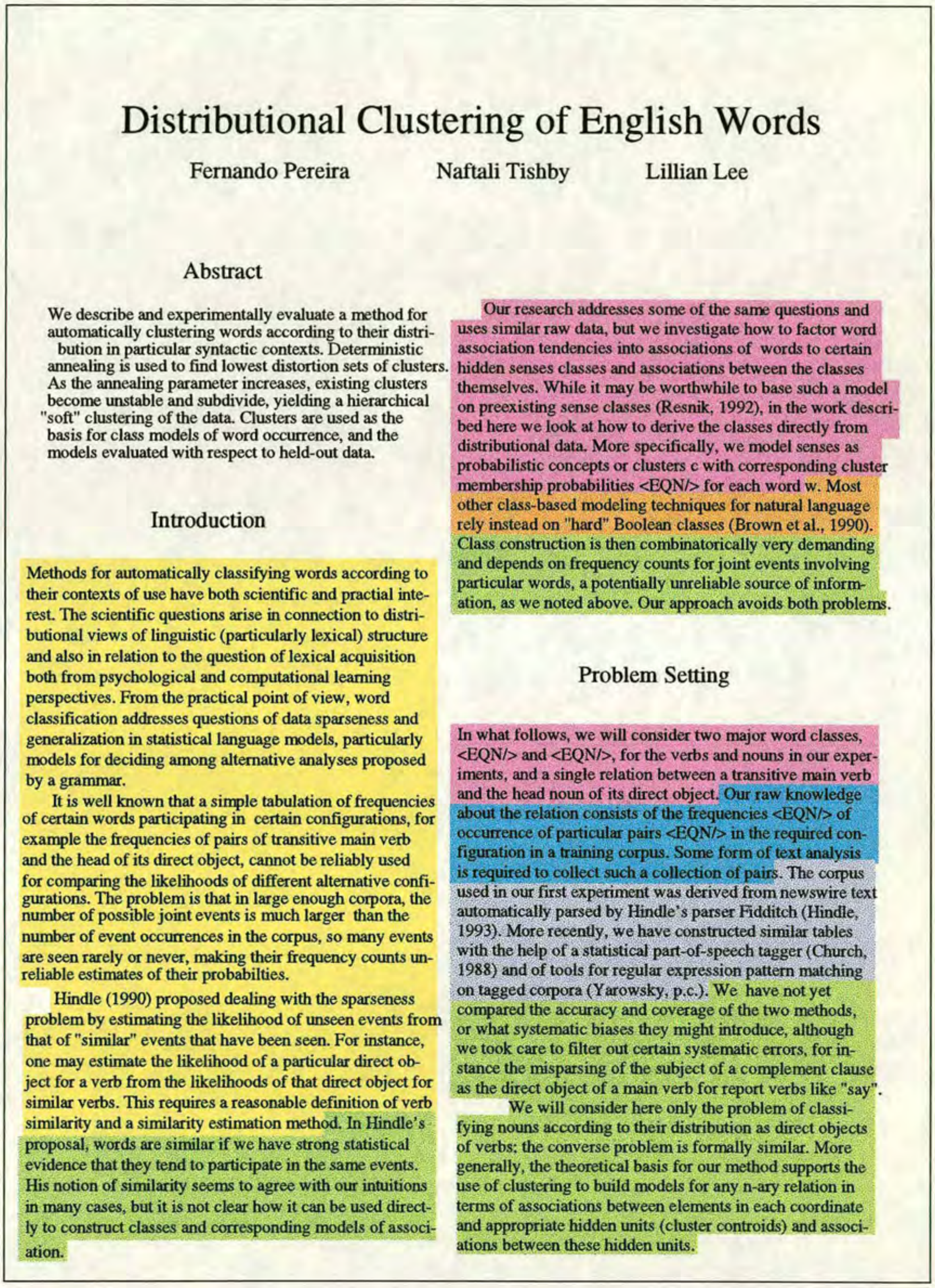
Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster centroids) and associations between these hidden units.

B.9. Automatic Annotation (N-Gram)



Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster controids) and associations between these hidden units.

Appendix C

Annotation Materials

C.1. Study I: Guidelines for Human Annotation of Basic Scheme

Principles of annotation

These guidelines describe a classification scheme for scientific papers which annotates the *ownership* of scientific ideas. Segmentation of ownership identifies segments in the paper where authors describe general statements about the field, other researcher's work and their own work, cf. C.1.

BACKGROUND	Generally accepted background knowledge
OTHER	Specific other work
OWN	Own work: method, results, future work…

Figure C.1: Overview of annotation scheme

Each of the classes is associated with a colour, and these colours are matched with marker pens. Please use these to mark your judgement on the printout of the papers.

Annotate from the author's perspective and their opinion about what is general, specific and their own claim, even if you might not agree with the portrayal of the situation as presented in the paper.

The unit of annotation is always the whole sentence. Annotation is mutually exclusive and proceeds sentence by sentence: once you have decided to assign a certain class, you can immediately go to the next sentence, as a sentence cannot have more than one class.

Please annotate all sentences in the abstract, and all sentences in the document except acknowledgement sentences.

Description of classes

BACKGROUND

BACKGROUND knowledge marks sentences which are presented as uncontroversial in the field. In such sentences, the research context is established. This includes statements of general capacity of the field, general problems, research goals, methodologies and general solutions (*"In recent years, there has been a growing interest in the field of X in the subject of Y"*). The most prototypical use of BACKGROUND is in the beginning of the paper.

Examples for general problems:

- *One of the difficult problems in machine translation from Japanese to English or other European languages is the treatment of articles and numbers.*
- *Complications arise in spelling rule application from the fact that, at compile time, neither the lexical nor the surface form of the root, nor even its length, is known.*
- *Collocations present specific problems in translation, both in human and automatic contexts.*

Examples for generally accepted/old solutions or claims:

- *Tagging by means of a Hidden Markov Model (HMM) is widely recognised as an effective technique for assigning parts of speech to a corpus in a robust and efficient manner.*
- *Current research in lexical aquisition is eminently knowledge-based.*
- *Literature in psychology has amply demonstrated that children do not acquire [. . .]*

In linguistics papers, mark the description of the linguistic phenomena being covered as BACKGROUND. This includes example sentences. In contrast, the *analysis* of the phenomena are typically either own or other work.

It may be that there is a BACKGROUND segment somewhere in the middle of the paper. It may then not be easy to decide if it is BACKGROUND or OWN. Use the following test: if you think that this segment could have been used as an introductory text at the beginning of the paper, and if it does not contain material that is individualized to the authors themselves, then it should be marked as BACKGROUND.

References to "pioneers" in the field are also BACKGROUND material—sentences which describe other work in an introductory way without any criticism. These are usually older references.

Sometimes there is no BACKGROUND segment, namely if the authors start directly by describing one specific individualized approach.

OTHER

The difference between BACKGROUND and OTHER is only in degree of *specificity*.

OTHER are descriptions of other work which is described *specifically* enough to contrast the own work to it, to criticize it or to mention that it provides support for own idea. For some work to be considered specific other work, it must be clearly attributable to some other researchers, otherwise it might be too general to count as specific other work. Often such segments are started by markers of specific work, citations:

- *<REF> argues that children don't acquire grammar frames until they have a lexicon […]*
- *<REF> 's solution solves the problem of data-sparseness.*
- *<REF> 's formalism allows the treatment of coordinated structures.*
- *The bilingual dual-coding theory <REF> partially answers the above questions.*
- *<REF> introduced the notion of temporal anaphora, to account for ways in which temporal expressions depend on surrounding elements in the discourse for their semantic contribution to the discourse.*

Named solutions can also count as specificity markers for other work:

- *Similarity-based models suggest an appealing approach for dealing with data sparseness.*

The distinction between BACKGROUND and OTHER might be difficult to make. Stop marking as BACKGROUND when you reach a point where ideas, solutions, or tasks are clearly being individualized, i.e. attributed to researchers in such a way that they can get criticized. Often the breaking point looks like this: "*<General problem description> Recently, some researchers have tried to tackle this by doing <More specific description with references>*" In that case, the border is before *"Recently"*.

When authors give specific information about research, but express no stance towards that work, particularly if it happens in the beginning, they seem to imply the statements are generally accepted in the field. You might in this case decide to mark it as BACKGROUND.

OWN

Own work in the context of this paper means work presented as performed by the authors *in the given paper*, i.e. as new research. This includes a description of the own solution, results, discussion, limitations and future work.

Previous own research, i.e. research done by the authors before and published elsewhere, does *not* count as own work. Sometimes the fact that previous work is discussed is specifically marked ("*we have previously*"), sometimes it can only be inferred because there is a reference indicating the author's name. Check the reference list to make sure that the string "*et al.*" in a citation (cited paper) does not "hide" one of the authors of the current paper. Unfortunately, authors tend to talk about previous own work in much the same way as they do about the current (own) work. This might constitute a problem here. It is your job to decide if certain statements are presented as if they were the contribution of the paper. There is one exception: PhD or MSc theses do not count as published work (otherwise, some entire papers would have to be marked as other work if the paper is a short version of a PhD or MSc thesis).

Sometimes, short descriptions of own work (statements of opinion) appear within sections talking about other work (background or specific). For example, an author might describe a general problem, then individualize the present research by setting the scope within the current work ("*We will here only be interested in VP gapping as opposed to NP gapping*"), then continue describing general specific to VP gapping. These scope declarations should be considered as own work because they talk about the given work/opinions. The grammatical subject in a sentence does not always tell you whether it's own work or not. Sometimes the criticism of other work might look like own opinion (*"However, we are convinced that this is wrong [...]"*). Cases like this should *not* be considered as own work, but as a description of the weaknesses of other work, i.e. it should be marked as OTHER.

In particular, watch out for the first mention of the own work, typically two thirds down in the introduction. Most of the information under the Summary or Conclusion section is normally own work. Sometimes, individual sentences in the conclusion section make direct comparisons with other work, e.g. detailing advantages of the approach. Only mark these as OTHER if the other work is described again, using more than one sentence of description, else mark as OWN.

When it gets difficult

There are several reasons why the annotation scheme might not work well for a given paper. The writing style in some papers might make it difficult to see the trisection according to intellectual ownership. In some papers however, the scheme's assumptions that research with different ownership (own/other/background) is indeed presented in separate segments in the paper are violated:

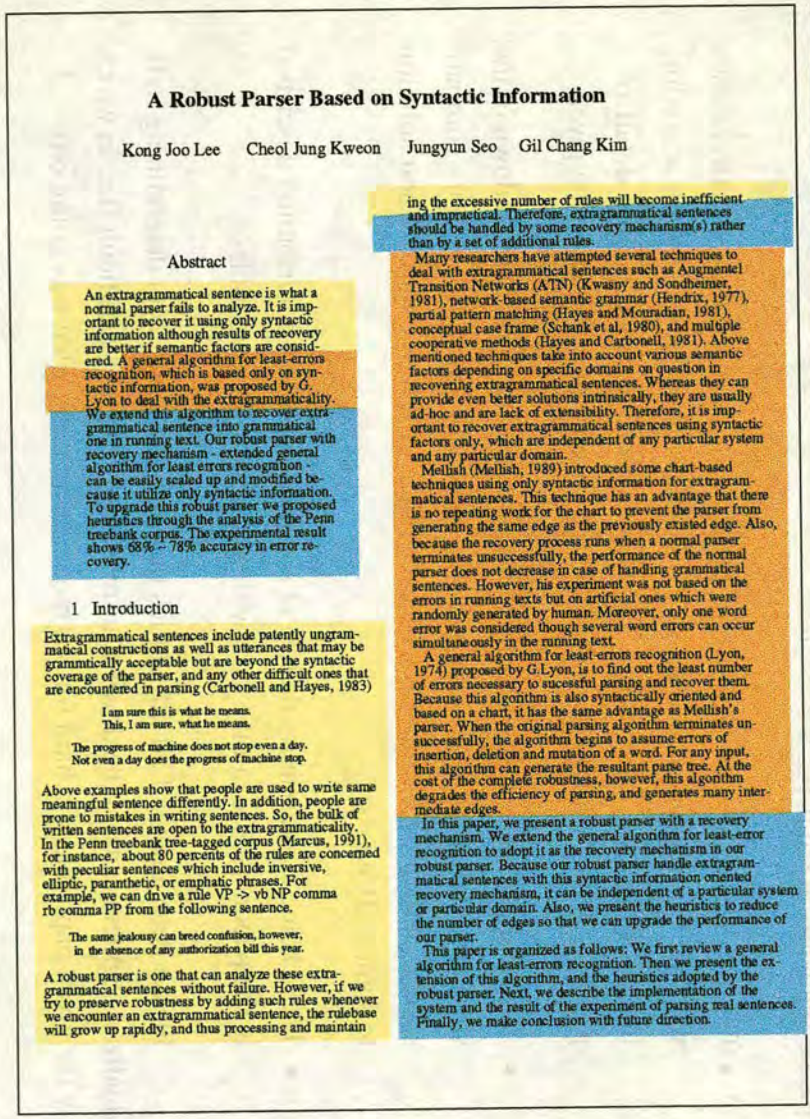
- Our model assumes that the author perceives a clear separation between own work and work outside the scope of the paper, and presents work according to that separation. However, if the paper describes some minute detail of a previous, larger work of the author, then this separation might not be given.

- A specialized case of this, and another example of a potential breakdown of the simple model is for evaluation papers, especially where the authors compare several of their own solutions with each other, or if they compare their solution to somebody else's.

- The scheme also assumes that there is *really* some new contribution described in the paper. This is not the case with position or review articles.

Please keep a note of all difficulties that you encounter with determining individualized segments, and write down your reasons for finding it difficult (i.e. in which way the given paper made it hard for our model to describe what was going on).

Example annotation



A Robust Parser Based on Syntactic Information

Kong Joo Lee Cheol Jung Kweon Jungyun Seo Gil Chang Kim

Abstract

An extragrammatical sentence is what a normal parser fails to analyze. It is important to recover it using only syntactic information although results of recovery are better if semantic factors are considered. A general algorithm for least-errors recognition, which is based only on syntactic information, was proposed by G. Lyon to deal with the extragrammaticality. We extend this algorithm to recover extragrammatical sentence into grammatical one in running text. Our robust parser with recovery mechanism - extended general algorithm for least errors recognition - can be easily scaled up and modified because it utilize only syntactic information. To upgrade this robust parser we proposed heuristics through the analysis of the Penn treebank corpus. The experimental result shows 68% ~ 78% accuracy in error recovery.

1 Introduction

Extragrammatical sentences include patently ungrammatical constructions as well as utterances that may be grammtically acceptable but are beyond the syntactic coverage of the parser, and any other difficult ones that are encountered in parsing (Carbonell and Hayes, 1983)

I am sure this is what he means.
This, I am sure, what he means.

The progress of machine does not stop even a day.
Not even a day does the progress of machine stop.

Above examples show that people are used to write same meaningful sentence differently. In addition, people are prone to mistakes in writing sentences. So, the bulk of written sentences are open to the extragrammaticality. In the Penn treebank tree-tagged corpus (Marcus, 1991), for instance, about 80 percents of the rules are concerned with peculiar sentences which include inversive, elliptic, paranthetic, or emphatic phrases. For example, we can drive a rule VP -> vb NP comma rb comma PP from the following sentence.

The same jealousy can breed confusion, however,
in the absence of any authorization bill this year.

A robust parser is one that can analyze these extragrammatical sentences without failure. However, if we try to preserve robustness by adding such rules whenever we encounter an extragrammatical sentence, the rulebase will grow up rapidly, and thus processing and maintaining the excessive number of rules will become inefficient and impractical. Therefore, extragrammatical sentences should be handled by some recovery machanism(s) rather than by a set of additional rules.

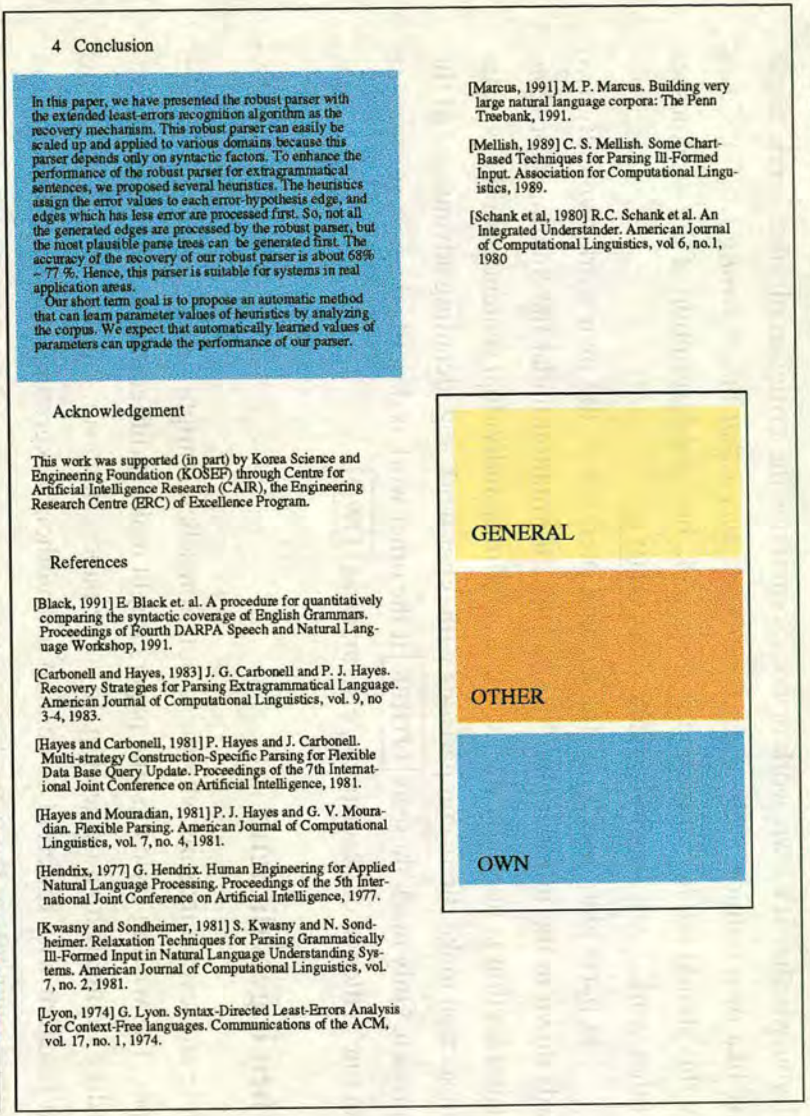
Many researchers have attempted several techniques to deal with extragrammatical sentences such as Augmentel Transition Networks (ATN) (Kwasny and Sondheimer, 1981), network-based semantic grammar (Hendrix, 1977), partial pattern matching (Hayes and Mouradian, 1981), conceptual case frame (Schank et al, 1980), and multiple cooperative methods (Hayes and Carbonell, 1981). Above mentioned techniques take into account various semantic factors depending on specific domains on question in recovering extragrammatical sentences. Whereas they can provide even better solutions intrinsically, they are usually ad-hoc and are lack of extensibility. Therefore, it is important to recover extragrammatical sentences using syntactic factors only, which are independent of any particular system and any particular domain.

Mellish (Mellish, 1989) introduced some chart-based techniques using only syntactic information for extragrammatical sentences. This technique has an advantage that there is no repeating work for the chart to prevent the parser from generating the same edge as the previously existed edge. Also, because the recovery process runs when a normal parser terminates unsuccessfully, the performance of the normal parser does not decrease in case of handling grammatical sentences. However, his experiment was not based on the errors in running texts but on artificial ones which were randomly generated by human. Moreover, only one word error was considered though several word errors can occur simultaneously in the running text.

A general algorithm for least-errors recognition (Lyon, 1974) proposed by G.Lyon, is to find out the least number of errors necessary to sucessful parsing and recover them. Because this algorithm is also syntactically oriented and based on a chart, it has the same advantage as Mellish's parser. When the original parsing algorithm terminates unsuccessfully, the algorithm begins to assume errors of insertion, deletion and mutation of a word. For any input, this algorithm can generate the resultant parse tree. At the cost of the complete robustness, however, this algorithm degrades the efficiency of parsing, and generates many intermediate edges.

In this paper, we present a robust parser with a recovery mechanism. We extend the general algorithm for least-error recognition to adopt it as the recovery mechanism in our robust parser. Because our robust parser handle extragrammatical sentences with this syntactic information oriented recovery mechanism, it can be independent of a particular system or particular domain. Also, we present the heuristics to reduce the number of edges so that we can upgrade the performance of our parser.

This paper is organized as follows: We first review a general algorithm for least-errors recognition. Then we present the extension of this algorithm, and the heuristics adopted by the robust parser. Next, we describe the implementation of the system and the result of the experiment of parsing real sentences. Finally, we make conclusion with future direction.



4 Conclusion

In this paper, we have presented the robust parser with the extended least-errors recognition algorithm as the recovery mechanism. This robust parser can easily be scaled up and applied to various domains because this parser depends only on syntactic factors. To enhance the performance of the robust parser for extragrammatical sentences, we proposed several heuristics. The heuristics assign the error values to each error-hypothesis edge, and edges which has less error are processed first. So, not all the generated edges are processed by the robust parser, but the most plausible parse trees can be generated first. The accuracy of the recovery of our robust parser is about 68% ~ 77 %. Hence, this parser is suitable for systems in real application areas.

Our short term goal is to propose an automatic method that can learn parameter values of heuristics by analyzing the corpus. We expect that automatically learned values of parameters can upgrade the performance of our parser.

Acknowledgement

This work was supported (in part) by Korea Science and Engineering Foundation (KOSEF) through Centre for Artificial Intelligence Research (CAIR), the Engineering Research Centre (ERC) of Excellence Program.

References

[Black, 1991] E. Black et. al. A procedure for quantitatively comparing the syntactic coverage of English Grammars. Proceedings of Fourth DARPA Speech and Natural Language Workshop, 1991.

[Carbonell and Hayes, 1983] J. G. Carbonell and P. J. Hayes. Recovery Strategies for Parsing Extragrammatical Language. American Journal of Computational Linguistics, vol. 9, no 3-4, 1983.

[Hayes and Carbonell, 1981] P. Hayes and J. Carbonell. Multi-strategy Construction-Specific Parsing for Flexible Data Base Query Update. Proceedings of the 7th International Joint Conference on Artificial Intelligence, 1981.

[Hayes and Mouradian, 1981] P. J. Hayes and G. V. Mouradian. Flexible Parsing. American Journal of Computational Linguistics, vol. 7, no. 4, 1981.

[Hendrix, 1977] G. Hendrix. Human Engineering for Applied Natural Language Processing. Proceedings of the 5th International Joint Conference on Artificial Intelligence, 1977.

[Kwasny and Sondheimer, 1981] S. Kwasny and N. Sondheimer. Relaxation Techniques for Parsing Grammatically Ill-Formed Input in Natural Language Understanding Systems. American Journal of Computational Linguistics, vol. 7, no. 2, 1981.

[Lyon, 1974] G. Lyon. Syntax-Directed Least-Errors Analysis for Context-Free languages. Communications of the ACM, vol. 17, no. 1, 1974.

[Marcus, 1991] M. P. Marcus. Building very large natural language corpora: The Penn Treebank, 1991.

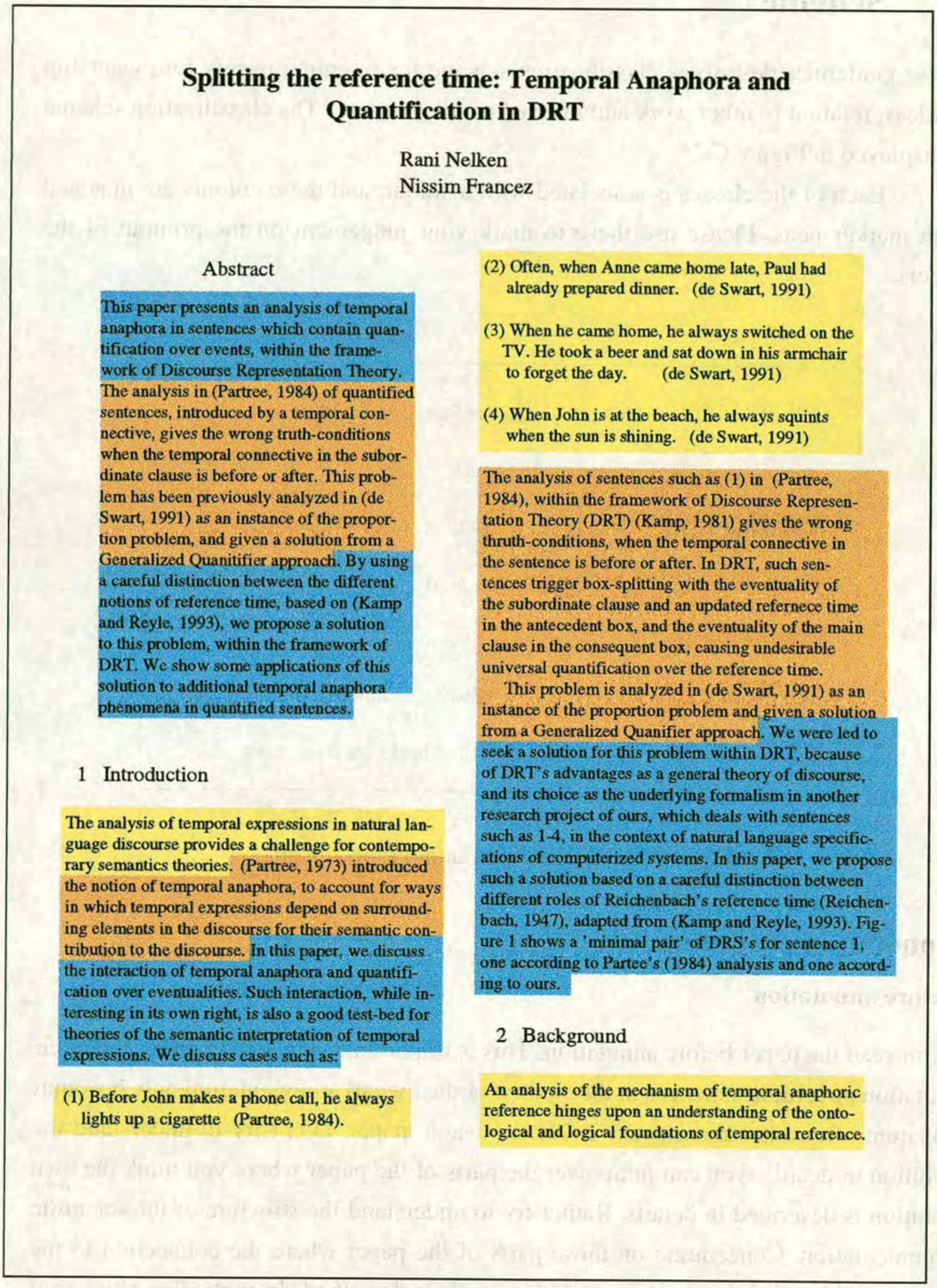
[Mellish, 1989] C. S. Mellish. Some Chart-Based Techniques for Parsing Ill-Formed Input. Association for Computational Linguistics, 1989.

[Schank et al, 1980] R.C. Schank et al. An Integrated Understander. American Journal of Computational Linguistics, vol 6, no.1, 1980

GENERAL

OTHER

OWN



Splitting the reference time: Temporal Anaphora and Quantification in DRT

Rani Nelken
Nissim Francez

Abstract

This paper presents an analysis of temporal anaphora in sentences which contain quantification over events, within the framework of Discourse Representation Theory. The analysis in (Partree, 1984) of quantified sentences, introduced by a temporal connective, gives the wrong truth-conditions when the temporal connective in the subordinate clause is before or after. This problem has been previously analyzed in (de Swart, 1991) as an instance of the proportion problem, and given a solution from a Generalized Quanitifier approach. By using a careful distinction between the different notions of reference time, based on (Kamp and Reyle, 1993), we propose a solution to this problem, within the framework of DRT. We show some applications of this solution to additional temporal anaphora phenomena in quantified sentences.

1 Introduction

The analysis of temporal expressions in natural language discourse provides a challenge for contemporary semantics theories. (Partree, 1973) introduced the notion of temporal anaphora, to account for ways in which temporal expressions depend on surrounding elements in the discourse for their semantic contribution to the discourse. In this paper, we discuss the interaction of temporal anaphora and quantification over eventualities. Such interaction, while interesting in its own right, is also a good test-bed for theories of the semantic interpretation of temporal expressions. We discuss cases such as:

(1) Before John makes a phone call, he always lights up a cigarette (Partree, 1984).

(2) Often, when Anne came home late, Paul had already prepared dinner. (de Swart, 1991)

(3) When he came home, he always switched on the TV. He took a beer and sat down in his armchair to forget the day. (de Swart, 1991)

(4) When John is at the beach, he always squints when the sun is shining. (de Swart, 1991)

The analysis of sentences such as (1) in (Partree, 1984), within the framework of Discourse Representation Theory (DRT) (Kamp, 1981) gives the wrong thruth-conditions, when the temporal connective in the sentence is before or after. In DRT, such sentences trigger box-splitting with the eventuality of the subordinate clause and an updated refernece time in the antecedent box, and the eventuality of the main clause in the consequent box, causing undesirable universal quantification over the reference time.

This problem is analyzed in (de Swart, 1991) as an instance of the proportion problem and given a solution from a Generalized Quanifier approach. We were led to seek a solution for this problem within DRT, because of DRT's advantages as a general theory of discourse, and its choice as the underlying formalism in another research project of ours, which deals with sentences such as 1-4, in the context of natural language specifications of computerized systems. In this paper, we propose such a solution based on a careful distinction between different roles of Reichenbach's reference time (Reichenbach, 1947), adapted from (Kamp and Reyle, 1993). Figure 1 shows a 'minimal pair' of DRS's for sentence 1, one according to Partee's (1984) analysis and one according to ours.

2 Background

An analysis of the mechanism of temporal anaphoric reference hinges upon an understanding of the ontological and logical foundations of temporal reference.

C.2. Study II: Guidelines for Human Annotation of Full Scheme

These guidelines describe a classification scheme for scientific papers for ownership of ideas, relation to other work and internal paper structure. The classification scheme is displayed in Figure C.2.

Each of the classes is associated with a colour, and these colours are matched with marker pens. Please use these to mark your judgement on the printout of the papers.

BACKGROUND	Generally accepted background knowledge
OTHER	Specific other work
OWN	Own work: method, results, future work...
AIM	Specific research goal
TEXTUAL	Textual section structure
CONTRAST	Contrast, comparison, weakness of other solution
BASIS	Other work provides basis for own work

Figure C.2: Overview of annotation scheme

Annotation procedure

Before annotation

Skim-read the paper before annotation. This is important, as in some papers, the interpretation of certain sentences in the context of the overall argumentation only becomes apparent after one has an overview of the whole paper. Don't try to understand the solution in detail—you can jump over the parts of the paper where you think the own solution is described in details. Rather try to understand the structure of the scientific argumentation. Concentrate on those parts of the paper where the connection to the subject field and the connection to other work is described. In particular, skim-read the abstract, the introduction, the conclusions (if it is summary-style), and sections reviewing other research (often after introduction or before conclusions; they could be marked sections with headlines like "Relation to other work", "Prior research", "X in the literature" etc.).

Annotation procedure

Annotation proceeds sentence by sentence, and is mutually exclusive: Each sentence can have only one category. The main decision procedure is given in Figure C.3. For each sentence, the following questions have to be answered.

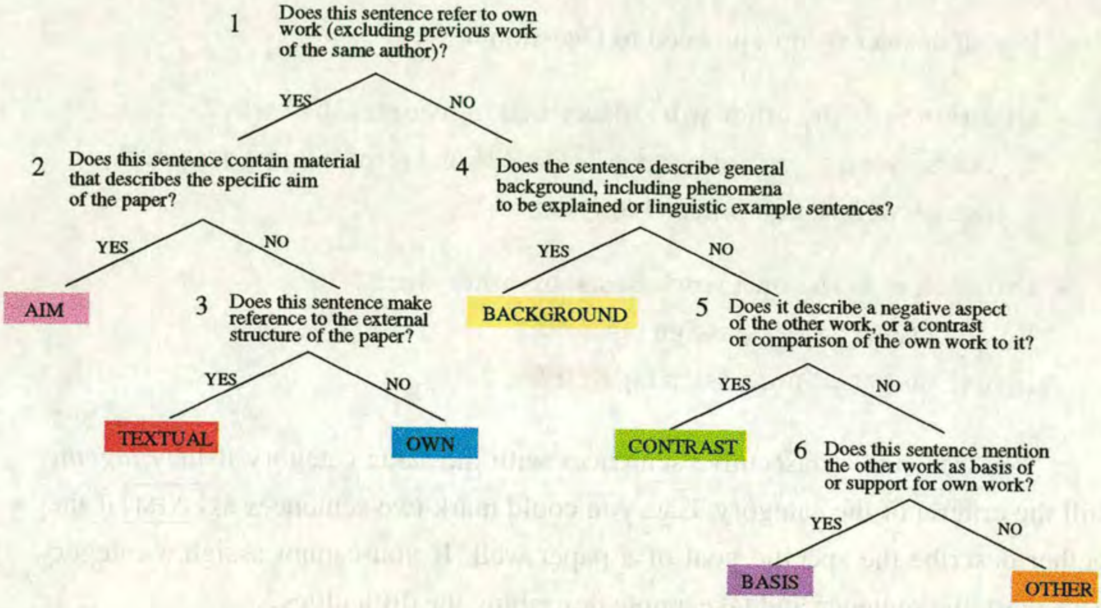

1 Does this sentence refer to own work (excluding previous work of the same author)?
YES
NO
2 Does this sentence contain material that describes the specific aim of the paper?
4 Does the sentence describe general background, including phenomena to be explained or linguistic example sentences?
YES
NO
AIM
3 Does this sentence make reference to the external structure of the paper?
YES
NO
BACKGROUND
5 Does it describe a negative aspect of the other work, or a contrast or comparison of the own work to it?
YES
NO
TEXTUAL
OWN
YES
NO
CONTRAST
6 Does this sentence mention the other work as basis of or support for own work?
YES
NO
BASIS
OTHER

Figure C.3: Decision process

Therefore, if there is a conflict, the "higher" classes in the decision tree (the ones that you reach first) will win over the "lower" classes. These guidelines will give details about the questions.

When interpreting the role of a sentence, you should treat the sentence in the way in which you think the *author* intended it in their argumentation. Context and location of a sentence are important.

- **Question 1: Does this sentence talk about own work?**
  If your answer is 'yes', proceed to Question 2.
  If your answer is 'no', proceed to Question 4.

- **Question 2: Does it contain a goal statement?**
  If your answer is 'yes', assign class AIM and move to next sentence.
  If your answer is 'no', proceed to Question 3.

- **Question 3: Does it contain a textual overview?**
  If your answer is 'yes', assign tag TEXTUAL and move to the next sentence.
  If your answer is 'no', assign tag OWN and move to the next sentence.

- **Question 4: Does it describe background?**
  If your answer is 'yes', assign tag BACKGROUND and move to the next sentence.
  If your answer is 'no', proceed to Question 6.

- **Question 5: Is the other work described in a contrastive way?**
  If your answer is 'yes', assign tag CONTRAST and move to next sentence.
  If your answer is 'no', proceed to Question 5.

- **Question 6: Is the own work based on other work?**
  If your answer is 'yes', assign tag BASIS.
  If your answer is 'no', assign tag OTHER.

You can mark consecutive sentences with the same category if they *together* fulfill the criteria of the category. E.g. you could mark two sentences as AIM if they together describe the specific goal of a paper well. If you cannot assign a category, please mark the sentence and take a note describing the difficulties.

As soon as you have reached a leaf, assign the corresponding category to the sentence. Please annotate all sentences in the abstract, and all sentences in the document except acknowledgement sentences. Also mark (linguistic) example sentences.

After annotation

Check a few things, and rectify your annotation if necessary:

- There must be at least one AIM sentence. If this is not the case, reclassify some other candidate sentences, until you have found at least one sentence that represents the specific aim of the given paper.

- There must not be more than 5 AIM sentences per paper. The only exception is if each of them is a straight hit, i.e. they are indisputably goal statements, particularly if the sentences are paraphrases of each other.

  If you have to eliminate AIM sentences, do the following:

  - Prefer explicit AIM statements (prefer 'direct' goal statements and 'functionality-provided' to 'solved' and other types).
  - Prefer AIM sentences towards the periphery (e.g. at the beginning of summarizing conclusions), and in the border area with OTHER or Background segments;
  - If all fails, pick the ones you think are most relevant in the context of distinguishing this piece of research from others.

The questions

Question 1: Does this sentence talk about own work?

Own work in the context of this paper means work presented as performed by the authors *in the given paper*, i.e. as new research.

Description of own work should make up a large part of the paper—it includes descriptions of the own solution, method, results, discussion, limitations and future work.

Previous own research, i.e. research done by the authors before and published elsewhere, does *not* count as own work. Sometimes the fact that previous work is discussed is specifically marked ("*we have previously*"), sometimes it can only be inferred because there is a reference indicating the author's name. Check the reference list to make sure that the string "*et al.*" in a citation (cited paper) does not "hide" one of the authors of the current paper. Unfortunately, authors tend to talk about previous own work in much the same way as they do about the current (own) work. This might constitute a problem here. It is your job to decide if certain statements are presented as if they were the contribution of the paper. There is one exception: PhD or MSc theses do not count as published work (otherwise, some entire papers would have to be marked as other work if the paper is a short version of a PhD or MSc thesis). In that case, the sentence first citing the thesis is to be marked as BASIS. In all other contexts, reference to the thesis/research is to be considered as own.

Sometimes, short descriptions of own work (statements of opinion) appear within sections talking about other work (background or specific). For example, an author might describe a general problem, then individualize the present research by setting the scope within the current work ("*We will here only be interested in VP gapping as opposed to NP gapping*"), then continue describing general specific to VP gapping. These scope declarations should be considered as own work because they talk about the given work/opinions. The grammatical subject in a sentence does not always tell you whether it's own work or not. Sometimes the criticism of other work might look like own opinion ("*However, we are convinced that this is wrong [...]*"). Cases like this should *not* be considered as own work, but as weaknesses of other work, i.e. OTHER.

In particular, watch out for the first mention of the own work, typically two thirds down in the introduction. Most of the information under the Summary or Conclusion section is normally own work. Sometimes, individual sentences in the conclusion section make direct comparisons with other work, e.g. detailing advantages of the approach. Only mark these as OTHER if the other work is described again, using more than one sentence of description, else mark as OWN.

Question 2: Does this sentence contain a goal statement?

Two kinds of sentences count as goal statements:

- Goal statements (i.e. description of research goal)
- Scope statement (i.e. delimitation of research goal: what the goal is not)

If the sentence describes a general goal in the field, e.g. "*machine translation*", it should not be marked as AIM. AIM sentences describe *particular* goals of the paper. There are different ways of expressing the particular goal of the paper.

A prime location of AIM sentences is around the first 2/3 of the introduction, when the authors are mentioned for the first time.

Direct aim/goal description:

- *Our aim in this paper is to [...]*
- *We, in contrast, aim at defining categories that help us [...]*

Also descriptions of phenomena plus the statement that current work tries to explain them, e.g.:

- *We aim to find a method of inducing grammar rules.*
- *Our goal, however, is to develop a mechanism for [...]*
- *We will introduce PHENOMENON X that we seek to explain*
- *I show how grammar rules can be induced.*

Functionality provided: Another way of expressing the research goal is to say that one has accomplished doing a certain task.

- *This paper gives a syntactic-head-driven generation algorithm which includes a well-defined treatment of moved constituents.*
- *We have presented an analysis of the data sparseness problem*
- *I have presented an analysis of PHENOMENON X*
- *We have presented an analysis of why children cannot [...]* (PHENOMENON)

Hypothesis: In experimental papers the goal might be expressed as a hypothesis:

- *The hypothesis investigated in this paper is that children can acquire [...]*

Goal as focus: The declaration of a research interest can count as an $\boxed{\text{AIM}}$:

- *This paper focuses on inducing grammar rules.*
- *This paper concerns the formal definitions underlying synchronous tree-adjoining grammars.*
- *In this paper, we focus on the application of the developed techniques in the context of the comparatively neglected area of HPSG generation.*
- *This paper will focus on [...] our analysis of narrative progression, rhetorical structure, perfects and temporal expressions.*

Solutionhood: Sometimes a sentence states that the own solution works, i.e. solves a particular research task. Such sentences can under certain circumstances be AIMs, but they are AIMs of a lower quality. You must be sure that the announcement of the successful problem-solving process is indeed important enough to cover the goal of the whole paper, and you must be sure that the sentence refers to the *highest* level of problem solving. If it talks about a *sub*problem, don't consider the sentence an AIM. Often such statements are dressed as a claim.

Examples:

- *[we present an analysis] which automatically gives the right results for quantifier scope ambiguities and interactions with bound anaphora.*
- *In this paper we presented a new model that implements the similarity-based approach to provide estimates for the conditional probabilities of unseen word cooccurrences*
- *Our technique segments continuous speech into words using only distributional and phonotactic information*
- *The Spoken Language Translator (SLT) is a prototype system that translates air travel (ATIS) queries from spoken English to spoken Swedish and to French.*

Definition of a desired property or as necessity: The goal can be given by describing a hypothetical, desired mechanism or a desired outcome. This is not a typical way to describe the paper's AIM, but the context can still make this the "best AIM around".

Examples:

- *A robust Natural Language Processing (NLP) system must be able to process sentences that contain words unknown to its lexicon.*
- *The importance of a method for SPECIFIC-TASK grows as the coverage of [...] improves.*
- *and I demonstrate the importance of having a Y tool which allows for X.*

Advantage of a solution: Sometimes the description of an advantage of a solution can provide an acceptable AIM:

- *Our method yields polynomial complexity in an elegant way.*
- *Our method avoids problems of non-determinacy.*

- *First, it is in certain respects simpler, in that it requires no postulation of otherwise unmotivated ambiguities in the source clause.*
- *The traditional problems of training times do not arise.*

Scope statement: These sentences define the goal as *part* of previous goal, e.g. *"here we will look only at relative pronouns"*, excluding some other, similar goals.

Indirect aim/goal description: In some cases, if you find nothing better, you can also look for more indirect ways of expressing what the goal might have been.

- *In this paper we address two issues relating to the application of preference functions.*
- *[...] and make a specific proposal concerning the interface between these and the syntactic and semantic representations they utilize.*
- *In addition, we have taken a few steps towards determining the relative importance of different factors to the successful operation of discourse modules.*

Question 3: Does this sentence contain a textual overview?

All statements whose primary function it is to give us an overview of the section structure (*"in the next section we will [...]"*). Several such sentences often occur at the end of the introduction.

Mark also backward looking pointers at the beginning of a section (first sentence) (*"In the previous section we have implemented a model"*) or before the end of the section ("*in the next section, we will turn our attention to [...]* ". Some authors give an overview of the section at the beginning of the section ("*in this section I will [dots]*"), or summarize after each section ("*in this section I have [dots]*" or "*this concludes my discussion of X"*.

Caveat: Sentences referring to figures or tables are not meant here (*"figure 3 shows [...]*")!

Sentences summing up main conclusions from *previous* sections are also not meant here:

- *"In chapter 3, we have seen that children cannot reliably form generalizations about [...]".*

Question 4: Does this sentence describe background?

BACKGROUND knowledge marks sentences which are presented as uncontroversial in the field. In such sentences, the research context is established. This includes statements of general capacity of the field, general problems, research goals, methodologies and general solutions ("*In recent years, there has been a growing interest in the field of X in the subject of Y*"). The most prototypical use of BACKGROUND is in the beginning of the paper.

Examples for general problems:

- *One of the difficult problems in machine translation from Japanese to English or other European languages is the treatment of articles and numbers.*
- *Complications arise in spelling rule application from the fact that, at compile time, neither the lexical nor the surface form of the root, nor even its length, is known.*
- *Collocations present specific problems in translation, both in human and automatic contexts.*

Examples for generally accepted/old solutions or claims:

- *Tagging by means of a Hidden Markov Model (HMM) is widely recognised as an effective technique for assigning parts of speech to a corpus in a robust and efficient manner.*
- *Current research in lexical aquisition is eminently knowledge-based.*
- *Literature in psychology has amply demonstrated that children do not acquire [. . .]*

In linguistics papers, mark the description of the linguistic phenomena being covered as BACKGROUND. This includes example sentences. In contrast, the *analysis* of the phenomena are typically either own or other work.

It may be that there is a BACKGROUND segment somewhere in the middle of the paper. It may then not be easy to decide if it is BACKGROUND or OWN. Use the following test: if you think that this segment could have been used as an introductory text at the beginning of the paper, and if it does not contain material that is individualized to the authors themselves, then it should be marked as BACKGROUND.

References to "pioneers" in the field are also BACKGROUND material—sentences which describe other work in an introductory way without any criticism. These are usually older references.

Sometimes there is no BACKGROUND segment, namely if the authors start directly by describing one specific individualized approach.

The difference between BACKGROUND and OTHER is only in degree of *specificity*.

OTHER are descriptions of other work which is described *specifically* enough to contrast the own work to it, to criticize it or to mention that it provides support for own idea. For some work to be considered specific other work, it must be clearly attributable to some other researchers, otherwise it might be too general to count as specific other work. Often such segments are started by markers of specific work, citations:

- *<REF> argues that children don't acquire grammar frames until they have a lexicon [...]*
- *<REF> 's solution solves the problem of data-sparseness.*
- *<REF> 's formalism allows the treatment of coördinated structures.*
- *The bilingual dual-coding theory <REF> partially answers the above questions.*
- *<REF> introduced the notion of temporal anaphora, to account for ways in which temporal expressions depend on surrounding elements in the discourse for their semantic contribution to the discourse.*

Named solutions can also count as specificity markers for other work:

- *Similarity-based models suggest an appealing approach for dealing with data sparseness.*

The distinction between BACKGROUND and OTHER might be difficult to make. Stop marking as BACKGROUND when you reach a point where ideas, solutions, or tasks are clearly being individualized, i.e. attributed to researchers in such a way that they can get criticized. Often the breaking point looks like this: "*<General problem description> Recently, some researchers have tried to tackle this by doing <More specific description with references>*" In that case, the border is before *"Recently"*.

When authors give specific information about research, but express no stance towards that work, particularly if it happens in the beginning, they seem to imply the statements are generally accepted in the field. You might in this case decide to mark it as BACKGROUND.

Question 5: Is the other work described in a contrastive way?

These sentences make one type of connection between specific other work and own work. Comparative sentences might occur within segments describing other work or own work (e.g. in conclusions).

Mark sentences which contain mentions of:

- Weaknesses of other people's solutions
- The absence of a solution for a given problem
- Difference in approach/solution
- Superiority of own solution
- Statements of direct comparisons with other work or between several other approaches (these appear mostly in evaluation papers)
- Incompatibility between own and other claims or results

Weaknesses of other solutions:

- *<REF>'s solution is problematic for several reasons.*
- *The results suggest that a completely unconstrained initial model does not produce good quality results.*
- *Here, we will produce experimental evidence suggesting that this simple model leads to serious overestimates of system error rates.*
- *The analysis of sentences such as <CREF> in <REF>, within the framework of Discourse Representation Theory (DRT) <REF> gives the wrong truth-conditions, when the temporal connective in the sentence is "before" or "after".*
- *A limiting factor of this method is the potentially large number of distinct parse trees.*

Absence of a solution:

- *While we know of previous work which associates scores with feature structures <REF> we are not aware of any previous treatment which makes explicit the link to classical probability theory.*
- *First, although much work has been done on how agents request clarifications, or respond to such requests, little attention has been paid to the collaborative aspects of clarification discourse.*

Difference in approach/solution:

- *In contrast to standard approaches, we use a statistical model.*
- *In this paper, we propose an alternative approach in which a performance-oriented (behaviour-based) perspective is taken instead of a competence-oriented (knowledge-based) one.*
- *Namely, since we use semantic/pragmatic roles instead of grammatical roles in constraints [...]*

Superiority of own solution:

- *Our model outperforms simple pattern-matching models by 25%.*
- *Our results indicate that our full integrated heuristic scheme for selecting the best parse out-performs the simple heuristic [...]*
- *We have also argued that an architecture that uses obligations provides a much simpler implementation than the strong plan-based approaches.*

Direct comparisons with other work:

- *In this paper, we will compare two tagging algorithms, one based on classifying word types, and one based on classifying words-plus-context.*
- *[...] and a comparison with manual scaling in section <CREF>.*
- *The performance of both implementations is evaluated and compared on a range of artificial and real data.*

Incompatibility between own and other claims or results:

- *This result challenges the claims of recent discourse theories (<REF>, <REF>) which argue for a the close relation between cue words and discourse structure.*
- *It is implausible that children learn grammar on the fly.*

There is a conflict between AIM and CONTRAST when goals are introduced contrastively, as in the following examples. These sentences would normally be tagged AIM, unless there are too many better AIM sentences around.

- *Until now, research has focused on demonstrations of infants' sensitivity to various sources; we have begun to provide quantitative measures of the usefulness of those sources.*
- *However our objective is not to propose a faster algorithm, but is to show the possibility of distributed processing of natural languages.*
- *This article proposes a method for automatically finding the appropriate tree-cutting criteria in the EBG scheme, rather than having to hand-code them.*

If the sentence expresses no sentential content other than the fact that there is a contrast ("*however, our approach is quite different*") mark this sentence only as CONTRAST if you don't find a better one.

If authors compare their own work contrastively to somebody else's (e.g. a linguistic analysis) to explain in which aspects their own work is superior, you might be undecided as to whether to mark it as CONTRAST or OWN (or even AIM, in some cases!). Assign AIM only if the authors specifically say that they did something differently in order to achieve a (different?) goal. Assign CONTRAST if you believe that the main function of the sentence is to mention a negative aspect of the other work. Assign OWN if the focus is on their own work rather than on the other work.

Question 6: Is the own work based on other work?

There are 5 different classes of how work could be based or positively related:

- Direct Based
- Adaptation

- Consistency
- Similarity
- Quality

Consistency, Similarity and Quality cases should be marked only if the approaches are important to the paper, i.e. if some more discussion about that work is given in the paper.

Direct Based: It is explicitly stated that the own solution builds on another solution (intellectual ancestry).

- *We base our model on <REF>'s backup model.*
- *Our approach is in the spirit of <REF> 's approach*
- *We choose to use Link Grammar <REF>*

The last example describes a BASIS describing intellectual ancestry with more than one other approach.

Adaptation: The authors have adapted a solution, contributed by somebody else. As the solution was not initially invented for the current research task, and needs to be adapted.

- *The main aim is to show how existing text planning techniques can be adapted for this particular application.*
- *We extend the model for doing X by allowing it to do Y, too.*
- *We have suggested some ways in which LFs can be enriched with lexical semantic information to improve translation quality.*
- *This model draws upon <REF>, but adapts it to the collaborative situation.*
- *In our work, we have taken <REF>'s descriptive model and recast it into a computational one [...]*

Consistency: Statements about consistency with another theoretical framework or other people's results can be BASIS, even if the own solution is not directly based on it:

- *Our account [...] fits within a general framework for [...]*

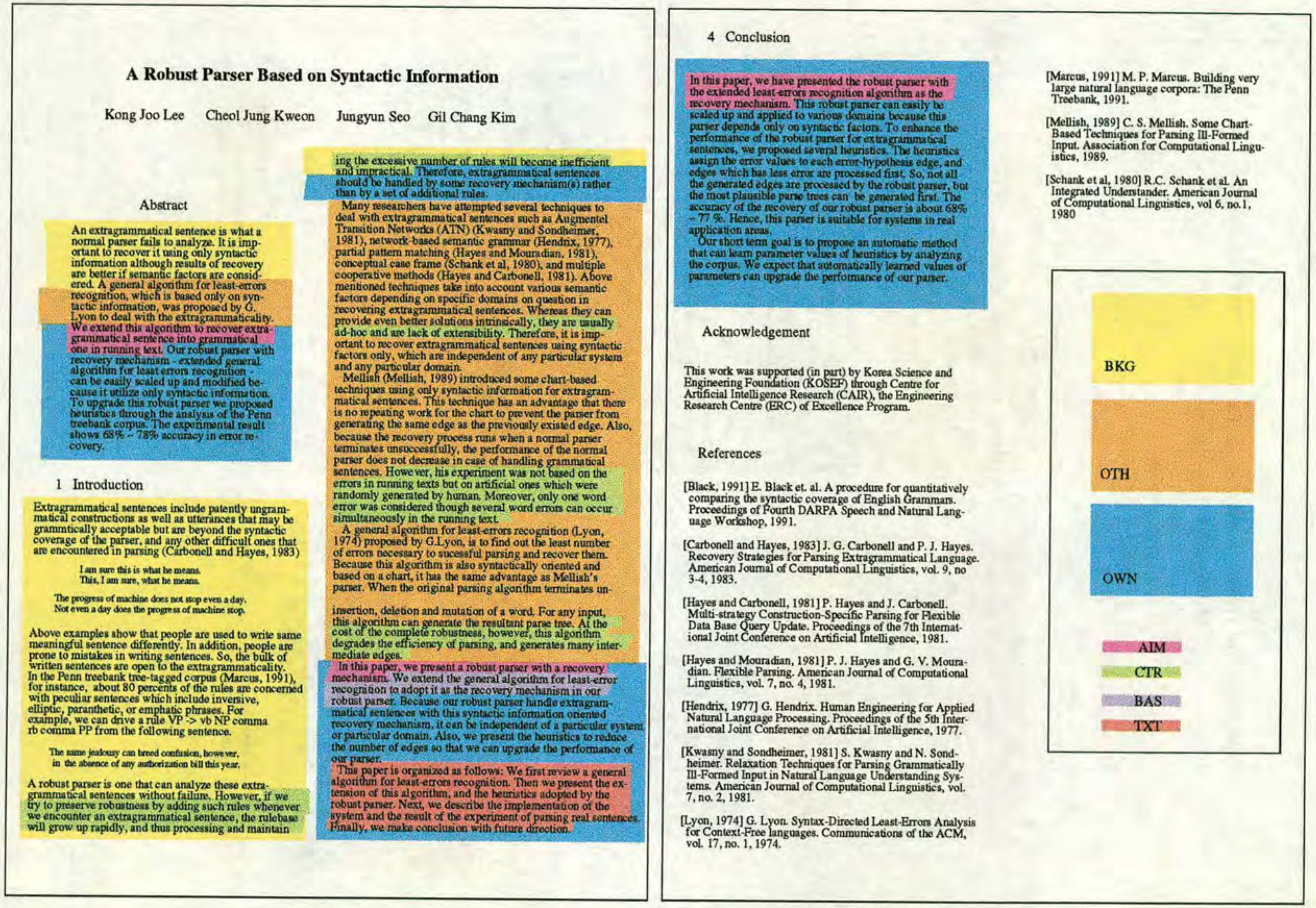
Similarity: Statements about similarities between the own and other approaches can be a BASIS, if these similarities are not "cancelled" later by mentioning a contrasting property.

- *The analysis presented here has strong similarities to analyses of the same phenomena discussed by <REF> and <REF>.*
- *The method, which is related to that of <REF>,*
- *In this section we define a grammar similar to <REF>'s first grammar.*

Quality of other approach: If you think that an approach provides a basis, and is important enough to be marked up as a BASIS, but you can find no explicit sentence expressing it, you can mark up statements about the quality of the approach.

- *We discuss the advantages of <REF>'s model.*
- *[...] the success of an abstract model such as <REF>'s [...]*
- *[...] thus demonstrating the computational feasibility of their work and its compatibility with current practices in artificial intelligence.*
- *Earley deduction is a very attractive framework for natural language processing because it has the following properties and applications.*

Example annotation



BKG
OTH
OWN
AIM
CTR
BAS
TXT

A Robust Parser Based on Syntactic Information

Kong Joo Lee   Cheol Jung Kweon   Jungyun Seo   Gil Chang Kim

Abstract

An extragrammatical sentence is what a normal parser fails to analyze. It is important to recover it using only syntactic information although results of recovery are better if semantic factors are considered. A general algorithm for least-errors recognition, which is based only on syntactic information, was proposed by G. Lyon to deal with the extragrammaticality. We extend this algorithm to recover extragrammatical sentence into grammatical one in running text. Our robust parser with recovery mechanism - extended general algorithm for least errors recognition - can be easily scaled up and modified because it utilize only syntactic information. To upgrade this robust parser we proposed heuristics through the analysis of the Penn treebank corpus. The experimental result shows 68% ~ 78% accuracy in error recovery.

1 Introduction

Extragrammatical sentences include patently ungrammatical constructions as well as utterances that may be grammtically acceptable but are beyond the syntactic coverage of the parser, and any other difficult ones that are encountered in parsing (Carbonell and Hayes, 1983)

I am sure this is what he means.
This, I am sure, what he means.

The progress of machine does not stop even a day.
Not even a day does the progress of machine stop.

Above examples show that people are used to write same meaningful sentence differently. In addition, people are prone to mistakes in writing sentences. So, the bulk of written sentences are open to the extragrammaticality. In the Penn treebank tree-tagged corpus (Marcus, 1991), for instance, about 80 percents of the rules are concerned with peculiar sentences which include inversive, elliptic, paranthetic, or emphatic phrases. For example, we can drive a rule VP -> vb NP comma rb comma PP from the following sentence.

The same jealousy can breed confusion, however,
in the absence of any authorization bill this year.

A robust parser is one that can analyze these extragrammatical sentences without failure. However, if we try to preserve robustness by adding such rules whenever we encounter an extragrammatical sentence, the rulebase will grow up rapidly, and thus processing and maintaining the excessive number of rules will become inefficient and impractical. Therefore, extragrammatical sentences should be handled by some recovery mechanism(s) rather than by a set of additional rules.

Many researchers have attempted several techniques to deal with extragrammatical sentences such as Augmentel Transition Networks (ATN) (Kwasny and Sondheimer, 1981), network-based semantic grammar (Hendrix, 1977), partial pattern matching (Hayes and Mouradian, 1981), conceptual case frame (Schank et al, 1980), and multiple cooperative methods (Hayes and Carbonell, 1981). Above mentioned techniques take into account various semantic factors depending on specific domains on question in recovering extragrammatical sentences. Whereas they can provide even better solutions intrinsically, they are usually ad-hoc and are lack of extensibility. Therefore, it is important to recover extragrammatical sentences using syntactic factors only, which are independent of any particular system and any particular domain.

Mellish (Mellish, 1989) introduced some chart-based techniques using only syntactic information for extragrammatical sentences. This technique has an advantage that there is no repeating work for the chart to prevent the parser from generating the same edge as the previously existed edge. Also, because the recovery process runs when a normal parser terminates unsuccessfully, the performance of the normal parser does not decrease in case of handling grammatical sentences. However, his experiment was not based on the errors in running texts but on artificial ones which were randomly generated by human. Moreover, only one word error was considered though several word errors can occur simultaneously in the running text.

A general algorithm for least-errors recognition (Lyon, 1974) proposed by G.Lyon, is to find out the least number of errors necessary to sucessful parsing and recover them. Because this algorithm is also syntactically oriented and based on a chart, it has the same advantage as Mellish's parser. When the original parsing algorithm terminates un-

insertion, deletion and mutation of a word. For any input, this algorithm can generate the resultant parse tree. At the cost of the complete robustness, however, this algorithm degrades the efficiency of parsing, and generates many intermediate edges.

In this paper, we present a robust parser with a recovery mechanism. We extend the general algorithm for least-error recognition to adopt it as the recovery mechanism in our robust parser. Because our robust parser handle extragrammatical sentences with this syntactic information oriented recovery mechanism, it can be independent of a particular system or particular domain. Also, we present the heuristics to reduce the number of edges so that we can upgrade the performance of our parser.

This paper is organized as follows: We first review a general algorithm for least-errors recognition. Then we present the extension of this algorithm, and the heuristics adopted by the robust parser. Next, we describe the implementation of the system and the result of the experiment of parsing real sentences. Finally, we make conclusion with future direction.

4 Conclusion

In this paper, we have presented the robust parser with the extended least-errors recognition algorithm as the recovery mechanism. This robust parser can easily be scaled up and applied to various domains because this parser depends only on syntactic factors. To enhance the performance of the robust parser for extragrammatical sentences, we proposed several heuristics. The heuristics assign the error values to each error-hypothesis edge, and edges which has less error are processed first. So, not all the generated edges are processed by the robust parser, but the most plausible parse trees can be generated first. The accuracy of the recovery of our robust parser is about 68% ~ 77 %. Hence, this parser is suitable for systems in real application areas.

Our short term goal is to propose an automatic method that can learn parameter values of heuristics by analyzing the corpus. We expect that automatically learned values of parameters can upgrade the perfomance of our parser.

Acknowledgement

This work was supported (in part) by Korea Science and Engineering Foundation (KOSEF) through Centre for Artificial Intelligence Research (CAIR), the Engineering Research Centre (ERC) of Excellence Program.

References

[Black, 1991] E. Black et. al. A procedure for quantitatively comparing the syntactic coverage of English Grammars. Proceedings of Fourth DARPA Speech and Natural Language Workshop, 1991.

[Carbonell and Hayes, 1983] J. G. Carbonell and P. J. Hayes. Recovery Strategies for Parsing Extragrammatical Language. American Journal of Computational Linguistics, vol. 9, no 3-4, 1983.

[Hayes and Carbonell, 1981] P. Hayes and J. Carbonell. Multi-strategy Construction-Specific Parsing for Flexible Data Base Query Update. Proceedings of the 7th International Joint Conference on Artificial Intelligence, 1981.

[Hayes and Mouradian, 1981] P. J. Hayes and G. V. Mouradian. Flexible Parsing. American Journal of Computational Linguistics, vol. 7, no. 4, 1981.

[Hendrix, 1977] G. Hendrix. Human Engineering for Applied Natural Language Processing. Proceedings of the 5th International Joint Conference on Artificial Intelligence, 1977.

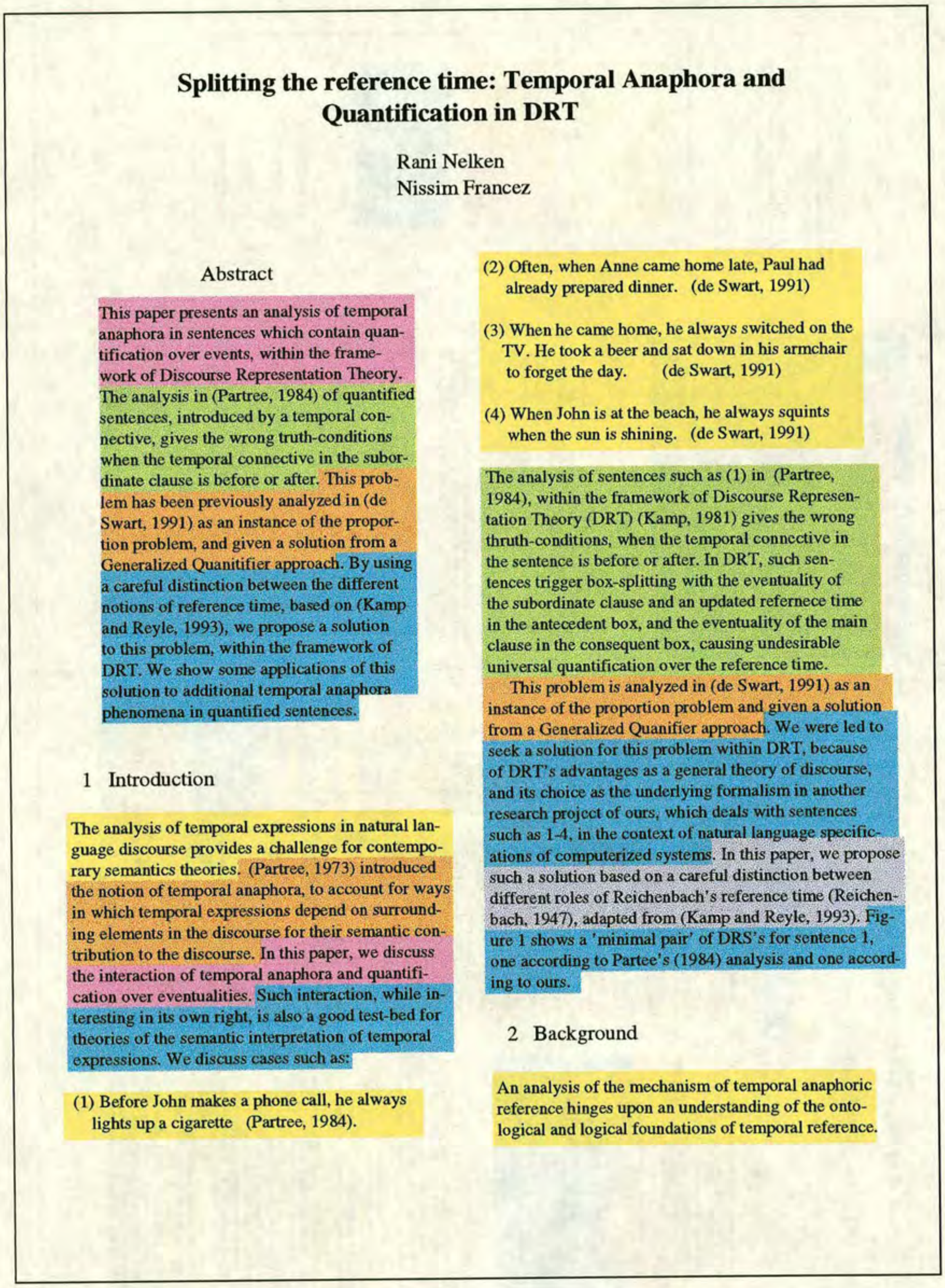
[Kwasny and Sondheimer, 1981] S. Kwasny and N. Sondheimer. Relaxation Techniques for Parsing Grammatically Ill-Formed Input in Natural Language Understanding Systems. American Journal of Computational Linguistics, vol. 7, no. 2, 1981.

[Lyon, 1974] G. Lyon. Syntax-Directed Least-Errors Analysis for Context-Free languages. Communications of the ACM, vol. 17, no. 1, 1974.

[Marcus, 1991] M. P. Marcus. Building very large natural language corpora: The Penn Treebank, 1991.

[Mellish, 1989] C. S. Mellish. Some Chart-Based Techniques for Parsing Ill-Formed Input. Association for Computational Linguistics, 1989.

[Schank et al, 1980] R.C. Schank et al. An Integrated Understander. American Journal of Computational Linguistics, vol 6, no.1, 1980



Splitting the reference time: Temporal Anaphora and Quantification in DRT

Rani Nelken
Nissim Francez

Abstract

This paper presents an analysis of temporal anaphora in sentences which contain quantification over events, within the framework of Discourse Representation Theory. The analysis in (Partree, 1984) of quantified sentences, introduced by a temporal connective, gives the wrong truth-conditions when the temporal connective in the subordinate clause is before or after. This problem has been previously analyzed in (de Swart, 1991) as an instance of the proportion problem, and given a solution from a Generalized Quanitifier approach. By using a careful distinction between the different notions of reference time, based on (Kamp and Reyle, 1993), we propose a solution to this problem, within the framework of DRT. We show some applications of this solution to additional temporal anaphora phenomena in quantified sentences.

1 Introduction

The analysis of temporal expressions in natural language discourse provides a challenge for contemporary semantics theories. (Partree, 1973) introduced the notion of temporal anaphora, to account for ways in which temporal expressions depend on surrounding elements in the discourse for their semantic contribution to the discourse. In this paper, we discuss the interaction of temporal anaphora and quantification over eventualities. Such interaction, while interesting in its own right, is also a good test-bed for theories of the semantic interpretation of temporal expressions. We discuss cases such as:

(1) Before John makes a phone call, he always lights up a cigarette (Partree, 1984).

(2) Often, when Anne came home late, Paul had already prepared dinner. (de Swart, 1991)

(3) When he came home, he always switched on the TV. He took a beer and sat down in his armchair to forget the day. (de Swart, 1991)

(4) When John is at the beach, he always squints when the sun is shining. (de Swart, 1991)

The analysis of sentences such as (1) in (Partree, 1984), within the framework of Discourse Representation Theory (DRT) (Kamp, 1981) gives the wrong thruth-conditions, when the temporal connective in the sentence is before or after. In DRT, such sentences trigger box-splitting with the eventuality of the subordinate clause and an updated refernece time in the antecedent box, and the eventuality of the main clause in the consequent box, causing undesirable universal quantification over the reference time.

This problem is analyzed in (de Swart, 1991) as an instance of the proportion problem and given a solution from a Generalized Quanifier approach. We were led to seek a solution for this problem within DRT, because of DRT's advantages as a general theory of discourse, and its choice as the underlying formalism in another research project of ours, which deals with sentences such as 1-4, in the context of natural language specifications of computerized systems. In this paper, we propose such a solution based on a careful distinction between different roles of Reichenbach's reference time (Reichenbach, 1947), adapted from (Kamp and Reyle, 1993). Figure 1 shows a 'minimal pair' of DRS's for sentence 1, one according to Partee's (1984) analysis and one according to ours.

2 Background

An analysis of the mechanism of temporal anaphoric reference hinges upon an understanding of the ontological and logical foundations of temporal reference.

C.3. Study III: Short Instructions for Human Annotation

This coding scheme is about the ownership of ideas in scientific papers and about author's stance towards other work. Your intuitions about the structure of this paper will be useful input to help build better tools for information extraction from scientific papers, which in turn will improve automatic bibliographic search.

Read the complete paper first to get a sense of what it is about. You do not have to understand the details of the paper. Then, working from the beginning, mark each

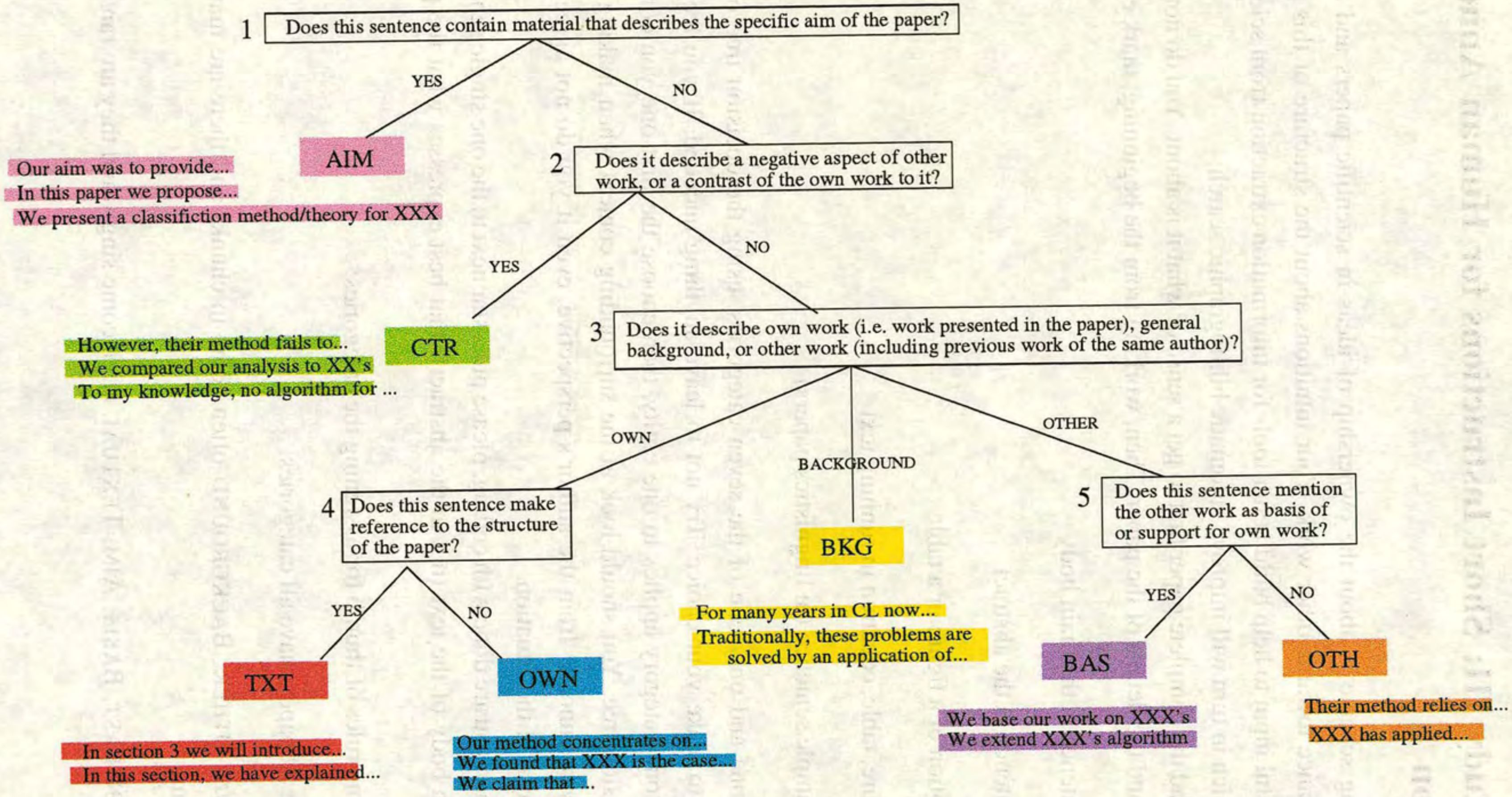
- sentence in the main body
- sentence in the abstract
- caption of a figure or a table
- figure, table, equation in running text
- example sentence (in linguistics papers)

as one and only one of the seven categories, using the decision tree on the other side to make your choice. Try not to leave anything uncoded. If you feel that more than one category applies to one entity, then choose the first one you come to in the decision tree. You should look at the surrounding context when making your choice. Try to annotate from the author's perspective, even if you do not agree with their portrayal of the situation.

When you are done with coding, please put a star next to the one single sentence in the main body of the text (not in the abstract!) that best expresses what the paper was about.

Some rules of thumb for assigning the categories:

- Not all papers have all categories.
- OWN, OTHER, BACKGROUND often come in chunks and there are many of them.
- CONTRAST, BASIS, AIM, TEXTUAL often come singly and they are rarer.


1 Does this sentence contain material that describes the specific aim of the paper?
YES
NO
AIM
Our aim was to provide...
In this paper we propose...
We present a classifiction method/theory for XXX
2 Does it describe a negative aspect of other work, or a contrast of the own work to it?
YES
NO
CTR
However, their method fails to...
We compared our analysis to XX's
To my knowledge, no algorithm for ...
3 Does it describe own work (i.e. work presented in the paper), general background, or other work (including previous work of the same author)?
OWN
BACKGROUND
OTHER
4 Does this sentence make reference to the structure of the paper?
YES
NO
TXT
In section 3 we will introduce...
In this section, we have explained...
OWN
Our method concentrates on...
We found that XXX is the case...
We claim that ...
BKG
For many years in CL now...
Traditionally, these problems are solved by an application of...
5 Does this sentence mention the other work as basis of or support for own work?
YES
NO
BAS
We base our work on XXX's
We extend XXX's algorithm
OTH
Their method relies on...
XXX has applied...

Appendix D

Lexical Resources

D.1. Formulaic Patterns

GENERAL_FORMULAIC	in @TRADITION_ADJ JJ ↑@WORK_NOUN
	in @TRADITION_ADJ used ↑@WORK_NOUN
	in @TRADITION_ADJ ↑@WORK_NOUN
	in @MANY JJ ↑@WORK_NOUN
	in @MANY ↑@WORK_NOUN
	in @BEFORE_ADJ JJ ↑@WORK_NOUN
	in @BEFORE_ADJ ↑@WORK_NOUN
	in other JJ ↑@WORK_NOUN
	in other ↑@WORK_NOUN
	in such ↑@WORK_NOUN
THEM_FORMULAIC	↑according to CITE
	along the ↑lines of CITE
	↑like CITE
	CITE ↑style
	a la ↑CITE
	CITE - ↑style
US_PREVIOUS_FORMULAIC	@SELF_NOM have ↑previously
	@SELF_NOM have ↑earlier
	@SELF_NOM have ↑elsewhere
	@SELF_NOM ↑elsewhere
	@SELF_NOM ↑previously
	@SELF_NOM ↑earlier
	↑elsewhere @SELF_NOM
	↑elswhere @SELF_NOM
	↑elsewhere , @SELF_NOM
	↑elswhere , @SELF_NOM
	presented ↑elswhere
	presented ↑elsewhere
	@SELF_NOM have shown ↑elsewhere
	@SELF_NOM have argued ↑elsewhere
	@SELF_NOM have shown ↑elswhere_NOM
	@SELF_NOM have argued ↑elswhere_NOM
	@SELF_NOM will show ↑elsewhere
	@SELF_NOM will show ↑elswhere

	@SELF_NOM will argue ↑elsewhere
	@SELF_NOM will argue ↑elswhere
	↑elsewhere SELFCITE
	↑elswhere SELFCITE
	in a @BEFORE_ADJ ↑@PRESENTATION_NOUN
	in an earlier ↑@PRESENTATION_NOUN
	another ↑@PRESENTATION_NOUN
TEXTSTRUCTURE_FORMULAIC	↑then @SELF_NOM describe
	↑then , @SELF_NOM describe
	↑next @SELF_NOM describe
	↑next , @SELF_NOM describe
	↑finally @SELF_NOM describe
	↑finally , @SELF_NOM describe
	↑then @SELF_NOM present
	↑then , @SELF_NOM present
	↑next @SELF_NOM present
	↑next , @SELF_NOM present
	↑finally @SELF_NOM present
	↑finally , @SELF_NOM present
	↑briefly describe
	↑briefly introduce
	↑briefly present
	↑briefly discuss
HERE_FORMULAIC	in this ↑@PRESENTATION_NOUN
	the present ↑@PRESENTATION_NOUN
	@SELF_NOM ↑here
	↑here @SELF_NOM
	↑here , @SELF_NOM
	@GIVEN ↑here
	@SELF_NOM ↑now
	↑now @SELF_NOM
	↑now , @SELF_NOM
	@GIVEN ↑now
	herein
METHOD_FORMULAIC	a new ↑@WORK_NOUN
	a novel ↑@WORK_NOUN
	a ↑@WORK_NOUN of
	an ↑@WORK_NOUN of
	a JJ ↑@WORK_NOUN of
	an JJ ↑@WORK_NOUN of
	a NN ↑@WORK_NOUN of
	an NN ↑@WORK_NOUN of
	a JJ NN ↑@WORK_NOUN of
	an JJ NN ↑@WORK_NOUN of
	a ↑@WORK_NOUN for
	an ↑@WORK_NOUN for
	a JJ ↑@WORK_NOUN for
	an JJ ↑@WORK_NOUN for
	a NN ↑@WORK_NOUN for
	an NN ↑@WORK_NOUN for
	a JJ NN ↑@WORK_NOUN for
	an JJ NN ↑@WORK_NOUN for
	↑@WORK_NOUN designed to VV

	↑@WORK_NOUN intended for
	↑@WORK_NOUN for VV_ING
	↑@WORK_NOUN for the NN
	↑@WORK_NOUN designed to VV
	↑@WORK_NOUN to the NN
	↑@WORK_NOUN to NN
	↑@WORK_NOUN to VV_ING
	↑@WORK_NOUN for JJ VV_ING
	↑@WORK_NOUN for the JJ NN
	↑@WORK_NOUN to the JJ NN
	↑@WORK_NOUN to JJ VV_ING
	the ↑problem of RB VV_ING
	the ↑problem of VV_ING
	the ↑problem of how to
CONTINUE_FORMULAIC	↑following CITE
	↑following the @WORK_NOUN of CITE
	↑following the @WORK_NOUN given in CITE
	↑following the @WORK_NOUN presented in CITE
	↑following the @WORK_NOUN proposed in CITE
	↑following the @WORK_NOUN discussed in CITE
	↑adopt CITE 's
	↑starting point for @REFERENTIAL @WORK_NOUN
	↑starting point for @SELF_POSS @WORK_NOUN
	as a ↑starting point
	as ↑starting point
	↑use CITE 's
	↑base @SELF_POSS
	↑supports @SELF_POSS
	↑supports @OTHERS_POSS
	↑support @OTHERS_POSS
	↑support @SELF_POSS
	lends ↑support to @SELF_POSS
	lends ↑support to @OTHERS_POSS
CONTRAST_FORMULAIC	however, nevertheless, nonetheless, unfortunately, yet, although
GAP_FORMULAIC	as far as @SELF_NOM ↑know
	to @SELF_POSS ↑knowledge
	to the best of @SELF_POSS ↑knowledge
FUTURE_FORMULAIC	in the ↑future
	in the near ↑future
	↑@FUTURE_ADJ @WORK_NOUN
	↑@FUTURE_ADJ @AIM_NOUN
	↑@FUTURE_ADJ development
	needs ↑further
	requires ↑further
	beyond the ↑scope
	↑avenue for improvement
	↑avenues for improvement
	↑avenues for @FUTURE_ADJ improvement
	↑areas for @FUTURE_ADJ improvement
	↑areas for improvement
	↑avenues of @FUTURE_ADJ research
	promising ↑avenue
	promising ↑avenues

SIMILARITY_FORMULAIC	along the same ↑lines
	in a ↑similar vein
	as in ↑@SELF_POSS
	as in ↑CITE
	as ↑did CITE
	like in ↑CITE
	↑like CITE 's
	similarity with ↑CITE
	similarity with ↑@SELF_POSS
	similarity with ↑@OTHERS_POSS
	↑similarity with @TRADITION_ADJ
	↑similarity with @MANY
	↑similarity with @BEFORE_ADJ
	in analogy to ↑CITE
	in analogy to ↑@SELF_POSS
	in analogy to ↑@OTHERS_POSS
	in ↑analogy to @TRADITION_ADJ
	in ↑analogy to @MANY
	in ↑analogy to @BEFORE_ADJ
	↑similar to that described here
	↑similar to that of
	↑similar to those of
	↑similar to CITE
	↑similar to @SELF_ACC
	↑similar to @SELF_POSS
	↑similar to @OTHERS_ACC
	↑similar to @TRADITION_ADJ
	↑similar to @MANY
	↑similar to @BEFORE_ADJ
	↑similar to @OTHERS_POSS
	↑similar to CITE
	a ↑similar NN to @SELF_POSS
	a ↑similar NN to @OTHERS_POSS
	a ↑similar NN to CITE
	↑analogous to that described here
	↑analogous to CITE
	↑analogous to @SELF_ACC
	↑analogous to @SELF_POSS
	↑analogous to @OTHERS_ACC
	↑analogous to @TRADITION_ADJ
	↑analogous to @MANY
	↑analogous to @BEFORE_ADJ
	↑analogous to @OTHERS_POSS
	↑analogous to CITE
	the ↑same NN as @SELF_POSS
	the ↑same NN as @OTHERS_POSS
	the ↑same NN as CITE
	the ↑same as @SELF_POSS
	the ↑same as @OTHERS_POSS
	the ↑same as CITE
	in ↑common with @OTHERS_POSS
	in ↑common with @SELF_POSS
	in ↑common with @TRADITION_ADJ

	in ↑common with @MANY
	in ↑common with @BEFORE_ADJ
	most ↑relevant to @SELF_POSS
COMPARISON_FORMULAIC	↑against CITE
	↑against @SELF_ACC
	↑against @SELF_POSS
	↑against @OTHERS_ACC
	↑against @OTHERS_POSS
	↑against @BEFORE_ADJ @WORK_NOUN
	↑against @MANY @WORK_NOUN
	↑against @TRADITION_ADJ @WORK_NOUN
	↑than CITE
	↑than @SELF_ACC
	↑than @SELF_POSS
	↑than @OTHERS_ACC
	↑than @OTHERS_POSS
	↑than @TRADITION_ADJ @WORK_NOUN
	↑than @BEFORE_ADJ @WORK_NOUN
	↑than @MANY @WORK_NOUN
	point of ↑departure from @SELF_POSS
	points of ↑departure from @OTHERS_POSS
	↑advantage over @OTHERS_ACC
	↑advantage over @TRADITION_ADJ
	↑advantage over @MANY @WORK_NOUN
	↑advantage over @BEFORE_ADJ @WORK_NOUN
	↑advantage over @OTHERS_POSS
	↑advantage over CITE
	↑advantage to @OTHERS_ACC
	↑advantage to @OTHERS_POSS
	↑advantage to CITE
	↑advantage to @TRADITION_ADJ
	↑advantage to @MANY @WORK_NOUN
	↑advantage to @BEFORE_ADJ @WORK_NOUN
	↑advantages over @OTHERS_ACC
	↑advantages over @TRADITION_ADJ
	↑advantages over @MANY @WORK_NOUN
	↑advantages over @BEFORE_ADJ @WORK_NOUN
	↑advantages over @OTHERS_POSS
	↑advantages over CITE
	↑advantages to @OTHERS_ACC
	↑advantages to @OTHERS_POSS
	↑advantages to CITE
	↑advantages to @TRADITION_ADJ
	↑advantages to @MANY @WORK_NOUN
	↑advantages to @BEFORE_ADJ @WORK_NOUN
	↑benefit over @OTHERS_ACC
	↑benefit over @OTHERS_POSS
	↑benefit over CITE
	↑benefit over @TRADITION_ADJ
	↑benefit over @MANY @WORK_NOUN
	↑benefit over @BEFORE_ADJ @WORK_NOUN
	↑difference to CITE
	↑difference to @TRADITION_ADJ

↑difference to CITE
↑difference to @TRADITION_ADJ
↑difference to @MANY @WORK_NOUN
↑difference to @BEFORE_ADJ @WORK_NOUN
↑difference to @OTHERS_ACC
↑difference to @OTHERS_POSS
↑difference to @SELF_ACC
↑difference to @SELF_POSS
↑differences to CITE
↑differences to @TRADITION_ADJ
↑differences to @MANY @WORK_NOUN
↑differences to @BEFORE_ADJ @WORK_NOUN
↑differences to @OTHERS_ACC
↑differences to @OTHERS_POSS
↑differences to @SELF_ACC
↑differences to @SELF_POSS
↑difference between CITE
↑difference between @TRADITION_ADJ
↑difference between @MANY @WORK_NOUN
↑difference between @BEFORE_ADJ @WORK_NOUN
↑difference between @OTHERS_ACC
↑difference between @OTHERS_POSS
↑difference between @SELF_ACC
↑difference between @SELF_POSS
↑differences between CITE
↑differences between @TRADITION_ADJ
↑differences between @MANY @WORK_NOUN
↑differences between @BEFORE_ADJ @WORK_NOUN
↑differences between @OTHERS_ACC
↑differences between @OTHERS_POSS
↑differences between @SELF_ACC
↑differences between @SELF_POSS
↑contrast with CITE
↑contrast with @TRADITION_ADJ
↑contrast with @MANY @WORK_NOUN
↑contrast with @BEFORE_ADJ @WORK_NOUN
↑contrast with @OTHERS_ACC
↑contrast with @OTHERS_POSS
↑contrast with @SELF_ACC
↑contrast with @SELF_POSS
↑unlike @SELF_ACC
↑unlike @SELF_POSS
↑unlike CITE
↑unlike @TRADITION_ADJ
↑unlike @BEFORE_ADJ @WORK_NOUN
↑unlike @MANY @WORK_NOUN
↑unlike @OTHERS_ACC
↑unlike @OTHERS_POSS
in ↑contrast to @SELF_ACC
in ↑contrast to @SELF_POSS
in ↑contrast to CITE
in ↑contrast to @TRADITION_ADJ
in ↑contrast to @MANY @WORK_NOUN

	in ↑contrast to @BEFORE_ADJ @WORK_NOUN
	in ↑contrast to @OTHERS_ACC
	in ↑contrast to @OTHERS_POSS
	as ↑opposed to @SELF_ACC
	as ↑opposed to @SELF_POSS
	as ↑opposed to CITE
	as ↑opposed to @TRADITION_ADJ
	as ↑opposed to @MANY @WORK_NOUN
	as ↑opposed to @BEFORE_ADJ @WORK_NOUN
	as ↑opposed to @OTHERS_ACC
	as ↑opposed to @OTHERS_POSS
	↑contrary to @SELF_ACC
	↑contrary to @SELF_POSS
	↑contrary to CITE
	↑contrary to @TRADITION_ADJ
	↑contrary to @MANY @WORK_NOUN
	↑contrary to @BEFORE_ADJ @WORK_NOUN
	↑contrary to @OTHERS_ACC
	↑contrary to @OTHERS_POSS
	↑whereas @SELF_ACC
	↑whereas @SELF_POSS
	↑whereas CITE
	↑whereas @TRADITION_ADJ
	↑whereas @BEFORE_ADJ @WORK_NOUN
	↑whereas @MANY @WORK_NOUN
	↑whereas @OTHERS_ACC
	↑whereas @OTHERS_POSS
	↑compared to @SELF_ACC
	↑compared to @SELF_POSS
	↑compared to CITE
	↑compared to @TRADITION_ADJ
	↑compared to @BEFORE_ADJ @WORK_NOUN
	↑compared to @MANY @WORK_NOUN
	↑compared to @OTHERS_ACC
	↑compared to @OTHERS_POSS
	in ↑comparison to @SELF_ACC
	in ↑comparison to @SELF_POSS
	in ↑comparison to CITE
	in ↑comparison to @TRADITION_ADJ
	in ↑comparison to @MANY @WORK_NOUN
	in ↑comparison to @BEFORE_ADJ @WORK_NOUN
	in ↑comparison to @OTHERS_ACC
	in ↑comparison to @OTHERS_POSS
	↑while @SELF_NOM
	↑while @SELF_POSS
	↑while CITE
	↑while @TRADITION_ADJ
	↑while @BEFORE_ADJ @WORK_NOUN
	↑while @MANY @WORK_NOUN
	↑while @OTHERS_NOM
	↑while @OTHERS_POSS
AFFECT_FORMULAIC	hopefully
	thankfully

	fortunately
	unfortunately
GOOD_FORMULAIC	@POS_ADJ
BAD_FORMULAIC	@NEG_ADJ
TRADITION_FORMULAIC	@TRADITIONAL_ADJ
IN_ORDER_TO_FORMULAIC	in ↑order to
DETAIL_FORMULAIC	@SELF_NOM have ↑also
	@SELF_NOM ↑also
	this @PRESENTATION_NOUN ↑also
	this @PRESENTATION_NOUN has ↑also
NO_TEXTSTRUCTURE_FORMULAIC	(↑TXT_NOUN CREF)
	as explained in ↑@TXT_NOUN CREF
	as explained in the @BEFORE_ADJ ↑@TXT_NOUN
	as ↑@GIVEN earlier in this @TXT_NOUN
	as ↑@GIVEN below
	as @GIVEN in ↑@TXT_NOUN CREF
	as @GIVEN in the @BEFORE_ADJ ↑@TXT_NOUN
	as @GIVEN in the next ↑@TXT_NOUN
	NN @GIVEN in ↑@TXT_NOUN CREF
	NN @GIVEN in the @BEFORE_ADJ ↑@TXT_NOUN
	NN @GIVEN in the next ↑@TXT_NOUN
	NN @GIVEN ↑below
	cf. ↑@TXT_NOUN CREF
	cf. ↑@TXT_NOUN below
	cf. the ↑@TXT_NOUN below
	cf. the @BEFORE_ADJ ↑@TXT_NOUN
	cf. ↑@TXT_NOUN above
	cf. the ↑@TXT_NOUN above
	e. g. , ↑@TXT_NOUN CREF
	e. g , ↑@TXT_NOUN CREF
	e. g. ↑@TXT_NOUN CREF
	e. g ↑@TXT_NOUN CREF
	compare ↑@TXT_NOUN CREF
	compare ↑@TXT_NOUN below
	compare the ↑@TXT_NOUN below
	compare the @BEFORE_ADJ ↑@TXT_NOUN
	compare ↑@TXT_NOUN above
	compare the ↑@TXT_NOUN above
	see ↑@TXT_NOUN CREF
	see the @BEFORE_ADJ ↑@TXT_NOUN
	recall from the @BEFORE_ADJ ↑@TXT_NOUN
	recall from the ↑@TXT_NOUN above
	recall from ↑@TXT_NOUN CREF
	@SELF_NOM shall see ↑below
	@SELF_NOM will see ↑below
	@SELF_NOM shall see in the ↑next @TXT_NOUN
	@SELF_NOM will see in the ↑next @TXT_NOUN
	@SELF_NOM shall see in ↑@TXT_NOUN CREF
	@SELF_NOM will see in ↑@TXT_NOUN CREF
	example in ↑@TXT_NOUN CREF
	example CREF in ↑@TXT_NOUN CREF
	examples CREF and CREF in ↑@TXT_NOUN CREF
	examples in ↑@TXT_NOUN CREF

D.2. Agent Patterns

US_AGENT	@SELF_NOM
	@SELF_POSS JJ ↑@WORK_NOUN
	@SELF_POSS JJ ↑@PRESENTATION_NOUN
	@SELF_POSS JJ ↑@ARGUMENTATION_NOUN
	@SELF_POSS JJ ↑@SOLUTION_NOUN
	@SELF_POSS JJ ↑@RESULT_NOUN
	@SELF_POSS ↑@WORK_NOUN
	@SELF_POSS ↑@PRESENTATION_NOUN
	@SELF_POSS ↑@ARGUMENTATION_NOUN
	@SELF_POSS ↑@SOLUTION_NOUN
	@SELF_POSS ↑@RESULT_NOUN
	↑@WORK_NOUN @GIVEN here
	↑@WORK_NOUN @GIVEN below
	↑@WORK_NOUN @GIVEN in this @PRESENTATION_NOUN
	↑@WORK_NOUN @GIVEN in @SELF_POSS @PRESENTATION_NOUN
	the ↑@SOLUTION_NOUN @GIVEN here
	the ↑@SOLUTION_NOUN @GIVEN in this @PRESENTATION_NOUN
	the first ↑author
	the second ↑author
	the third ↑author
	one of the ↑authors
	one of ↑us
REF_US_AGENT	this ↑@PRESENTATION_NOUN
	the present ↑@PRESENTATION_NOUN
	the current ↑@PRESENTATION_NOUN
	the present JJ ↑@PRESENTATION_NOUN
	the current JJ ↑@PRESENTATION_NOUN
	the ↑@WORK_NOUN @GIVEN
OUR_AIM_AGENT	@SELF_POSS ↑@AIM_NOUN
	the point of this ↑@PRESENTATION_NOUN
	the ↑@AIM_NOUN of this @PRESENTATION_NOUN
	the ↑@AIM_NOUN of the @GIVEN @WORK_NOUN
	the ↑@AIM_NOUN of @SELF_POSS @WORK_NOUN
	the ↑@AIM_NOUN of @SELF_POSS @PRESENTATION_NOUN
	the most important feature of ↑@SELF_POSS @WORK_NOUN
	contribution of this ↑@PRESENTATION_NOUN
	contribution of the @GIVEN ↑@WORK_NOUN
	contribution of ↑@SELF_POSS @WORK_NOUN
	the question @GIVEN in this ↑PRESENTATION_NOUN
	the question @GIVEN ↑here
	@SELF_POSS @MAIN ↑@AIM_NOUN
	@SELF_POSS ↑@AIM_NOUN in this @PRESENTATION_NOUN
	@SELF_POSS ↑@AIM_NOUN here
	the JJ point of this ↑@PRESENTATION_NOUN
	the JJ purpose of this ↑@PRESENTATION_NOUN
	the JJ ↑@AIM_NOUN of this @PRESENTATION_NOUN
	the JJ ↑@AIM_NOUN of the @GIVEN @WORK_NOUN
	the JJ ↑@AIM_NOUN of @SELF_POSS @WORK_NOUN
	the JJ ↑@AIM_NOUN of @SELF_POSS @PRESENTATION_NOUN
	the JJ question @GIVEN in this ↑PRESENTATION_NOUN

	the JJ question @GIVEN ↑here
AIM_REF_AGENT	its ↑@AIM_NOUN
	its JJ ↑@AIM_NOUN
	@REFERENTIAL JJ ↑@AIM_NOUN
	contribution of this ↑@WORK_NOUN
	the most important feature of this ↑@WORK_NOUN
	feature of this ↑@WORK_NOUN
	the ↑@AIM_NOUN
	the JJ ↑@AIM_NOUN
US_PREVIOUS_AGENT	SELFCITE
	this @BEFORE_ADJ ↑@PRESENTATION_NOUN
	@SELF_POSS @BEFORE_ADJ ↑@PRESENTATION_NOUN
	@SELF_POSS @BEFORE_ADJ ↑@WORK_NOUN
	in ↑SELFCITE , @SELF_NOM
	in ↑SELFCITE @SELF_NOM
	the ↑@WORK_NOUN @GIVEN in SELFCITE
REF_AGENT	@REFERENTIAL JJ ↑@WORK_NOUN
	@REFERENTIAL ↑@WORK_NOUN
	this sort of ↑@WORK_NOUN
	this kind of ↑@WORK_NOUN
	this type of ↑@WORK_NOUN
	the current JJ ↑@WORK_NOUN
	the current ↑@WORK_NOUN
	the ↑@WORK_NOUN
	the ↑@PRESENTATION_NOUN
	the ↑author
	the ↑authors
THEM_PRONOUN_AGENT	@OTHERS_NOM
THEM_AGENT	CITE
	CITE 's NN
	CITE 's ↑@PRESENTATION_NOUN
	CITE 's ↑@WORK_NOUN
	CITE 's ↑@ARGUMENTATION_NOUN
	CITE 's JJ ↑@PRESENTATION_NOUN
	CITE 's JJ ↑@WORK_NOUN
	CITE 's JJ ↑@ARGUMENTATION_NOUN
	the CITE ↑@WORK_NOUN
	the ↑@WORK_NOUN @GIVEN in CITE
	the ↑@WORK_NOUN of CITE
	@OTHERS_POSS ↑@PRESENTATION_NOUN
	@OTHERS_POSS ↑@WORK_NOUN
	@OTHERS_POSS ↑@RESULT_NOUN
	@OTHERS_POSS ↑@ARGUMENTATION_NOUN
	@OTHERS_POSS ↑@SOLUTION_NOUN
	@OTHERS_POSS JJ ↑@PRESENTATION_NOUN
	@OTHERS_POSS JJ ↑@WORK_NOUN
	@OTHERS_POSS JJ ↑@RESULT_NOUN
	@OTHERS_POSS JJ ↑@ARGUMENTATION_NOUN
	@OTHERS_POSS JJ ↑@SOLUTION_NOUN
GAP_AGENT	none of these ↑@WORK_NOUN
	none of those ↑@WORK_NOUN
	no ↑@WORK_NOUN
	no JJ ↑@WORK_NOUN

	none of these ↑@PRESENTATION_NOUN
	none of those ↑@PRESENTATION_NOUN
	no ↑@PRESENTATION_NOUN
	no JJ ↑@PRESENTATION_NOUN
GENERAL_AGENT	@TRADITION_ADJ JJ ↑@WORK_NOUN
	@TRADITION_ADJ used ↑@WORK_NOUN
	@TRADITION_ADJ ↑@WORK_NOUN
	@MANY JJ ↑@WORK_NOUN
	@MANY ↑@WORK_NOUN
	@BEFORE_ADJ JJ ↑@WORK_NOUN
	@BEFORE_ADJ ↑@WORK_NOUN
	@BEFORE_ADJ JJ ↑@PRESENTATION_NOUN
	@BEFORE_ADJ ↑@PRESENTATION_NOUN
	other JJ ↑@WORK_NOUN
	other ↑@WORK_NOUN
	such ↑@WORK_NOUN
	these JJ ↑@PRESENTATION_NOUN
	these ↑@PRESENTATION_NOUN
	those JJ ↑@PRESENTATION_NOUN
	those ↑@PRESENTATION_NOUN
	@REFERENTIAL ↑authors
	@MANY ↑authors
	↑researchers in @DISCIPLINE
	@PROFESSIONAL_NOUN
PROBLEM_AGENT	@REFERENTIAL JJ ↑@PROBLEM_NOUN
	@REFERENTIAL ↑@PROBLEM_NOUN
	the ↑@PROBLEM_NOUN
SOLUTION_AGENT	@REFERENTIAL JJ ↑@SOLUTION_NOUN
	@REFERENTIAL ↑@SOLUTION_NOUN
	the ↑@SOLUTION_NOUN
	the JJ ↑@SOLUTION_NOUN
TEXTSTRUCTURE_AGENT	↑@TXT_NOUN CREF
	↑@TXT_NOUN CREF and CREF
	this ↑@TXT_NOUN
	next ↑@TXT_NOUN
	next CD ↑@TXT_NOUN
	concluding ↑@TXT_NOUN
	@BEFORE_ADJ ↑@TXT_NOUN
	↑@TXT_NOUN above
	↑@TXT_NOUN below
	following ↑@TXT_NOUN
	remaining ↑@TXT_NOUN
	subsequent ↑@TXT_NOUN
	following CD ↑@TXT_NOUN
	remaining CD ↑@TXT_NOUN
	subsequent CD ↑@TXT_NOUN
	↑@TXT_NOUN that follow
	rest of this ↑@PRESENTATION_NOUN
	remainder of this ↑@PRESENTATION_NOUN
	in ↑@TXT_NOUN CREF , @SELF_NOM
	in this ↑@TXT_NOUN , @SELF_NOM
	in the next ↑@TXT_NOUN , @SELF_NOM
	in @BEFORE_ADJ ↑@TXT_NOUN , @SELF_NOM

in the @BEFORE_ADJ ↑@TXT_NOUN , @SELF_NOM
in the ↑@TXT_NOUN above , @SELF_NOM
in the ↑@TXT_NOUN below , @SELF_NOM
in the following ↑@TXT_NOUN , @SELF_NOM
in the remaining ↑@TXT_NOUN , @SELF_NOM
in the subsequent ↑@TXT_NOUN , @SELF_NOM
in the ↑@TXT_NOUN that follow , @SELF_NOM
in the rest of this ↑@PRESENTATION_NOUN , @SELF_NOM
in the remainder of this ↑@PRESENTATION_NOUN , @SELF_NOM
↑below , @SELF_NOM
the ↑@AIM_NOUN of this @TXT_NOUN

D.3. Action Lexicon

AFFECT	afford, believe, decide, feel, hope, imagine, regard, trust, think
ARGUMENTATION	agree, accept, advocate, argue, claim, conclude, comment, defend, embrace, hypothesize, imply, insist, posit, postulate, reason, recommend, speculate, stipulate, suspect
AWARE	be unaware, be familiar with, be aware, be not aware, know of
BETTER_SOLUTION	boost, enhance, defeat, improve, go beyond, perform better, outperform, outweigh, surpass
CHANGE	adapt, adjust, augment, combine, change, decrease, elaborate, expand, extend, derive, incorporate, increase, manipulate, modify, optimize, optimise, refine, render, replace, revise, substitute, tailor, upgrade
COMPARISON	compare, compete, evaluate, test
CONTINUE	adopt, agree with CITE, base, be based on, be derived from, be originated in, be inspired by, borrow, build on, follow CITE, originate from, originate in, side with
CONTRAST	be different from, be distinct from, conflict, contrast, clash, differ from, distinguish @RFX, differentiate, disagree, disagreeing, dissent, oppose
FUTURE_INTEREST	plan on, plan to, expect to, intend to
INTEREST	aim, ask @SELF_RFX, ask @OTHERS_RFX, address, attempt, be concerned, be interested, be motivated, concern, concern @SELF_ACC, concern @OTHERS_ACC, consider, concentrate on, explore, focus, intend to, like to, look at how, motivate @SELF_ACC, motivate @OTHERS_ACC, pursue, seek, study, try, target, want, wish, wonder
NEED	be dependent on, be reliant on, depend on, lack, need, necessitate, require, rely on
PRESENTATION	describe, discuss, give, introduce, note, notice, point out, present, propose, put forward, recapitulate, remark, report, say, show, sketch, state, suggest, talk about
PROBLEM	abound, aggravate, arise, be cursed, be incapable of, be forced to, be limited to, be problematic, be restricted to, be troubled, be unable to, contradict, damage, degrade, degenerate, fail, fall prey, fall short, force @SELF_ACC, force @OTHERS_ACC, hinder, impair, impede, inhibit, misclassify, misjudge, mistake, misuse, neglect, obscure, overestimate, over-estimate, overfit, over-fit, overgeneralize, over-generalize, overgeneralise, over-generalise, overgenerate, over-generate, overlook, pose, plague, preclude, prevent, remain, resort to, restrain, run into, settle for, spoil, suffer from, threaten, thwart, underestimate, under-estimate, undergenerate, under-generate, violate, waste, worsen
RESEARCH	apply, analyze, analyse, build, calculate, categorize, categorise, characterize, characterise, choose, check, classify, collect, compose, compute, conduct, confirm, construct, count, define, delineate, detect, determine, equate, estimate, examine, expect, formalize, formalise, formulate, gather, identify, implement,

indicate, inspect, integrate, interpret, investigate, isolate, maximize, maximise, measure, minimize, minimise, observe, predict, realize, realise, reconfirm, simulate, select, specify, test, verify

SIMILAR bear comparison, be analogous to, be alike, be related to, be closely related to, be reminiscent of, be the same as, be similar to, be in a similar vein to, have much in common with, have a lot in common with, pattern with, resemble

SOLUTION accomplish, account for, achieve, apply to, answer, alleviate, allow for, allow @SELF_ACC, allow @OTHERS_ACC, avoid, benefit, capture, clarify, circumvent, contribute, cope with, cover, cure, deal with, demonstrate, develop, devise, discover, elucidate, escape, explain, fix, gain, go a long way, guarantee, handle, help, implement, justify, lend itself, make progress, manage, mend, mitigate, model, obtain, offer, overcome, perform, preserve, prove, provide, realize, realise, rectify, refrain from, remedy, resolve, reveal, scale up, sidestep, solve, succeed, tackle, take care of, take into account, treat, warrant, work well, yield

TEXTSTRUCTURE begin by, illustrate, conclude by, organize, organise, outline, return to, review, start by, structure, summarize, summarise, turn to

USE apply, employ, use, make use, utilize

D.4. Concept Lexicon

NEGATION	no, not, nor, non, neither, none, never, aren't, can't, cannot, hadn't, hasn't, haven't, isn't, didn't, don't, doesn't, n't, wasn't, weren't, nothing, nobody, less, least, little, scant, scarcely, rarely, hardly, few, rare, unlikely
3RD PERSON PRONOUN (NOM)	they, he, she, theirs, hers, his
3RD PERSON PRONOUN (ACC)	her, him, them
3RD POSS PRONOUN	their, his, her
3RD PERSON REFLEXIVE	themselves, himself, herself
1ST PERSON PRONOUN (NOM)	we, i, ours, mine
1ST PERSON PRONOUN (ACC)	us, me
1ST POSS PRONOUN	my, our
1ST PERSON REFLEXIVE	ourselves, myself
REFERENTIAL	this, that, those, these
REFLEXIVE	itself ourselves, myself, themselves, himself, herself
QUESTION	?, how, why, whether, wonder
GIVEN	noted, mentioned, addressed, illustrated, described, discussed, given, outlined, presented, proposed, reported, shown, taken
PROFESSIONALS	collegues, community, computer scientists, computational linguists, discourse analysts, expert, investigators, linguists, logicians, philosophers, psycholinguists, psychologists, researchers, scholars, semanticists, scientists
DISCIPLINE	computer science, computer linguistics, computational linguistics, discourse analysis, logics, linguistics, psychology, psycholinguistics, philosophy, semantics, several disciplines, various disciplines
TEXT_NOUN	paragraph, section, subsection, chapter
SIMILAR_NOUN	analogy, similarity
COMPARISON_NOUN	accuracy, baseline, comparison, competition, evaluation, inferiority, measure, measurement, performance, precision, optimum, recall, superiority
CONTRAST_NOUN	contrast, conflict, clash, clashes, difference, point of departure
AIM_NOUN	aim, goal, intention, objective, purpose, task, theme, topic
ARGUMENTATION_NOUN	assumption, belief, hypothesis, hypotheses, claim, conclusion, confirmation, opinion, recommendation, stipulation, view
PROBLEM_NOUN	Achilles heel, caveat, challenge, complication, contradiction, damage, danger, deadlock, defect, detriment, difficulty, dilemma, disadvantage, disregard, doubt, downside, drawback, error, failure, fault, foil, flaw, handicap, hindrance, hurdle, ill, inflexibility, impediment, imperfection, intractability, inefficiency, inadequacy, inability, lapse, limitation, malheur, mishap, mischance, mistake, obstacle, oversight, pitfall, problem, shortcoming, threat, trouble, vulnerability, absence, dearth, deprivation, lack, loss, fraught, proliferation, spate
QUESTION_NOUN	question, conundrum, enigma, paradox, phenomena, phenomenon, puzzle, riddle
SOLUTION_NOUN	answer, accomplishment, achievement, advantage, benefit, breakthrough, contribution, explanation, idea, improvement, innovation, insight, justification, proposal, proof, remedy, solution, success, triumph, verification, victory

INTEREST_NOUN	attention, quest
RESEARCH_NOUN	evidence, experiment, finding, progress, observation, outcome, result
CHANGE_NOUN	alternative, adaptation, extension, development, modification, refinement, version, variant, variation
PRESENTATION_NOUN	article, draft, paper, project, report, study
NEED_NOUN	necessity, motivation
WORK_NOUN	account, algorithm, analysis, analyses, approach, approaches, application, architecture, characterization, characterisation, component, design, extension, formalism, formalization, formalisation, framework, implementation, investigation, machinery, method, methodology, model, module, moduls, process, procedure, program, prototype, research, researches, strategy, system, technique, theory, tool, treatment, work
TRADITION_NOUN	acceptance, community, convention, disciples, disciplines, folklore, literature, mainstream, school, tradition, textbook
CHANGE_ADJ	alternate, alternative
GOOD_ADJ	adequate, advantageous, appealing, appropriate, attractive, automatic, beneficial, capable, cheerful, clean, clear, compact, compelling, competitive, comprehensive, consistent, convenient, convincing, constructive, correct, desirable, distinctive, efficient, elegant, encouraging, exact, faultless, favourable, feasible, flawless, good, helpful, impeccable, innovative, insightful, intensive, meaningful, neat, perfect, plausible, positive, polynomial, powerful, practical, preferable, precise, principled, promising, pure, realistic, reasonable, reliable, right, robust, satisfactory, simple, sound, successful, sufficient, systematic, tractable, usable, useful, valid, unlimited, well worked out, well, enough
BAD_ADJ	absent, ad-hoc, adhoc, ad hoc, annoying, ambiguous, arbitrary, awkward, bad, brittle, brute-force, brute force, careless, confounding, contradictory, defect, defunct, disturbing, elusive, erraneous, expensive, exponential, false, fallacious, frustrating, haphazard, ill-defined, imperfect, impossible, impractical, imprecise, inaccurate, inadequate, inappropriate, incomplete, incomprehensible, inconclusive, incorrect, inelegant, inefficient, inexact, infeasible, infelicitous, inflexible, implausible, inpracticable, improper, insufficient, intractable, invalid, irrelevant, labour-intensive, labor-intensive, labour intensive, labor intensive, limited-coverage, limited coverage, limited, limiting, meaningless, modest, misguided, misleading, non-existent, NP-hard, NP-complete, NP hard, NP complete, questionable, pathological, poor, prone, protracted, restricted, scarce, simplistic, suspect, time-consuming, time consuming, toy, unacceptable, unaccounted for, unaccounted-for, unaccounted, unattractive, unavailable, unavoidable, unclear, uncomfortable, unexplained, undecidable, undesirable, unfortunate, uninnovative, uninterpretable, unjustified, unmotivated, unnatural, unnecessary, unorthodox, unpleasant, unpractical, unprincipled, unreliable, unsatisfactory, unsound, unsuccessful, unsuited, unsystematic, untractable, unwanted, unwelcome, useless, vulnerable, weak, wrong, too, overly, only
BEFORE_ADJ	earlier, past, previous, prior
CONTRAST_ADJ	different, distinguishing, contrary, competing, rival
TRADITION_ADJ	better known, better-known, cited, classic, common, conventional, current, customary, established, existing, extant, available, favourite, fashionable, general, obvious, long-standing, mainstream, modern, naive, orthodox, popular, prevailing, prevalent, published, quoted, seminal, standard, textbook, traditional, trivial, typical, well-established, well-known, widely-assumed, unanimous, usual

MANY	a number of, a body of, a substantial number of, a substantial body of, most, many, several, various
COMPARISON_ADJ	evaluative, superior, inferior, optimal, better, best, worse, worst, greater, larger, faster, weaker, stronger
PROBLEM_ADJ	demanding, difficult, hard, non-trivial, nontrivial
RESEARCH_ADJ	empirical, experimental, exploratory, ongoing, quantitative, qualitative, preliminary, statistical, underway
AWARE_ADJ	unnoticed, understood, unexplored
NEED_ADJ	necessary
NEW_ADJ	new, novel,state-of-the-art, state of the art, leading-edge, leading edge, enhanced
FUTURE_ADJ	further, future
MAIN_ADJ	main, key, basic, central, crucial, essential, eventual, fundamental, great, important, key, largest, main, major, overall, primary, principle, serious, substantial, ultimate

Appendix E

Reviewed Publications about Work Presented in this Thesis

E.1. Teufel and Moens (1997)

Sentence extraction as a classification task

Simone Teufel
Centre for Cognitive Science
and Language Technology Group
University of Edinburgh
S.Teufel@ed.ac.uk

Marc Moens
Language Technology Group
University of Edinburgh
M.Moens@ed.ac.uk

Abstract

A useful first step in document summarisation is the selection of a small number of 'meaningful' sentences from a larger text. Kupiec et al. (1995) describe this as a classification task: on the basis of a corpus of technical papers with summaries written by professional abstractors, their system identifies those sentences in the text which also occur in the summary, and then acquires a model of the 'abstract-worthiness' of a sentence as a combination of a limited number of properties of that sentence.

We report on a replication of this experiment with different data: summaries for our documents were not written by professional abstractors, but by the authors themselves. This produced fewer alignable sentences to train on. We use alternative 'meaningful' sentences (selected by a human judge) as training and evaluation material, because this has advantages for the subsequent automatic generation of more flexible abstracts. We quantitatively compare the two different strategies for training and evaluation (viz. alignment vs. human judgement); we also discuss qualitative differences and consequences for the generation of abstracts.

1 Introduction

A useful first step in the automatic or semi-automatic generation of abstracts from source texts is the selection of a small number of 'meaningful' sentences from the source text. To achieve this, each sentence in the source text is scored according to some measure of importance, and the best-rated sentences are selected. This results in collections of the N most 'meaningful' sentences, in the order in which they appeared in the source text – we will call these *excerpts*. An excerpt can be used to give readers an idea of what the longer text is about, or it can be used as input into a process to produce a more coherent abstract.

It has been argued for almost 40 years that it is possible to automatically create excerpts which meet basic information compression needs (Luhn, 1958). Since then, different measurements for the importance of a sentence have been suggested, in particular stochastic measurements for the significance of key words or phrases (Luhn, 1958; Zechner, 1995). Other research, starting with (Edmundson, 1969), stressed the importance of heuristics for the location of the candidate sentence in the source text (Baxendale, 1958) and for the occurrence of cue phrases (Paice and Jones, 1993; Johnson et al., 1993).

Single heuristics tend to work well on documents that resemble each other in style and content. For the more robust creation of excerpts, combinations of these heuristics can be used. The crucial question is how to combine the different heuristics. In the past, the relative usefulness of single methods had to be balanced manually. Kupiec et al. (1995) use supervised learning to automatically adjust feature weights, using a corpus of research papers and corresponding summaries.

Humans have good intuition about what makes a sentence 'abstract-worthy', i.e. suitable for inclusion in a summary. Abstract-worthiness is a high-level quality, comprising notions such as semantic content, relative importance and appropriateness for representing the contents of a document. For the automatic evaluation of the quality of machine generated excerpts, one has to find an operational approximation to this subjective notion of abstract-worthiness, i.e. a definition of a desired result. We will call the criteria of what constitutes success the *gold standard*, and the set of sentences that fulfill these criteria the *gold standard sentences*. Apart from evaluation, a gold standard is also needed for supervised learning.

In Kupiec et al. (1995), a gold standard sentence is a sentence in the source text that is matched with a summary sentence on the basis of semantic and syntactic similarity. In their corpus of 188 engineering papers with summaries written by professional abstractors, 79% of sentences occurred in both summary and source text with at most minor modifications.

However, our collection of papers, whose abstracts were written by the authors themselves, shows a significant difference: these abstracts have significantly fewer alignable sentences (31.7%). This does not mean that there are fewer abstract-worthy sentences in the source text. We used a simple (labour-intensive) way of defining this alternative gold standard, viz. asking a human judge to identify additional abstract-worthy sentences in the source text.

Our main question was whether Kupiec et al.'s methodology could be used for our kind of gold standard sentences also, and if there was a fundamental difference in extraction performance between sentences in both gold standards or between documents with higher or lower alignment. We also conducted an experiment to see how additional training material would influence the statistical model.

The remainder of this paper is organized as follows: in the next section, we summarize Kupiec et al.'s method and results. Then, we describe our data and discuss the results from three experiments with different evaluation strategies and training material. Differences between our and Kupiec et al.'s data with respect to the alignability of document and summary sentences, and consequences thereof are considered in the discussion.

2 Sentence selection as classification

In Kupiec et al.'s experiment, the gold standard sentences are those summary sentences that can be aligned with sentences in the source texts. Once the alignment has been carried out, the system tries to determine the characteristic properties of aligned sentences according to a number of features, viz. presence of particular cue phrases, location in the text, sentence length, occurrence of thematic words, and occurrence of proper names. Each document sentence receives scores for each of the features, resulting in an estimate for the sentence's probability to also occur in the summary. This probability is calculated as follows:

$$P(s \in S | F_1, \ldots, F_k) \approx \frac{P(s \in S) \prod_{j=1}^{k} P(F_j | s \in S)}{\prod_{j=1}^{k} P(F_j)}$$

$P(s \in S|F_1, \ldots, F_k)$: Probability that sentence s in the source text is included in summary S, given its feature values;

$P(s \in S)$: compression rate (constant);

$P(F_j|\, s \in S)$: probability of feature-value pair occurring in a sentence which is in the summary;

$P(F_j)$: probability that the feature-value pair occurs unconditionally;

k: number of feature-value pairs;

F_j: j-th feature-value pair.

Assuming statistical independence of the features, $P(F_j|s \in S)$ and $P(F_j)$ can be estimated from the corpus.

Evaluation relies on cross-validation. The model is trained on a training set of documents, leaving one document out at a time (the current test document). The model is then used to extract candidate sentences from the test document, allowing evaluation of precision (sentences selected correctly over total number of sentences selected) and recall (sentences selected correctly over alignable sentences in summary). Since from any given test text as many sentences are selected as there are alignable sentences in the summary, precision and recall are always the same.

Kupiec et al. reports that precision of the individual heuristics ranges between 20–33%; the highest cumulative result (44%) was achieved using paragraph, fixed phrases and length cut-off features.

3 Our experiment

3.1 Data and gold standards

Our corpus is a collection of 202 papers from different areas of computational linguistics, with summaries written by the authors.[1] The average length of the summaries is 4.7 sentences; the average length of the documents 210 sentences.

We semi-automatically marked up the following structural information: title, summary, headings, paragraph structure and sentences. Tables, equations, figures, captions, references and cross references were removed and replaced by place holders.

We decided to use two gold standards:

- **Gold standard A: Alignment.** Gold standard sentences are those occurring in both author summary and source text, in line with Kupiec et al.'s gold standard.
- **Gold standard B: Human Judgement.** Gold standard sentences are non-alignable source text sentences which a human judge identified as relevant, i.e. indicative of the contents of the source text. Exactly how many human-selected sentence candidates were chosen was the human judge's decision.

Alignment between summary and document sentences was assisted by a simple surface similarity measure (longest common subsequence of non-stop-list words). Final alignment was decided by a human judge. The criterion was similarity of semantic contents of the compared sentences. The following sentence pair illustrates a *direct match*:

> **Summary:** *In understanding a reference, an agent determines his confidence in its adequacy as a means of identifying the referent.*
>
> **Document:** *An agent understands a reference once he is confident in the adequacy of its (inferred) plan as a means of identifying the referent.*

Our data show an important difference with Kupiec et al.'s data: we have significantly lower alignment rates. Only 17.8% of the summary sentences in our corpus could be automatically aligned with a document sentence with a certain degree of reliability, and only 3% of all summary sentences are identical matches with document sentences.

We created three different sets of training material:

- **Training set 1:** The 40 documents with the highest rate of overlap; 84% of the summary sentences could be semi-automatically aligned with a document sentence.
- **Training set 2:** 42 documents from the year 1994 were arbitrarily chosen out of the remaining 163 documents and semi-automatically aligned. They showed a much lower rate of overlap; only 36% of summary sentences could be mapped into a document sentence.
- **Training set 3:** 42 documents from the year 1995 were arbitrarily chosen out of the remaining documents and semi-automatically aligned. Again, the overlap was rather low: 42%.
- **Training set 123:** Conjunction of training sets 1, 2 and 3. The average document length is 194

[1] The corpus was drawn from the computation and language archive (`http://xxx.lanl.gov/cmp-lg`), converted from LaTeX source into HTML in order to extract raw text and minimal structure automatically, then transformed into our SGML format with a perl script, and manually corrected. Data collection took place collaboratively with Byron Georgantopolous.

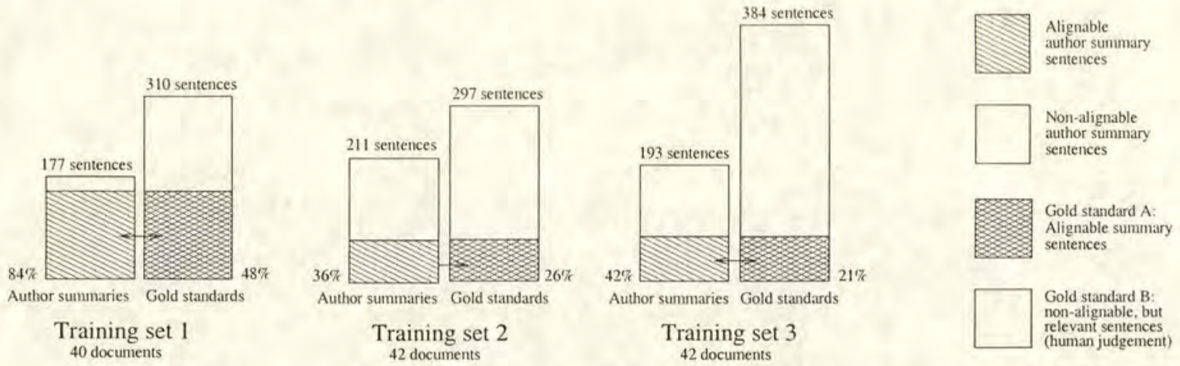


Figure 1: Composition of gold standards for training sets

sentences; the average summary length is 4.7 sentences.

A human judge provided a mark-up of additional abstract-worthy sentences for these 3 training sets (124 documents). The remaining 78 documents remain as unseen test data. Figure 1 shows the composition of gold standards for our training sets. Gold standard sentences for training set 1 consist of an approximately balanced mixture of aligned and human-selected candidates, whereas training set 2 contains three times as many human-selected as aligned gold standard sentences, training set 3 even four times as many. Each document in training set 1 is associated with an average of 7.75 gold standard sentences (A+B), compared to an average of 7.07 gold standard sentences in training set 2, and an average of 9.14 gold standard sentences in training set 3.

3.2 Heuristics

We employed 5 different heuristics: 4 of the methods used by Kupiec et al. (1995), viz. cue phrase method, location method, sentence length method and thematic word method, and another well-known method in the literature, viz. title method.

1. Cue phrase method: The cue phrase method seeks to filter out meta-discourse from subject matter. We advocate the cue phrase method as our main method because of the additional 'rhetorical' context these meta-linguistic markers make available. This context of the extracted sentences – along with their propositional content – can be used to generate more flexible abstracts.

We use a list of 1670 negative and positive cues and indicator phrases or formulaic expressions, 707 of which occur in our training sets. For simplicity and efficiency, these cue phrases are fixed strings.

Our cue phrase list was manually created by a cycle of inspection of extracted sentences, identification of as yet unaccounted-for expressions, addition of these expressions to the cue phrase list, and possibly inclusion of overlooked abstract-worthy sentences in the gold standard. Cue phrases were manually classified into 5 classes, which we expected to correspond to the likelihood of a sentence containing the given cue to be included in the summary: a score of -1 means 'very unlikely'; $+3$ means 'very likely to be included in a summary'.[2] We found it useful to assist the decision process with corpus frequencies. For each cue phrase, we compiled its relative frequency in the gold standard sentences and in the overall corpus. If a cue phrase proved general (i.e. it had a high relative corpus frequency) and distinctive (i.e. it had a high frequency within the gold standard sentences), we gave it a high score, and included other phrases that are syntactically and semantically similar to it into the cue list. We scanned the data and found the following tendencies:

- Certain communicative verbs are typically used to describe the overall goals; they occur frequently in the gold-standard sentences (*argue*, *propose*, *develop* and *attempt*). Others are predominantly used for describing communicative sub-goals (detailed steps and sub-arguments) and should therefore be in a different equivalence class (*prove*, *show* and *conclude*). Within the class of communicative verbs, tense and mode seem to be relevant for abstract-worthiness. Verbs in past tense or present perfect (as used in the conclusion) are more likely to refer to global achievements/goals, and thus to be included in the summary. In the body of the text, present and

[2] We experimented with larger and smaller numbers of classes, but obtained best results with the 5-way distinction.

future forms tend to be used to introduce subtasks.

- Genre specific nominal phrases like *this paper* are more distinctive when they occur at the beginning of the sentence (as an approximation to subject/topic position) than their non-subject counterparts.

- Explicit summarisation markers like *in sum*, *concluding* did occur frequently, but quite unexpectedly almost always in combination with communicative sub-tasks. They were therefore less useful at signalling abstract-worthy material.

Sentences in the source text are matched against expressions in the list. Matching sentences are classified into the corresponding class, and sentences not containing cue phrases are classified as 'neutral' (score 0). Sentences with competing cue phrases are classified as members of the class with the higher numerical score, unless one of the competing classes is negative.

Sentences occurring directly after headings like *Introduction* or *Results* are valuable indicators of the general subject area of papers. Even though one might argue that this property should be handled within the location method, we perceive this information as meta-linguistic (and thus logically belonging to the cue phrase method). Thus, scores for these sentences receive a prior score of +2 ('likely to occur in a summary').

In a later section, we show how this method performs on unseen data of the same kind (viz. texts in the genre of computational linguistics research papers of about ~6–8 pages long). Even though the cue phrase method is well tuned to these data, we are aware that the list of phrases we collected might not generalize to other genres. Some kind of automation seems desirable to assist a possible adaptation.

2. Location method. Paragraphs at the start and end of a document are more likely to contain material that is useful for a summary, as papers are organized hierarchically. Paragraphs are also organized hierarchically, with crucial information at the beginning and the end of paragraphs. Therefore, sentences in document peripheral paragraphs should be good candidates, and even more so if they occur in the periphery of the paragraph.

Our algorithm assigns non-zero values only to sentences which are in document peripheral sections; sentences in the middle of the document receive a 0 score. The algorithm is sensitive to prototypical headings (*Introduction*); if such headings cannot be found, it uses a fixed range of paragraphs (first 7 and last 3 paragraphs). Within these document peripheral paragraphs, the values 'i_f' and 'm' (for paragraph initial-or-final and paragraph medial sentences, respectively) are assigned.

3. Sentence Length method. All sentences under a certain length (current threshold: 15 tokens including punctuation) receive a 0 score, all sentences above the threshold a 1 score.

Kupiec et al. mention this method as useful for filtering out captions, titles and headings. In our experiment, this was not necessary as our format encodes headings and titles as such, and captions are removed. As expected, it turns out that the sentence length method is our least effective method.

4. Thematic word method. This method tries to identify key words that are characteristic for the contents of the document. It concentrates on non-stop-list words which occur frequently in the document, but rarely in the overall collection. In theory, sentences containing (clusters of) such thematic words should be characteristic for the document. We use a standard term-frequency*inverse-document-frequency (tf*idf) method:

$$score(w) = f_{loc} * log(\frac{100*N}{f_{glob}})$$

f_{loc}: frequency of word w in document
f_{glob}: number of documents containing word w
N: number of documents in collection

The 10 top-scoring words are chosen as thematic words; sentence scores are then computed as a weighted count of thematic word in sentence, meaned by sentence length. The 40 top-rated sentences get score 1, all others 0.

5. Title method. Words occurring in the title are good candidates for document specific concepts. The title method score of a sentence is the mean frequency of title word occurrences (excluding stop-list words). The 18 top-scoring sentences receive the value 1, all other sentences 0. We also experimented with taking words occurring in all headings into account (these words were scored according to the tf*idf method) but received better results for title words only.

3.3 Results

Training and evaluation took place as in Kupiec et al.'s experiment. As a baseline we chose sentences from the beginning of the source text, which obtained a recall and precision of 28.0% on training set 123. This from-top baseline (which is also used by Kupiec et al.) is a more conservative baseline

	Indiv.	Cumul.
Method 1 (cue)	55.2	55.2
Method 2 (location)	32.1	65.3
Method 3 (length)	28.9	66.3
Method 4 (tf*idf)	17.1	66.5
Method 5 (title)	21.7	68.4
Baseline	28.0	

Figure 2: First experiment: Impact of individual heuristics; training set 123, gold standards A+B

	Seen	Unseen
Cue Phrase Method	60.9	54.9
Heuristics Combination	71.6	65.3
Baseline	29.1	

Figure 3: First Experiment: Difference between unseen and seen data; training set 3, gold standards A+B

than random order: it is more difficult to beat, as prototypical document structure places a high percentage of relevant information in the beginning.

3.3.1 First experiment

Figure 2 summarizes the contribution of the individual methods.[3] Using the cue phrase method (method 1) is clearly the strongest single heuristic. Note that the contribution of a method cannot be judged by the individual precision/recall for that method. For example, the sentence length method (method 3) with a recall and precision over the baseline contributes hardly anything to the end result, whereas the title method (method 5), which is below the baseline if regarded individually, performs much better in combination with methods 1 and 2 than method 3 does (67.3% for heuristics 1, 2 and 5; not to be seen from this table). The reason for this is the relative independence of the methods. If method 5 identifies a successful candidate, it is less likely that this candidate has also been identified by method 1 or 2. Method 4 (tf*idf) decreased results slightly in some of the experiments, but not in the experiments with our final/largest training set 123 where it led to a (non-significant) increase.

We also checked how much precision and recall decrease for unseen data. This decrease applies only to the cue phrase method, because the other heuristics are fixed and would not change by seeing more data. After the manual mark-up of gold standard sentences and additions to the cue phrase list for training set 3, we treated training set 3 as if it was unseen: we used only those 1423 cue phrases for extraction that were compiled from training set 1 and 2. A comparison of this 'unseen' result to the end result (Figure 3) shows that our cue phrase list, even though hand-crafted, is robust and general enough for our purposes; it generalizes reasonably well to texts of a similar kind.

	comb	cue	base
TS 1	66.1	49.0	29.6
TS 2	62.2	54.5	24.9
TS 3	71.6	60.9	29.1
TS 123	68.4	55.2	28.0

Figure 4: First experiment: Baseline, best single heuristic and combination; gold standards A+B

	Evaluation strategy					
	Gold standard A			Gold standard B		
TS	comb	cue	base	comb	cue	base
1	36.9	27.5	21.4	45.3	30.4	10.8
2	25.0	18.4	9.2	53.8	47.9	20.3
3	27.1	13.5	13.5	64.3	54.4	25.7
123	31.6	23.2	16.3	57.2	46.7	20.4

Figure 5: Second experiment: Impact of type of gold standard

Figure 4 shows mean precision and recall for our different training sets for three different extraction methods: a combination of all 5 methods ('comb.'); the best single heuristic ('cue'); and the baseline ('base'). We used both gold standards A+B. These results reconfirm the usefulness of Kupiec et al.'s method of heuristic combination. The method increases precision for the best method by around 20%. It is worth pointing out that this method produces very short excerpts, with compressions as high as 2–5%, and with a precision equal to the recall. Thus this is a different task from producing long excerpts, e.g. with a compression of 25%, as usually reported in the literature. Using this compression, we achieved a recall of 96.0% (gold standard A), 98.0% (gold standard B) and 97.3% (gold standards A+B) for training set 123. For comparison, Kupiec et al. report a 85% recall.

3.3.2 Second experiment

In order to see how the different gold standards contribute to the results, we used only one gold standard (A or B) at a time for training and for extraction. Figure 5 summarizes the results.

Looking at Gold standard A, we see that training set 1 is the only training set which obtains a recall

[3] All figures in tables are precision percentages.

that is comparable to Kupiec et al.'s. Incidentally, training set 1 is also the only training set that is comparable to Kupiec et al.'s data with respect to alignability. The bad performance of training set 2 and 3 under evaluation with gold standard A is not surprising, as there are too few aligned gold standard sentences to train on: 50% of the documents in these training sets contain no or only one aligned sentence.

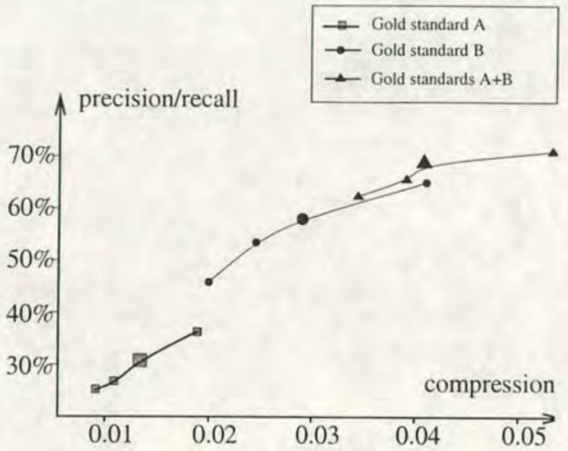


Figure 6: Second experiment: Impact of type of gold standard on precision and recall, as a function of compression

Overall, performance seems to correspond to the ratio of gold standard sentences to source text sentences, i.e. the compression of the task.[4] The dependency between precision/recall and compression is depicted in Figure 6. Taking both gold standards into account increases performance considerably compared to either of the gold standards alone, because of the lower compression. As we don't have training sets with exactly the same number of gold standard A and B sentences, we cannot directly compare the performance, but the graph is suggestive of a similar behaviour of both gold standards. The results for training set 123 fall between the results of the individual training sets (symbolized by the large data points).

From this second experiment we conclude that for our task, there is no difference between gold standard A and B. The crucial factor that precision and recall depends on is the compression of the task.

		Extraction			
	TS	1	2	3	123
Training	1	66.1	61.2	69.7	66.3
	2	65.8	62.2	69.5	66.0
	3	65.1	62.9	71.6	66.1
	123	66.4	62.9	70.8	68.4

Figure 7: Third experiment: Impact of training material on precision and recall; gold standards A+B

3.3.3 Third experiment

In order to evaluate the impact of the training material on precision and recall, we computed each possible pair of training and evaluation material (cf. figure 7).

In this experiment, all documents of the training set are used to train the model; this model is then evaluated against each document in the test set, and the mean precision and recall is reported. Importantly, in this experiment none of the other documents in the test set is used for training.

These experiments show a surprising uniformity within test sets: overall extraction results for each training set are very similar. Training on different data does not change the statistical model much. In most cases, extraction for each training set worked best when the model was trained on the training set itself, rather than on more data. Thus, the difference in results between individual training sets is not an effect of data sparseness at the level of heuristics combination.

We conclude from this third experiment that improvement in the overall results can primarily be achieved by improving single heuristics, and not by providing more training data for our simple statistical model.

4 Discussion

Comparing our experiment to Kupiec et al.'s the most obvious difference is the difference in data.

Our texts are likely to be more heterogeneous, coming from areas of computational linguistics with different methodologies and thus having an argumentative, experimental, or implementational orientation. Also, as they are not journal articles, they are not heavily edited. There is also less of a prototypical article structure in computational linguistics than in experimental disciplines like chemical

[4] The difference in performance between training sets in the first experiment is thus probably mainly attributable to differences in compression between the training sets.

engineering. This makes our texts more difficult to extract from.

The major difference, however, is that we use summaries which are not written by trained abstractors, but by the authors themselves. In only around 20% of documents in our original corpus, sentence selection had been used as a method for summary generation, whereas professional abstractors rely more heavily and systematically on sentences in the source text when creating their abstracts.

Using aligned sentences as gold standard has two main advantages. First, it makes the definition of the gold standard less labour intensive. Second, it provides a higher degree of objectivity. It is a much simpler task for a human judge to decide if two sentences convey the same propositional content, than to decide if a sentence is qualified for inclusion in a summary or not.

However, using alignment as the sole definition for gold standard implies that a sentence is only a good extraction candidate if its equivalent occurs in the summary, an assumption we believe to be too restrictive. Document sentences other than the aligned ones might have been similar in quality to the chosen sentences, but will be trained on as a negative example with Kupiec et al.'s method. Kupiec et al. also recognize that there is not only one optimal excerpt, and mention Rath et al.'s (1961) research which implies that the agreement between human judges is rather low. We argue that it makes sense to complement aligned sentences with manually determined supplementary candidates. This is not solely motivated by the data we work with but also by the fact that we envisage a different task than Kupiec et al. (who use the excerpts as indicative abstracts). We see the extraction of a set of sentences as an intermediate step towards the eventual generation of more flexible and coherent abstracts of variable length. For this task, a whole range of sentences other than just the summary sentences might qualify as good candidates for further processing.[5] One important subgoal is the reconstruction of approximated document structure (cf. rhetorical structure, as defined in RST (Mann et al., 1992)). One of the reasons why we concentrated on cue phrases was that we believe that cue phrases are an obvious and easily accessible source of rhetorical information.

Another important question was if there were other properties following from the main difference between our training sets, alignability. Are documents with a high degree of alignability *inherently* more suitable for abstraction by our algorithm? It might be suspected that alignability is correlated with a better internal structure of the papers, but our experiments suggest that, for the purpose of sentence extraction, this is either not the case or not relevant. Our results show that our training sets 1, 2 and 3 behave very similarly under evaluation taking aligned gold standards *or* human-selected gold standards into account. The only definite factor influencing the results was the compression rate. With respect to the quality of abstracts, this implies that the strategy which authors use for summary generation – be it sentence selection or complete regeneration of the summary from semantic representation – is a matter of authorial choice and not an indicator of style, text quality, or any aspect that our extraction program is particularly sensitive to. This means that Kupiec et al.'s method of classificatory sentence selection is not restricted to texts which have high-quality summaries created by human abstractors. We claim that adding human-selected gold standards will be useful for generation of more flexible and coherent abstracts, than training on just a fixed number of author-provided summary sentences would allow.

5 Conclusions

We have replicated Kupiec et al.'s experiment for automatic sentence extraction using several independent heuristics and supervised learning. The summaries for our documents were not written by professional abstractors, but by the authors themselves. As a result, our data demonstrated considerably lower overlap between sentences in the summary and sentences in the main text. We used an alternative evaluation that mixed aligned sentences with other good candidates for extraction, as identified by a human judge.

We obtained a 68.4% recall and precision on our text material, compared to a 28.0% baseline and a best individual method of 55.2%. Combining individually weaker methods results in an increase of around 20% of the best method, in line with Kupiec et al.'s results. This shows the usefulness of Kupiec et al.'s methodology for a different type of data and evaluation strategy. We found that there was no difference in performance between our evaluation strategies (alignment or human judgement), apart from external constraints on the task like the compression rate. We also show that increased training did not significantly improve the sentence extraction results, and conclude that there is more room for improvement in the extraction methods themselves.

With respect to our ultimate goal of generating of

[5] This is mirrored by the fact that in our gold standards, the number of human-selected sentence candidates outweighed aligned sentences by far.

higher quality abstracts (more coherent, more flexible variable-length abstracts), we argue that the use of human-selected extraction candidates is advantageous to the task. Our favourite heuristic includes meta-linguistic cue phrases, because they can be used to detect rhetorical structure in the document, and because they provide a rhetorical context for each extracted sentence in addition to its propositional content.

6 Acknowledgements

The authors would like to thank Chris Brew, Janet Hitzeman and two anonymous referees for comments on earlier drafts of the paper. The first author is supported by an EPSRC studentship.

References

Baxendale, P. B. (1958). Man-made index for technical literature – an experiment. *IBM Journal of Research and Development*, 2(4):354–361.

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285.

Johnson, F. C., Paice, C. D., Black, W. J., and Neal, A. P. (1993). The application of linguistic processing to automatic abstract generation. *Journal of Document and Text Management*, 1(3):215–241.

Kupiec, J., Pedersen, J. O., and Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference*, pages 68–73.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Mann, W. C., Matthiesen, C. M. I. M., and Thompson, S. A. (1992). Rhetorical structure theory and text analysis. In Mann, W. C. and Thompson, S. A., editors, *Discourse description*. J. Benjamins Pub. Co., Amsterdam.

Paice, C. D. and Jones, A. P. (1993). The identification of important concepts in highly structured technical papers. In *Proceedings of the Sixteenth Annual International ACM-SIGIR conference on research and development in IR, Association for Computing Machinery, Special Interest Group Information Retrieval.*

Zechner, K. (1995). Automatic text abstracting by selecting relevant passages. Master's thesis, Centre for Cognitive Science, University of Edinburgh.

E.2. Teufel and Moens (1998)

Sentence extraction and rhetorical classification for flexible abstracts

Simone Teufel
Centre for Cognitive Science
and Language Technology Group
University of Edinburgh
S.Teufel@ed.ac.uk

Marc Moens
Language Technology Group
University of Edinburgh
M.Moens@ed.ac.uk

Abstract

Knowledge about the discourse-level structure of a scientific article is useful for flexible and sub-domain independent automatic abstraction. We are interested in the automatic identification of content units ("argumentative entities" or "rhetorical roles") such as GOAL OR PROBLEM STATEMENT, CONCLUSIONS and RESULTS in the source text. In this paper, we present an extension of Kupiec et al.'s methodology for trainable statistical sentence extraction (1995). Our extension additionally classifies the extracted sentences according to their rhetorical status; because it takes only low-level properties of the sentence into account and uses no external knowledge sources other than meta-level linguistic ones, it achieves robustness and partial domain-independence. This is necessary, because in the domain we are working in (conference papers in various sub-domains of computational linguistics), the document structure with respect to these content units is unpredictable compared to certain other domains, e.g. experimental sciences.

1 Introduction

Until recently, the world of research publications was heavily paper-oriented. One of the roles of abstracts of research articles was to act as a decision tool: on the basis of the abstract a researcher could decide whether the paper was worth a visit to the library, whether it was worth a letter to the author requesting a copy of the full paper, whether it was worth postponing finishing one's own paper, etc.

For reasons of consistency (and copyright) these abstracts often were not the abstracts produced by the original authors, but by professional abstractors, and written according to agreed guidelines and recommendations. In particular, because the abstract was a pointer to a paper not immediately available, the abstracts had to be self-contained: the reader should be able to grasp the main goals and achievements of the full paper without needing the source text for clarification.

To agree on the required level of detail in an abstract, abstracts were aimed at the partially informed reader (Kircz, 1991): someone who knows enough about the field to understand basic methodology and general goals but does not have enough of an overview of previous work and may not know where a certain article is situated in the field or how articles are related to each other. For a novice reader an abstract would be too terse. For experienced researchers the abstract would provide unnecessary detail; for them the name of the author and the title of article would often be sufficient to decide whether the article was worth the trip to the library.

Research articles are now increasingly being made available on-line. Indeed, the goal of automated summarization presupposes the on-line existence and machine-readability of the full paper. An abstract will therefore no longer be a tool for the partially informed reader to decide whether to put extra effort into obtaining the full article, since the full article is available with no extra effort. Instead, automatically generated abstracts will be playing different roles, for example helping researchers find their way through large collections of research papers returned in reply to a search request.

This has a number of consequences for the abstracts generated. For one, it is no longer just the partially informed reader who needs to be catered for; the novice may also find abstracts a useful tool to find her way through large collections of papers, as indeed may the experienced reader, although they will all need different abstracts. Also, if the full papers are a reply to a query, something more is known about the users' information needs, and this could be taken into account when formulating the abstract. Indeed, in some situations it may no longer be necessary for the abstract to be fully self-contained, since relevant excerpts of the full paper might be displayed simultaneously with the abstract (or as a part of the abstract).

We see this kind of flexible automated generation of abstracts as our long-term goal. The problematic side-effect is that very little is known about what such abstracts should look like. Most of the information on good abstracts deal with the world of paper, not with the use of on-line research publications. More generally, very little research has been done so far on how people deal with research papers on screen as opposed to paper (O'Hara and Sellen, 1997).

That means that we cannot take existing guidelines on how to produce balanced, informative, concise abstracts at face value; we will need to fall back on a different set of intuitions as to what constitutes a good abstract. We take the rhetorical structure of research papers and their abstracts as our starting point.

The communicative function of a scientific article is one of a narrow range of things: for example, to report on own research, to provide an overview of a research area or to review a piece of work. This small range of communicative functions has lead to well-established prototypical rhetorical divisions in scientific articles, like Introduction, Purpose, Experimental Design, Results, Discussion, Conclusions, etc. Especially in experimental subgenres (like experimental psychology) these rhetorical divisions are very clearly marked in section headers (cf. Kintsch and van Dijk 1978). In non-experimental papers, the rhetorical divisions are still there, but implicitly so.

The communicative function of an abstract is similarly one of a narrow range of things: it can be an indicative abstract, reporting the topic of the full paper, or an informative abstract, also reporting main findings and conclusions (Cremmins 1996, Rowley 1982). Like in the case of research articles, the possible communicative function of abstracts has lead to a consensus on their rhetorical building blocks, like General Background, Specific Problem tackled by full paper, Main Results, Recommendations, etc.

We believe that automatically tracking the global level rhetorical structure of an article is possible. We seek to identify excerpts from the full article which fall into one of the typical rhetorical divisions, and to use these to compose an abstract. The abstract consists of a global rhetorical frame whose fillers are the rhetorically justified excerpts of the full paper.

This is different from methods which extract sentences based on heuristics about the *contents* of sentences (e.g. using a tf/idf model or lexical cohesion) and linking these together into an abstract (Salton et al., 1994). We want to extract meaningful sentences *with* information about their rhetorical role in the full article. We then want to use the sentence (or the information in the sentence) together with the rhetorical information to compose a rhetorically structured abstract. We believe this will provide greater flexibility in the generated abstract. Our goal of modular rhetorical abstracts is thus more related to the "structured abstracts" described in Broer (1971) and Rennie et al. (1991).

The rhetorical frame of the abstract will allow us generate longer or shorter abstracts where needed, e.g. by adding or suppressing BACKGROUND information, depending on whether users have been identified as novices or experienced readers. Other rhetorical roles for which only low-probability evidence was found in the source document, can be likewise pruned until the right length is reached. Our approach also promises to be adaptive to the existence of different content units in text, and thus also error tolerant to weaknesses in the information extraction process.

In the rest of this paper we discuss the rhetorical building blocks or "content units" (Tibbo 1992) that we have identified as important for the generation of flexible, modular abstracts. We then report on experiments to train a system to automatically detect meaningful sentences in the full paper together with their rhetorical role.

2 Rhetorical Structure of Abstracts

Although we argued in the previous section that most guidelines for abstracts cannot be taken at face value when designing a high-level framework for online abstracts, there is ample information in the literature which can be used to inform decisions about the rhetorical structure of abstracts.

Some of it is very specific. For example, for Archives of dermatology Arndt (1992) offers the following, highly domain-specific, subdivision: BACKGROUND/DESIGN, RESULTS, CONCLUSIONS (CLINICAL), BACKGROUND/OBSERVATIONS/CONCLUSIONS (OBSERVATIONAL). We want a more generic structure.

The description of the components of abstracts in Liddy (1991) is based on professional abstractors' intuitions and a corpus of abstracts. Nevertheless, it is very specific to the domain of empirical abstracts. In subsequent work (Francis and Liddy, 1991) less robust results were reported for abstracts of theoretical papers. Since the rhetorical structure of abstracts we want to develop has to be less domain dependent, we cannot directly use her results. Sometimes there are also technical reasons why we could not adopt suggestions in the literature. Liddy defines an abstract's constituent components in a recursive fashion (i.e. they can be contained within other components), and most of them span parts of sentences rather than whole sentences. Neither of these options are available with our machine-learning technology, and so her suggestions could not be taken on board without change.

In our search for non-recursive rhetorical building blocks for abstracts, we found a general consensus on most content units of informative abstracts. Whereas indicative abstract report on the topic of the source paper, informative abstracts also report findings and results (although not discussion or interpretation). Manning (1990) argues that informative abstracts are not a miniature version of the full paper in the sense of offering "a paraphrase of every rhetorical section" of the source article. Instead, most authors agree that informative abstracts should mention the Purpose or Problem of the full paper, Scope or Methodology, Results, and Conclusions or Recommendations (Borko and Chatman, 1963; American National Standards Institute, 1979; Day, 1995; Rowley, 1982; Cremmins, 1996).

There is more disagreement about "peripheral" content units, such as Related work, Background, Incidental findings, Future work and Data.

Of particular interest is the content unit Background. According to Alley (1996), Background is a useful content unit in an abstract if it is restricted to being the first sentence of the abstract. Other authors (Rowley 1982, Cremmins 1996) recommend not to include any background information at all. We believe that background information is potentially important, especially for self-contained abstracts and for abstracts for novice readers.

There is similar disagreement over the content unit Previous work. Cremmins (1996) states that this should not be included in an abstract unless the studies are replications or evaluations of the earlier work. However, depending on the information need, previous work might actually have been central to the query the user started off with, and we therefore want to include it in our modular abstract.

Apart from identifying the constituent units of abstracts, another important issue is the order in which these constituent units should be displayed. For this, the ANSI (1979) guidelines distinguish between *findings-oriented* abstracts and *purpose-oriented* abstracts. In findings-oriented abstracts major results, conclusions and possibly recommendations are given in a topical first sentence, followed by sentences that contain further results, conclusions or recommendations, and supporting details on methodology, purpose or scope). Purpose-oriented abstracts lead with the source article's purpose.

Primary content units	Feature extensions
Background	—
Topic/Aboutness	—
Related Work	[own/others] [prev/pres/future]
Purpose/Problem	[global/local] [new/weakness_prev_method] [concrete_contribution]
Solution/Method	[global/local] [limitation/advantage] [properties] [use] [assumpt]
Result	—
Conclusion/Claim	[global/local] [hypothesis/recom]

Table 1: Taxonomy of rhetorical roles

Our resulting hierarchy of rhetorical building blocks can be summarised as in Table 2, where the right column shows a feature-based subdivision of the primary content units (left column).

One important aspect is that we distinguish between plans/argumentation steps that describe research phases (e.g. *I first classified the phonemes*), and moves that describe textual steps (e.g. *we will then present*; not displayed here). We intend to filter texual information (e.g. sentences containing *This paper is organized as follows* or *in chapter 3, we will present our results*) separately, in order to support the information searching process. Also, we will determine the role *Type of Work*, i.e. Problem/Purpose[concrete_contribution] (implementation, theory, experiment, evaluation...) with pattern matching techniques, because they can often be recognized easily (*we present a method for...*).

What we were trying to annotate and subsequently automatically learn is a high-level quality, namely "which rhetorical role, if any, is expressed by the following sentence?" This can be a difficult question but we found that humans find it easier to answer that question for a given sentence than to answer the related question "is this sentence a good candidate for inclusion in an abstract?"

The task to decide on a certain role is nevertheless not easy. Often, the rhetorical role of a statement is dependent on the local context of the line of argument. For example, if the authors mention a weakness of their solution, it might be classified as Solution[limit] or as Purpose/Problem[local], depending on whether they solve the problem in that paper or not. Or, if somebody mentions in the conclusions that a certain problem does not occur, this might be viewed as a description of a tackled problem or as an advantage of the solution.

The following sentence with its judgements illustrates the type of markup:

> Repeating the argument of Section 2, we conclude that a construction grammar that encodes the formal language [EQN] is at least an order of magnitude more compact that any lexicalized grammar that encodes this language.
> **Conclusion/Claim**

We allowed for multiple annotation in ambiguous cases, but still faced problems, most of them having to do with the large unit of annotation (a whole sentence as opposed to a clause or even smaller unit) enforced by our annotation and machine-learning technology. The following sentence shows a difficult case, an ambiguous role (SOLUTION or PURPOSE/PROBLEM in a case where the sentence covers more than one role.[1]

> We also examined how utterance type related to topic shift and found that few interruptions introduced a new topic.
>
> **(Solution/Method[local] OR**
> **Purpose/Problem[local]) AND**
> **Claim/Conclusion[local]**

3 The classification experiment

The basic procedure for the sentence extraction experiment (to extract "meaningful" sentences with information about their rhetorical contribution) is to annotate a text with the sentences that we would want the system to extract (sentences worth being included in the abstract), and to additionally annotate these sentences with the role we would like the system to associate with them. The system is then trained to learn what the significant features of these sentences are.

Over the years there have been many suggestions as to which features contribute to making a sentence "meaningful" or abstract-worthy, such as its location in the source text (Baxendale, 1958; Edmundson, 1969), the presence of keywords or phrases (Paice, 1981), or the stochastic significance of cue phrases in the sentence (Luhn, 1958). The problem is that none of these features by themselves suffice, and weighted combinations need to be found.

Kupiec et al. (1995) describes supervised learning techniques to adjust the weights of some predefined features in a data-driven way. Kupiec et al.'s gold standard of abstract-worthy sentences is defined as the set of sentences in the source text that "align" with a sentence in the abstract—i.e. sentences that show sufficient semantic and syntactic similarity with a sentence in the abstract. The underlying reason is that a sentence in the source text is abstract-worthy if professional abstractors used it or parts of it when producing their abstract. In Kupiec et al.'s corpus of 188 engineering papers with summaries written by professional abstractors, 79% of sentences in the abstract also occurred in the source text with at most minor modifications.

Kupiec et al. (1995) then try to determine the characteristic properties of abstract-worthy sentences according to a number of features, viz. presence of particular cue phrases, location in the text, sentence length, occurrence of thematic words, and occurrence of proper names. Each document sentence receives scores for each of the features, resulting in an estimate for the sentence's probability to also occur in the summary. This probability is calculated for each feature value as a combination of the probability of the feature-value pair occurring in a sentence which is in the summary (successful case) and the probability that the feature-value pair occurs unconditionally.

Evaluation of the training relies on cross-validation: the model is trained on a training set of documents, leaving one document out at a time (the current test document). The model is then used to extract candidate sentences from the test document, allowing evaluation of precision (sentences selected correctly over total number of sentences selected) and recall (sentences selected correctly over gold standard sentences). Since from any given test text as many sentences are selected as there are gold standard sentences, precision and recall are always the same. Kupiec et al. report that precision of the individual heuristics ranges between 20–33%; the highest cumulative result (44%) was achieved using paragraph, fixed phrases and length cut-off features.

Teufel and Moens (1997) reimplemented this technique for a different collection of articles, whose abstracts displayed a substantially lower level of alignable sentences (31.7%). Because of this low alignment, they annotated source texts with additional abstract-worthy sentences, as selected by a human judge. They showed that Kupiec et al.'s methodology can be applied to this different task and data, and that aligned gold standards show the same training behaviour as human-judged gold standards. On their training corpus, with a fine-tuned and very elaborate cue phrase list, they report a combined precision and recall value of close to 70%.

[1] The annotation AND is used for concatenations of roles within one sentence.

3.1 Data and annotation of gold standards

Our corpus (also used in (Teufel and Moens, 1997)) is a collection of 201 articles (mainly conference papers) from different areas of computational linguistics.[2] The average length of the summaries is 4.7 sentences; the average length of the documents is 210 sentences. The corpus contains about 800,000 words. The following structural information is marked up: title, summary, headings, paragraph structure and sentences. Tables, equations, figures, captions, references and cross references were removed and replaced by place holders.

Although all the papers in this collection deal with computational linguistics, the corpus displays huge variation as to sub-domain. The largest part (about 45%) are articles describing implementational work, but there are about 25% theoretical-linguistic articles, with an argumentative tenet, about 10% overview and general-opinion articles and 20% experimental papers (reporting corpus studies or psycholinguistic experiments). The writing style varies from extremely informal to formal. About a third of the articles were not written (or subsequently edited) by native speakers of English. The papers have no homogenous discourse structure. We assume that most of the articles had been accepted for publication, although this cannot be relied on as the archive is unmoderated. Abstracts were not provided by professional abstractors, but by the authors themselves.

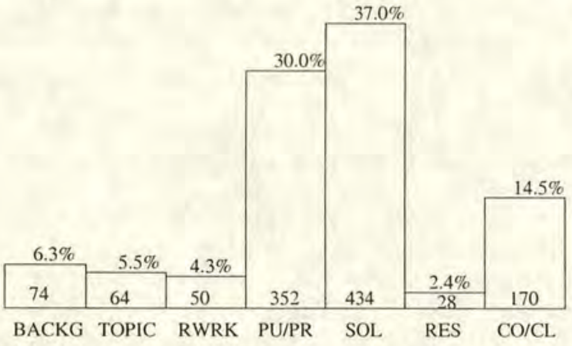


Figure 1: Composition of rhetorical roles for training set

We conducted an informal test to see if we could identify overall properties of the discourse level structure in the summaries. We applied our annotation scheme for rhetorical roles to the 123 summaries in our training corpus. The authors, in contrast to professional abstractors, had not used a prototypical scheme to write their abstracts. We could not confirm any pattern; abstracts vary widely, including everything from Background to Properties of Solution to Problem and vice versa, in almost any possible permutation. Some abstracts are extremely short, and many of the abstracts are not self-contained, and would thus be difficult to understand for the partially informed reader. In short, they are not the kind of abstracts we want to produce with our method.

Our next step was to manually markup up our gold standard sentences (i.e. the combination of aligned and human-judged abstract-worthy sentences) in the training set (123 documents) with their rhetorical role. The remaining 78 documents remain as unseen test data. Figure 1 shows the composition of 1172 instances of rhetorical roles for the 948 gold standard sentences in our training set. Our annotation scheme allowed for ambiguous mark-up of a sentence which was the case in 232 sentences (24%).

3.2 Heuristics

We employed 6 different heuristics: 4 of the methods used by Kupiec et al. (1995), viz. cue phrase method, location method, sentence length method and thematic word method, and 2 additional ones: title method and the cue phrase semantics method.

3.2.1 Cue phrase method: The cue phrase method uses linguistic text properties to identify meta-discourse (as opposed to subject matter) in a text. We use a list consisting of 1725 indicator phrases or formulaic expressions, like communicative verbs and research and argumentation related phrases. The largest part of these phrases is positive.

Our cue phrase list was manually created by a cycle of inspection of extracted sentences and addition to the list. Cue phrases were manually classified into 5 quality classes according to their occurrence frequencies within the gold standard sentences. Thus, the scores mirror the likelihood of a sentence containing the given cue to be included in the summary: a score of -1 means 'very unlikely'; $+3$ means 'very likely to be included in a summary'. For example, the phrase *we have given an account* received a high score of $+3$, whereas *supported by grant* receives a negative score.

3.2.2 Cue phrase semantics method: We also associated each indicator phrase with a semantic

[2] The corpus was drawn from the computation and language archive (`http://xxx.lanl.gov/cmp-lg`), converted from LaTeX source into HTML in order to extract raw text and minimal structure automatically, then transformed into SGML format with a `perl` script, and manually corrected. Data collection took place collaboratively with Byron Georgantopolous.

class, i.e. a guess as to which rhetorical role it is normally associated with. The positive example above would receive the score P (for problem). This feature returns 16 different scores, namely our rhetorical roles and the most common ambiguity/confusion classes between them. The phrase *our account* receives a SP score, for ambiguous between solution and problem. Again, this score was gained by corpus frequencies.

3.2.3 Location method: This feature distinguishes peripheral sentences in the document and within each paragraph, assuming a hierarchical organization of documents and of paragraphs. The algorithm is sensitive to prototypical headings (*Introduction*); if such headings cannot be found, it uses a fixed range of paragraphs (first 7 and last 3 paragraphs). Document final and initial areas receive different values, but paragraph initial and final sentences are collapsed into one group.

3.2.4 Sentence length method: All sentences under a certain length (current threshold: 15 tokens including punctuation) receive a 0 score, all sentences above the threshold a 1 score.

3.2.5 Thematic word method: This method is a variation of the tf/idf method, which tries to identify key words that are characteristic for the contents of the document, namely those of a medium range frequency relative to the overall collection. The 10 top-scoring words according to the tf/idf method are chosen as thematic words; sentence scores are then computed as a weighted count of thematic word in sentence, meaned by sentence length. The 40 top-rated sentences obtain score 1, all others 0.

3.2.6. Title method: Words occurring in the title are good candidates for document specific concepts. The title method score of a sentence is the mean frequency of title word occurrences (excluding stop-list words). The 18 top-scoring sentences receive the value 1, all other sentences 0. We also experimented with taking words occurring in all headings into account (these words were scored according to the tf/idf method) but received better results for title words only.

3.3 Classifier

The probability that a certain sentence is associated with a rhetorical role is calculated as follows:

$$P(s \in R_i | F_1, \ldots, F_k) \approx \frac{P(s \in R_i) \prod_{j=1}^{k} P(F_j | s \in R_i)}{\prod_{j=1}^{k} P(F_j)}$$

$P(s \in R_i | F_1, \ldots, F_k)$: Probability that sentence s in the source text is assigned the rhetorical role R_i (and thus included in summary S), given its feature values;

$P(s \in R_i)$: probability that role R_i occurs unconditionally;

$P(F_j | s \in R_i)$: probability of feature-value pair occurring in a sentence which has rhetorical role R_i;

$P(F_j)$: probability that the feature-value pair occurs unconditionally;

k: number of feature-value pairs;

F_j: j-th feature-value pair.

Assuming statistical independence of the features, $P(F_j | s \in R_i)$ and $P(F_j)$ can be estimated from the corpus for each F_j and each R_i.

These formulae are adaptions of Kupiec et al.'s estimations for the probability that a given sentence is contained in the abstract. We divide this probability into a vector of probabilities associated with each rhetorical role for a sentence. The sum over our vector thus results in the probability that Kupiec et al. use.

3.4 Extraction algorithms

Training (counting of frequencies) results in a vector of probabilities for each sentence in a document, as in Fig. 2. We employed three alternative extraction algorithms to interpret this huge mass of data.

Algorithm 1 extracts the best sentences for each rhetorical role, i.e. it reads the matrix in Fig. 2 vertically for each role and identifies the sentence with the highest probability for each role. If there are conflicts between roles, the role with the highest probability is assigned to that sentence. In our example, the algorithm would choose sentences 1 (RWRK), 235 (RES), 2 (PU/PR) and 0 (BACKG), in that order.

Algorithm 2 extracts a given number of sentences with highest probability for any role (and asigns it that role); there is no guarantee as with algorithm 1 that every role will be contained in the resulting set of sentences. The algorithm would assign RWRK to sentence 0, 1 and 2, and RES to 235.

Algorithm 3 extracts a given number of sentences with the highest overall probability (sum over all cells for one sentence). The difference to algorithm 2 lies in certain sentences that might not have a high *peak* of probabilities for a given role but yet a competitive *sum* of probabilities. This algorithm corresponds to the original one used by Kupiec et al. (with the exception that it can assign a rhetorical role). The algorithm would extract the same roles as algorithm 2, but prefer 235 over sentence 2

	BACKG	TOPIC	RWRK	PU/PR	SOLU	RES	CL/CO
Algorithm 1: Total extracted	130	126	132	124	126	127	123
Total extracted	5	2	0	225	690	0	26

Table 2: Number of extracted sentences per role

even though 2's highest probability is lower than the highest probability of 235, just because the sum is greater in 235.

0.3e-9	0.1e-9	0.2e-9	0.6e-10	0.9e-9	0.7e-10	0.9e-10	0
BACKG	TOPIC	RWRK	PU/PR	SOL	RES	CO/CL	

0.3e-12	0.9e-14	0.3e-14	0.6e-13	0.9e-12	0.7e-14	0.1e-17	1
BACKG	TOPIC	RWRK	PU/PR	SOL	RES	CO/CL	

0.6e-14	0.9e-11	0.3e-7	0.5e-10	0.4e-10	0.7e-8	0.1e-10	2
BACKG	TOPIC	RWRK	PU/PR	SOL	RES	CO/CL	

. . .

0.4e-8	0.9e-10	0.3e-9	0.6e-8	0.5e-10	0.7e-9	0.1e-10	235
BACKG	TOPIC	RWRK	PU/PR	SOL	RES	CO/CL	

Figure 2: Probability vectors for document sentences No. 0, 1, 2 and 235

3.5 Results

Evaluation is based on crossvalidation like in Kupiec et al.'s experiment. We measure two kinds of errors which correspond to the two kinds of tasks we are collapsing:

Extraction of abstract-worthy sentences ("extraction success"). The algorithms seperate sentences which carry *any* rhetorical roles (as defined by our annotation scheme) from irrelevant sentences (by far the larger part of the text). Failure to perform this task leads to the inclusion of irrelevant material in the abstracts (false positives), or the exclusion of relevant material from the abstract (false negatives).

Identification of the correct rhetorical role for a correctly extracted sentence ("classification success")[3]. We expect to do less well on this task, because the classification is inherently vague and ambiguous between certain classes (confusion classes). It is also possible that subsequent information extraction (and possibly reasoning) steps might still make sense of the sentence, given that we can be sure/should be able to be sure that they are abstract-worthy sentences.

In the following, we present success rates for the two kinds of tasks.[4]. In Table 3, 32% classification precision means that about a third of *all* extracted sentences were given the right role (as well as being correctly extracted). The classification success in relation to correctly extracted sentences would be higher (more than 50% for algorithm 2), but this has no practical meaning because we cannot guess which of the extracted sentences were correct.

	Extraction	Classification
	success	success
Algorithm 1	61%p./58%r.	22 %
Algorithm 2	62.6%	32 %
Algorithm 3	62.3%	32 %

Table 3: Success rates for the extraction algorithms

Algorithm 1 has a lower success rate for classification because it is *forced* to choose one sentence per role, even for those roles that occur with a low everall frequency (e.g. for TOPIC). The relatively high success of *Background* (which is rather low-frequency as well) is due to its prototypical location as first sentence in a text – an easily learnable property.). As a result, the probabilities of the sentences chosen for these roles are too. Algorithm 2 and 3 are very similar to each other in performance. However, even though they are better at classifying roles, their success builds on the fact that they operate mainly with the high frequency roles; most classifications are these roles (*safe guesses*, cf. Table 2). This table shows that algorithm 1 classifies too many sentences with infrequent roles, and algorithm 2 classifies too many sentences with high frequent roles. A better algorithm would have to combine information from both sources (success per role and success individually).

4 Discussion

The extracts in themselves are not good enough to serve as abstracts. This is mainly due to the fact that the classification is not reliable enough yet, due

[3] By definition, classification success implies extraction success.

[4] Due to the construction of algorithm 2 and 3, precision and recall are the same numerical value. 65% extraction success/precision means that 65% of all extracted sentences were gold standard sentences; 55% extraction success/recall means that 55% of all gold standard sentences were extracted.

to the suboptimality of the heuristics, the statistical classification and the extraction algorithms in their current implementation. Nevertheless, many of the classification errors are not as grave as they might look because many of the misclassifications were exactly of the kind where humans have problems as well. SOLUTION and PROBLEM are two roles that happen to be confused in certain situations, particularly where the status of the sentence is not linguistically marked. In that case, only inference on the argumentation in the paper as a whole might help ("is this a step towards the main goal or a goal in itself"). However, the rhetorical classification is useful at least insofar as it robustly identifies the potentially meaningful sentences, along with an indication as to their possible role – it thus identifies the candidate sentences that are *worth* further, deeper, more resource intensive analysis. Subsequent modules in the abstracting process must then decide how indispensable a given role is, and if the ambiguity needs to be resolved. Obviously, how much a given role is needed depends on the structure of the abstract frame and alternative information resources, e.g. textual cues (like *in chapter 4, we will give the goal statement*) or other extraction results.

We find the results encouraging in that they support our hypothesis that rhetorical document structure can be approximated by low-level properties of the sentence. We are aware that the list of phrases we collected might not generalize to other genres. We are experimenting with maximum entropy methods for determining cue phrases and possible rhetorical anchors for them automatically.

An important advantage of our method is its robustness towards the wide variety of subgenres present in our collection. As the argumentation steps we anchored in our annotation scheme are generic, we expect performance to be stable over different subgenres. We are currently experimenting with a subdivision of our corpus into subdomains. Our hope is that this study will reconfirm our hypothesis that there is enough overlap in the linguistic realizations of rhetorical roles to keep the classification stable.

5 Conclusion

We have argued that rhetorical classification of extracted material is a useful subtask for the production of a new kind of abstract that can be tailored in length and proportion to users' expertise and specific information needs.

Our goal is to recognize abstract-worthy sentences with respect to rhetorical structure, and to perform a simultaneous classification of these sentences into a set of predefined, generic rhetorical roles. We have presented a robust method which uses machine learning techniques to deduce rhetorical roles from lower-level properties of sentences.

The results are encouraging; one of our algorithms determines two out of three marked-up gold standards sentences in our training text and additionally associates the right role for every third sentence it extracts. Even though this level of precision in the classification is not reliable enough to use the extracts without further processing, our results seem to point to the general feasibility of a shallow processing of discourse structure.

References

Alley, M. (1996). *The craft of scientific writing*. Prentice-Hall, Englewood Cliffs, N.J. Third edition.

American National Standards Institute, I. (1979). *American National Standard for Writing Abstracts*. American National Standards Institute, Inc., New York. ANSI Z39.14.1979.

Arndt, K. A. (1992). The informative abstract. *Archives of Dermatology*, 128(1):101.

Baxendale, P. (1958). Man-made index for technical literature – an experiment. *IBM journal on research and development*, 2(4):354–361.

Borko, H. and Chatman, S. (1963). Criteria for acceptable abstracts: a survey of abstractors' instructions. *American Documentation*, 14(2):149–160.

Broer, J. W. (1971). Abstracts in block diagram form. *IEEE Transactions on Engineering Writing and Speech*, 14(2):64–67. ISA, 72-1626.

Cremmins, E. T. (1996). *The art of abstracting*. Information Resources Press.

Day, R. A. (1995). *How to write and publish a scientific paper*. Cambridge University Press, Cambridge. 4th edition.

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285.

Francis, H. and Liddy, E. D. (1991). Structured representation of theoretical abstracts: Implications for user interface design. In Dillon, M., editor, *Interfaces for Information Retrieval and Online Systems: The state of the art*. Greenwood Press.

Kintsch, W. and van Dijk, T. A. (1987). Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394.

Kircz, J. G. (1991). The rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of documentation*, 47(4):354–372.

Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference*, pages 68–73.

Liddy, E. D. (1991). The discourse-level structure of empirical abstracts: an exploratory study. *Information processing and management*, 27(1):55–81.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Manning, A. (1990). Abstracts in relation to larger and smaller discourse structures. *Journal of technical writing and communication*, 20(4):369–390.

O'Hara, K. and Sellen, A. (1997). A comparison of reading paper and on-line documents. In *Proceedings of CHI-97*.

Paice, C. D. (1981). The automatic generation of literary abstracts: an approach based on the identification of self-indicating phrases. In Norman, O. R., Robertson, S. E., van Rijsbergen, C. J., and Williams, P. W., editors, *Information Retrieval Research*, page 112. Butterworth, London.

Rennie, D. and Glass, R. M. (1991). Structuring abstracts to make them more informative. *Journal of the American Mecial Association*, 266(1).

Rowley, J. (1982). *Abstracting and indexing.* Bingley, London.

Salton, G., Allan, J., Buckley, C., and Singhal, A. (1994). Automatic analysis, theme generation, and summarisation of machine readable texts. *Science*, 264:1421–1426.

Teufel, S. and Moens, M. (1997). Sentence extraction as a classification task. In Mani, I. and Maybury, M. T., editors, *Proceedings of the workshop on Intelligent Scalable Text Summarization, in association with ACL/EACL-97.*

Tibbo, H. R. (1992). Abstracting across the disciplines: A content analysis of abstracts from the natural sciences, and the humanities with implications for abstracting standards and online information retrieval. *Library and Information Science Research*, 14(1):31–56.

E.3. Teufel (1998)

Meta-discourse markers and problem-structuring in scientific articles

Simone Teufel
Centre for Cognitive Science
University of Edinburgh
S.Teufel@ed.ac.uk

Abstract

Knowledge about the argumentative structure of scientific articles can, amongst other things, be used to improve automatic abstracts. We argue that the argumentative structure of scientific discourse can be automatically detected because reasoning about problems, research tasks and solutions follows predictable patterns.

Certain phrases explicitly mark the rhetorical status (communicative function) of sentences with respect to the global argumentative goal. Examples for such meta-discourse markers are *"in this paper, we have presented ..."* or *"however, their method fails to"*. We report on work in progress about recognizing such meta-comments automatically in research articles from two disciplines: computational linguistics and medicine (cardiology).

1 Motivation

We are interested in a formal description of the document structure of scientific articles from different disciplines. Such a description could be of practical use for many applications in document management; our specific motivation for detecting document structure is quality improvement in automatic abstracting.

Researchers in the field of automatic abstracting largely agree that it is currently not technically feasible to create automatic abstracts based on full text understanding [Spärck Jones(1994)]. As a result, many researchers have turned to sentence extraction [Kupiec et. al. (1995); Brandow et al. (1995); Hovy and Lin (1997)]. Sentence extraction, which does not involve any deep analysis, has the huge advantage of being robust with respect to individual writing style, discipline and text type (genre). Instead of producing abstracts, this results produces only *extracts*: document surrogates consisting of a number of sentences selected verbatim from the original text.

We consider a concrete document retrieval (DR) scenario in which a researcher wants to select one or more scientific articles from a large scientific database (or even from the Internet) for further inspection. The main task for the searcher is *relevance decision* for each paper: she needs to decide whether or not to spend more time on a paper (read or skim-read it), depending on how useful it presumably is to her current information needs. Traditional sentence extracts can be used as rough-and-ready relevance indicators for this task, but they are not doing a great job at representing the contents of the original document: searchers often get the wrong idea about what the text is about. Much of this has to do with the fact that extracts are typically incoherent texts, consisting of potentially unrelated sentences which have been taken out of their context. Crucially, extracts have no handle at revealing the text's logical and semantic organisation.

More sophisticated, user-tailored abstracts could help the searcher make a fast, informed relevance decision by taking factors like the searcher's expertise and current information need into account. If the searcher is dealing with research she knows well, her information needs might be quite concrete: during the process of writing her own paper she might want to find research which supports her own claims, find out if there are contradictory results to hers in the literature, or compare her results to those of researchers using a similar methodology. A different information need arises when she wants to gain an overview of a new research area – as an only "partially informed user" in this field [Kircz(1991)] she will need to find out about specific research goals, the names of the researchers who have contributed the main research ideas in a given time period, along with information of methodology and results in this research field.

There are new functions these abstracts could fulfil. In order to make an informed relevance decision, the searcher needs to judge differences and similarities between papers, e.g. how a given paper relates to similar papers with respect to research goals or methodology, so that she can place the research described in a given paper in the larger picture of the field, a function we call navigation *between* research articles. A similar operation is navigation *within* a paper, which supports searchers in non-linear reading and allows them to find relevant information faster, e.g. numerical results.

We believe that a document surrogate that aims at supporting such functions should characterize research articles in terms of the problems, research tasks and solutions/methodology presented in the specific paper, but it should also represent other researchers' problems, research tasks and solutions mentioned in the paper. Our long-term goal is to automatically reconstruct this problem-solution structure from unrestricted text for the searcher in the form of a *problem-structured* abstract.

But how can we find problems, research questions, research tasks and solutions in text without fully understanding the text? We take sentence extraction as a starting point due to its inherent robustness. Using additional information about each sentence, viz. their rhetorical status with respect to the entire paper, we are in a better position to perform shallow, but guided information extraction, in order to find the information units we are interested in.

In the next section we introduce the level of document structure we are talking about, and the kind of meta-comments we employ in discovering it. In the rest of the paper, we report on ongoing work on automatically filtering meta-comments from annotated and unannotated text. Finding meta-comments in text is an attractive task because it would allow for the automatic adaptation of sytems using such phrases to new domains.

2 Discourse structure and argumentation in scientific articles

Discourse linguistic theory suggests that texts serving a common purpose among a community of users eventually take on a predictable structure of presentation [Kintsch and van Dijk(1978)] – and scientific articles certainly serve a well-defined communicative purpose: they "present, retell and refer to the results of specific research" [Salager-Meyer(1992)]. Particularly in the life and experimental sciences, a rigid building plan for research articles has evolved over the years, where rhetorical divisions tend to be very clearly marked in section headers. Prototypical rhetorical divisions include *Introduction, Purpose, Experimental Design, Results, Discussion, Conclusions.* One of the reasons for this rigidly-defined structure seems to be that the scientific community in these fields has more or less agreed on how to do research: methodologies and evaluation methods are long-lived research entities that do not change often.

One of the corpora we are using is a good example of such texts. It consists of 129 articles in cardiology, taken from the *American Heart Journal*, which have a fixed structure with respect to rhetorical divisions and section headers. The other corpus, in contrast, consisting of 123 (mostly conference) articles in computational linguistics (CL), displays an heterogeneous mixture of methodologies and traditions of presentation one would expect in an interdisciplinary field. Most of the articles cover more than one single discipline, but as a rough estimate one can say that about 45% of the articles in the collection are predominantly technical in style, describing implementations (i.e. engineering solutions); about 25% report on research in theoretical linguistics, with an argumentative tenet; the remaining 30% are empirical (psycholinguistic or psychological experiments or corpus studies). Even though most of the articles have an introduction and conclusions (sometimes occurring under headers with different names), and almost all of them cite previous work, the presentation of the problem and the methodology/solution are idiosyncratic and depend on individual writing style. Very few of the headers in the computational linguistics articles correspond to prototypical rhetorical divisions; the rest contain content specific terminology (cf. Figure 1 which compares relative frequencies of headers for the two corpora).

Computational Linguistics		Cardiology	
Header	**Freq.**	**Header**	**Freq.**
Introduction	85%	Introduction	100%
Conclusion	46%	Results	94%
Conclusions	22%	Discussion	94%
Acknowledgments	17%	Methods	92%
Discussion	12%	Tables	79%
Results	11%	Statistics	40%
Experimental		Patients	29%
Results	9%	Limitations	28%
Related Work	7%	Conclusions	25%
Implementation	7%	Statistical	
Evaluation	7%	Analysis	22%
Example	6%	Conclusion	17%
Background	6%	Patient	
Summary	4%	Characteristics	9%

Figure 1: Highest-frequency headers (with relative occurrence frequencies)

Because the type of research reported in the computational linguistics corpus differs so much, the description of document structure we were looking for had to be flexible enough to generalize over differences in presentation, yet formal enough for the extraction of the information units which are useful for automatic abstracts. We base our model of argumentation in scientific articles on Swales' (1990) CARS model ("Create a Research Space"). Swales' claim is that the main communicative goal of an author of a research article is to convince readers (potential reviewers) that the research described in the paper constitutes an actual contribution to science, in order to have the paper reviewed positively and thus published; this is the case whether or not the paper tries to give the impression that it reports research in an objective, disinterested way. In order to successfully present their case, authors argue in a goal-directed and prototypical way about problem-solving activities — their own and other researchers'. Swales identified prototypical rhetorical building plans of introduction sections, along with linguistics surface cues that signal rhetorical moves. Examples for rhetorical moves include the claim that the paper addresses a *new* problem or, if it is a *well-known* problem, then the presented solution has to be better than that of other researchers.

A first analysis of the corpora confirmed many of the rhetorical building blocks suggested by Swales. We adapted Swales' scheme to the one shown in Figure 8 (at the end of this paper). In the medical corpus, almost all of the moves we found were of type I (*Explicit mention*); the rigid document structure seems to have replaced much of the "argument about problem-solving activities" (types II-V) for which we found ample evidence in the computational linguistics corpus.

We are interested in identifying these moves automatically and shallowly in text, and we believe that this is technically feasible, because the stereotypical, predictable overall structure of the argument can be exploited in doing so.

3 Meta-discourse markers

In this paper we focus on the linguistic realisations of rhetorical moves, i.e. the surfacy signals of the argumentative status of a given sentence. Consider the strings in boldface on the right hand side of Figure 8. They cover the activities of reporting about research (reporting and presenting verbs), the problem-solving process (problems, solutions, tasks); they also include other semantic links like necessity, causality and contrast. Due to explicit or implicit argumentation, many of these strings are evaluative (*"efficient, elegant, innovative, insightful*" vs. *"impossible, inadequate, inconclusive, insufficient"*). We call them *meta-comments* because they talk *about* information units, as opposed to being subject matter (scientific content). Such meta-comments are very frequent in our collection.

Our meta-comments are similar Paice's (1981) *indicator phrases* (he was the first to use such phrases for abstracting); they are less similar to *cue phrases*, the discourse markers usually studied in discourse analysis, because they are *not* sentence connectives (with some exceptions), and because they are typically considerably longer and far more varied.

The fact that the computational linguistics texts stem from an unmoderated medium (i.e. they are neither chosen for publication nor edited by a central authority), means that there were no external restrictions on how exactly to say things. Authors use idiosyncratic style, which can vary from formal to informal. There are meta-comments that tend to get used in a fixed, formulaic way, but interestingly, we observed a wide range of linguistic variability with respect to the realization of some of the meta-comments (whereas their semantics is usually perfectly unambiguous). This effect makes the meta-comments in this text type interesting linguistic objects to study.

We observed that there are certain meta-comments which are restricted to certain moves, mostly the evaluative and contrastive phrases and the phrases occurring in moves of type I (*Explicit mention*). Others occur frequently across moves, particularly general argumentative phrases and relevance markers such as *"important"*, *"in this paper, we"*. Argumentative phrases like *"we argue that"* appeared with solution and problem-related moves almost as often as with claims and conclusions. These phrases seemed to be the ones that were most formulaic/fixed across texts.

Our goal is to automatically find meta-markers and associate them with rhetorical moves (where this makes sense). In the next section, we report on a first experiment in that direction.

4 Our experiment

If it is true that most meta-comments are formulaic, recurring expressions, then frequency information should help us separate meta-comments from domain-specific parts of the sentence. Those strings which occur rarely in the corpus will most likely be domain-specific and will appear low on frequency listings of strings, whereas meta-comments should appear high on the lists.

We also used a lexicon of 433 *lexical seeds*. Lexical seeds are words which are semantically related to the activities of reporting, problem-solving, argumenting or evaluating, or expressions of deixis (*"we. . . "*) or other textual cues (e.g. literature references in text were marked up using the symbol [REF], which is a signal for mentions of other researchers' solutions, tasks or problems).

The computational linguistics corpus was drawn from the computation and language archive (`http://xxx.lanl.gov/cmp-lg`) and contains 123 articles; the 129 articles of the cardiology corpus appeared in the *American Heart Journal*. The medical corpus is smaller in overall size (436,909 words vs. 654,477; 14,770 sentences vs. 23,072).

For the computational linguistics corpus, we additionally had a collection of 948 sentences that had been identified as relevant by a human annotator in a prior experiment [Teufel and Moens(1997)]. A human judge annotated these with respect to the 23 rhetorical moves introduced in Figure 8.

4.1 Filtering

First, we compiled the two corpora into those bigrams, trigrams, 4-grams, 5-grams and 6-grams which did not cross sentence boundaries. We worked with a short stop-list compiled from the corpus (60 highest-frequent words) from which we had excluded those which we expected to be important in an argumentative domain, e.g. personal and demonstrative pronouns. We lowercased all words and counted punctuation (including brackets) as a full word.

We then filtered the n-grams through our seed lexicon, i.e. we retained those expressions which contain at least one of the words of the seed-lexicon (or a morphological variant of it). We also compiled and counted n-grams for the 948 computational linguistics target sentences, to see how similar these phrases from the annotated parts were to the filtered or unfiltered bigrams from the entire corpus.

Before filtering		After filtering	
354	[ref] , [ref] ,	118	in this paper ,
301	[ref]) .	111	in ([cref])
297	, [ref] , [ref]	110	can be used to
178	[cref]) .	106	in figure [cref] .
144	[ref] , [ref] .	100	([cref]) ,
139	on the other hand	99	on the basis of
134	for example , the	83	shown in figure [cref]
118	in this paper ,	83	in section [cref] .
116	the other hand ,	75	this paper , we
111	in ([cref])	75	in section [cref] ,
110	can be used to	71	it is possible to

Figure 2: 4-grams in entire CL corpus

The list of 4-grams for the computational linguistics corpus shows a typical picture of the outcome of this process (Figure 2). Before filtering, the frequent corpus n-grams contain general comments and expressions like *"for example"* but content specific expressions are already filtered out. (Very rarely, there are some content-specific phrases like "***natural language***" in the lists — this is due to the fact that the corpus, even though interdisciplinary in nature, is composed of papers focusing on language.) After filtering, the meta-comments on the lists have two properties: (a) they are frequent (b) they contain lexical items that *could* be related to argumentation about problems, research tasks, solutions — in the computational linguistics corpus, these conditions seem to be enough to produce expressions that are good candidates for meta-comments. Unfortunately, condition (a) means that a large number of meta-comments were lost, because they were of low frequency.

Target sentences			
49	in this paper ,	10	can be used to
37	this paper , we	9	in this paper is
25	[ref] , [ref] ,	7	[ref] , [ref] .
22	, [ref] , [ref]	7	described in this paper
18	in this paper we	7	this paper , i

Figure 3: 4-grams in CL target sentences

How similar are the lists for annotated text and entire corpus in the computational linguistics domain? Table 3 shows that they look very similar apart from minor differences, e.g. the fact that the list gained from annotated data contains more [CREF] items (internal cross reference like *"section [CREF]"*) which tend to appear frequently in the sentences where authors state the organisation of the paper. The expressions used in such organisation statements are typically formulaic and recurrent, but not many of these sentences were considered relevant when the 948 target sentences were determined.

Before filtering		After filtering	
120	$p < 0.05$)	85	on the basis of
102	$p < 0.01$)	66	of this study was
93	left ventricular ejection fraction	64	in this study ,
86	$p < 0.001$)	58	this study was to
86	$p < 0.0001$)	57	patients with heart failure
85	on the basis of	45	the purpose of this
72	at the time of	44	there were no significant
66	of this study was	40	new york heart association
64	in this study ,	39	purpose of this study
59	95 % confidence interval	35	there was no significant
58	this study was to	32	were no significant differences
55	coronary artery disease .	31	p = not significant
49	acute myocardial infarction .	30	has been shown to

Figure 4: 4-grams in medical corpus

The medical corpus shows significant differences (cf. Figure 4). Firstly, unfiltered bigrams do not separate content matter from meta-discourse. In these lists, there are phrases pertaining to statistical analyses ("$p < 0.01$") and several domain-specific phrases. Filtering (right hand side of Figure 4) forces the few meta-comments that *are* being used to the top of the list; they are linguistically invariant. For instance, "***study***" seems the only acceptable expression used for the current research, whereas the range is much wider in the other corpus (*"paper, article, study, work, research...* ", and all the meta-comment candidates in the top part of the list belonged to one single rhetorical move, viz. *Explicit Mention of the Research Task* (EX-T in Figure 8).

A certain amount of noise has been introduced through the seed-lexicon because word senses were not disambiguated: *"failure"* was included in the seed lexicon to indicate mentions of failure of other researchers' solution. Because this term obviously has the different meaning of *"heart failure"* in the cardiology context, the desired distinction between subject matter strings and meta-comments got lost; similarly *"New York"* was included because the word *"new"* could potentially point to novel approaches. This might mean that it is necessary to use different stop-lists and/or seed lexicons for different domains.

As we have seen before, associating meta-comments with rhetorical moves is a more difficult task for some meta-comments than for others. We tried to anchor the probable rhetorical move of a phrase in the lexical seed it contains, a simplification we are forced to make due to the small amount of annotated text we have available (which is reflected in the low numbers). We are thus in the process of working on a larger scale annotation.

We used the human judgements to count how often each word contained in the target sentences appears with a certain rhetorical move. If the difference in frequency between the best-scoring moves for that word was large enough, we assumed it was a good indicator for the highest-scoring move, and we then manually associated the given rhetorical move with the word if it was contained in the seed lexicon, or to semantically similar seeds. For example, seeds that are the most likely associated with the OWN SOLUTION BETTER (53 examples of this move in the target sentences) were *"than"* (39), *"better"* (36), *"results"* (21), *"method"* (19), *"using"* (15), *"significantly"* (14), *" outperforms"* (12) and *"more"* (12). Filtered meta-comments are then assigned the rhetorical move predicted by the first seed they contain. Figure 5 shows the meta-comments filtered for the seed *"better"* from both corpora. In the medical corpus, there is less argument about methodology/solutions, and as a result the phrases found are unfortunately *not* meta-comments but contain medical terminology.

Computational Linguistics		Cardiology	
64	better than	11	a better
50	a better	7	better than
23	better than the	6	to better
20	much better	6	significantly better
19	the better	6	better in
19	is better	5	better left
16	significantly better	5	better left ventricular
16	be better	5	and better
13	better performance	4	better preserved
11	are better	4	better in smokers
10	better the	3	to be somewhat better tolerated
9	significantly better than	3	failure symptoms in spite of better
6	better suited	3	better preserved left ventricular systolic function
6	better than that of	3	better in smokers than in nonsmokers

Figure 5: Potential meta-comments with *"better"* from both corpora

Also, we observed that it is not easy to predict the optimal length of a certain meta-comment which is indicative of a certain rhetorical move. For moves containing *other problems/solutions/tasks* the very short string "*[REF]*" is contained in all successful meta-comments, whereas for explicit mention of research goals, the maximal length 6 of meta-comments which we chose for these experiments might even be too short. As another example, for the STEP move (*"goal is achieved by doing solution"*), the best indicator we found was *"in order to"*.

4.2 Evaluation

We evaluate the quality of these automatically generated meta-comment lists by comparing them to a manually created meta-comment list used by a summarisation system, cf. [Teufel and Moens(1998)]. The performance of the system – with the two different meta-comment lists – is measured by precision and recall values of co-selection with the target extracts defined by human annotators mentioned earlier. The summarisation process consists of two consecutive steps, sentence extraction and rhetorical classification, and uses other heuristics like location and term frequency.

The summarisation system requires a list of meta-comments of arbitrary length, containing a *quality score* for each phrase which estimates how predictive these phrases are in pointing to extract-worthy sentences, and the most likely rhetorical label that sentences with this meta-comment will receive.

Quality Class	Rhetorical Move	Meta-comment
2	–	paper ,
3	–	this paper presents a
2	STEP	in order to
3	–	in this paper , we will
2	EX-T	in this paper we have
1	CO-S	unlike [ref]
1	EX-T	this paper is to
3	EX-T	in this paper , we describe
2	EX-P	paper is
2	–	paper we
1	EX-T	this paper has presented
1	EX-T	, we propose a method
1	–	in passage ([cref]
1	–	, we argue that
1	–	argue that
2	EX-T	method for
1	EX-C	we show that the
1	EX-T	show how
1	–	property and the number
1	EX-T	the advantages of
1	EX-E	the wall street journal
1	NEC-S-T	the importance of
1	CO-C	however , we
1	EX-T	be used to

Figure 6: Extracts from automatic list of meta-comments

We automatically built the meta-comment list in Figure 6 (containing 318 entries). We started from all n-grams compiled from the target sentences and took the following heuristics into account: Firstly, choose phrases with a high ratio of target frequency to corpus frequency, because these are *indicative* phrases. Set the quality value accordingly. Secondly, exclude phrases with a low overall frequency, or decrease their quality score, because including/overestimating them might construct a model that is over-fitted to the data. Thirdly, associate each phrase with its most likely rhetorical move, by taking the ratio between frequency in each rhetorical class and the frequency of the rhetorical label itself into account. If below a certain threshold, don't associate any move at all (e.g. *"paper ,"* in Figure 6).

The manual meta-comment list, in contrast, was compiled in an extremely labour intensive manner and refined over the months. It consists of 1791 meta-comments (some of which are much longer than the maximum of 6 words that the automatic phrases consisted of), along with their most plausible rhetorical moves and quality scores.

	Manual	Automatic
Extraction	66.4%	52.5%
Classification	66.3%	54.3%

Figure 7: Evaluation results: precison and recall of co-selection

As Figure 7 shows, using the automatic meta-comment list instead of the manually created one decreased the summarizer's performance from 66.4% to 52.5% precision and recall for extraction, and from 66.3% to 54.3% precision and recall for classification.

5 Discussion and further work

The evaluation indicates that the quality of the automatic meta-comment list is not yet high enough to replace the manual list in our summarization system. However, a look at the automatic list itself shows that, even though it is far from perfect, most of the high-frequent strings found are plausible candidates for meta-comments (or parts of meta-comments). In most cases, subject matter can be successfully filtered out.

We regard the simple method for automatic meta-comment identification discussed in this paper as a baseline for further work. We have simplified the problem of finding meta-comments enormously by only considering verbatim substrings. By doing so, we have ignored discontinuous strings, morphological variation and statistical interaction between the words in the string. In addition, the phrases considered so far have been short, and we have not collected many of them, because we wanted to rely only on the ones with reasonably high frequencies.

The biggest problem for now is that highly indicative, but infrequent meta-comments cannot be found with a simple method like ours. Therefore, it is essential to perform some generalization over similar phrases. One way would be the automatic clustering of similar concepts, e.g. to find out that *"argue"* and *"show"* are presentational verbs with similar semantics. Another idea would be to allow for more flexible patterns consisting of short n-grams and other words, in order to skip over intervening words like adjectives and adverbs. This might avoid the data sparseness problems encountered with the longer n-grams.

6 Summary

We have presented some baseline results from our ongoing work concerning automatic filtering of meta-comments (indicator phrases) from scientific papers. Meta-comments can vary considerably from one domain to another, as the comparison of the two corpora we considered shows. In the computational linguistics articles, authors argue explicitly about problems, solutions and research tasks, whereas this is less the case in the medical domain, where meta-comments are less frequent and more formulaic.

We have shown that lists of meta-comments acquired in a simple automatic process can be used to automatically identify a shallow document structure in scientific text, albeit with a certain quality loss when compared to manually constructed resources.

7 Acknowledgements

Data collection of the computational linguistics corpus took place collaboratively with Byron Georgantopolous. We thank Kathy McKeown, Vasileios Hatzivassiloglou and Olga Merport from Columbia University, NYC, for kindly allowing us to use their cardiography corpus.

References

Brandow, Ronald, Karl Mitze, and Lisa F. Rau. 1995. Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing and Management* 31(5): 675–685.

Hovy, Eduard H., and Chin-Yew Lin. 1997. Automated Text Summarization in SUMMARIST. In *Proceedings of the ACL/EACL-97 workshop on Intelligent Scalable Text Summarization.*

Kintsch, Walter, and Teun A. van Dijk. 1978. Toward a model of Text Comprehension and Production. *Psychological Review* 85(5): 363–394.

Kircz, Joost G. 1991. The rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of Documentation* 47(4): 354–372.

Kupiec, Julian, Jan O. Pedersen, and Francine Chen. 1995. A Trainable Document Summarizer. In *Proceedings of the 18th ACM-SIGIR Conference*, 68–73.

Salager-Meyer, Francoise. 1992. A text-type and move analysis study of verb tense and modality distributions in Medical English Abstracts. *English for Specific Purposes* 11: 93–113.

Spärck Jones, Karen. 1994. Discourse Modelling for Automatic Summarising. Tech. rep., Computer Laboratory, University of Cambridge. TR-290.

Swales, John. 1990. *Genre analysis: English in academic and research settings*, chap. Research Articles in English. Cambridge University Press.

Teufel, Simone, and Marc Moens. 1997. Sentence extraction as a classification Task. In *Proceedings of the ACL/EACL-97 workshop on Intelligent Scalable Text Summarization*, 58–65.

Teufel, Simone, and Marc Moens. 1998. Sentence extraction and rhetorical classification for flexible abstracts. In *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*, 16–25.

I. Introduction of problems, tasks, solutions by explicit mention

Ex-T	Own research task	**The aim of this paper is to** *examine the role that training plays in the tagging process.*
Ex-S	Own solution	**The basic idea for the analysis** *can be seen as a logical counterpart at the glue level of the standard type assignment for generalized quantifiers [REF].*
Ex-O-S	Other solution	**The traditional approach has been to** *plot isoglosses, delineating regions where the same word is used for the same concept.*
Ex-C	Own conclusions/ claims	*In our corpus study,* **we found that** *three types of utterances (prompts, repetitions and summaries) were consistently used to signal control shifts.*
Ex-O-C	Other conclusions/ claims	**It has often been stated that** *discourse is an inherently collaborative process.*
Ex-E	Own evaluation methodology	*In this section* **we evaluate the performance of** *the methodology implemented* by *predicting succeeding words by using preceding words.*
Ex-R	Own (numerical) results	**The evaluation of the accuracy of** *the rhetorical structure analysis carried out previously ([REF])* **showed** *74* **%.**

II. Contrastive introduction of problems, tasks, solutions

Co-S	Contrast between own and other solutions	*In this paper, we argue that* **instead of** *applying the arbitration process to the discourse level,* **it should** *be applied to the beliefs proposed by the discourse actions.*
Co-T	Contrast between own and other tasks/ problems	**Unlike most research in** *pragmatics that focuses on certain types of presuppositions or implicatures,* **we provide** *a global framework in which one can express all these types of pragmatic inferences.*
Co-C	Contrast between own and other claims	**Despite the hypothesis that** *the free word order of German leads to poor performance of low order HMM taggers when compared with a language like English,* **we have shown that** *the overall results for German are very much along the lines of comparable implementations for English, if not better.*

III. Attribution of properties to problems, tasks, solutions

Imp-T	Own research task is important	**The last decade has seen a growing interest in** *the application of machine learning to different kinds of linguistic domains.*
Hard-T	Own research task is hard	**One of the difficult problems in** *machine translation from Japanese to English or other European languages* **is** *the treatment of articles and numbers.*
New-T	Own problem/ research task is new	**No** *formal framework* **has been proposed, to our knowledge, to** *regulate the interaction between regular and exceptional grammatical resources.*
Good-S	Own solution is advantageous	*First, it is in certain respects* **simpler**, *in that* **it requires no** *postulation of otherwise unmotivated ambiguities in the source clause.*
Bad-O-S	Other solution is flawed	*However, we argue that such formalisms* **offer little help to** *computational linguists in practice.*

IV. Functional relations between problems, tasks, solutions

Solves	Own solution solves own research task	**This account** *also* **explains** *similar differences in felicity for other coordinating conjunctions as discussed in [REF].*
Avoids	Own solution avoids problems	*We have introduced a simple, natural definition of synchronous tree-adjoining derivation* **that avoids the** *expressivity and implementability* **problems of** *the original rewriting definition.*
Step	Own solution is a step towards research task	**We have** *thus* **developed** *an evaluation heuristic that combines several different measures,* **in order to** *select the parse that is deemed overall "best".*
Nec-S-T	Own solution necessary to achieve research task	*We have argued that obligations* **play an important role in** *accounting for the interactions in dialog.*
NotSo	Other solution does not solve problem/ task	*Dependency grammar* **runs into substantial difficulty trying to account for** *the proform one.*
New-P	Other solution introduces new problems	*Specifically,* **if a treatment such as [REF]'s is used** *to explain the forward progression of time in example [CREF],* **then it must be explained why** *sentence [CREF] is as felicitous as sentence [CREF].*

V. Direct comparison of problems, tasks, solutions

Better-S	Own solution is better than other solution	*We found that the MDL-based method* **performs better than** *the MLE-based method.*
Harder-T	Own research task is harder than other task	*. . . disambiguating word senses to the level of fine-grainedness found in WordNet* **is quite a bit more difficult than** *disambiguation to the level of homographs [REF], [REF].*

Figure 8: Rhetorical moves in scientific papers; examples from our corpus of computational linguistics

E.4. Teufel, Carletta and Moens (1999)

An annotation scheme for discourse-level argumentation in research articles

Simone Teufel[‡] and **Jean Carletta**[†] and **Marc Moens**[‡]
[‡]HCRC Language Technology Group and
[†]Human Communication Research Centre
Division of Informatics
University of Edinburgh
S.Teufel@ed.ac.uk, J.Carletta@ed.ac.uk, M.Moens@ed.ac.uk

Abstract

In order to build robust automatic abstracting systems, there is a need for better training resources than are currently available. In this paper, we introduce an annotation scheme for scientific articles which can be used to build such a resource in a consistent way. The seven categories of the scheme are based on rhetorical moves of argumentation. Our experimental results show that the scheme is stable, reproducible and intuitive to use.

1 Introduction

Current approaches to automatic summarization cannot create coherent, flexible automatic summaries. Sentence selection techniques (e.g. Brandow et al., 1995; Kupiec et al. 1995) produce extracts which can be incoherent and which, because of the generality of the methodology, can give under-informative results; fact extraction techniques (e.g. Rau et al., 1989, Young and Hayes, 1985) are tailored to particular domains, but have not really scaled up from restricted texts and restricted domains to larger domains and unrestricted text. Spärck Jones (1998) argues that taking into account the structure of a text will help when summarizing the text.

The problem with sentence selection is that it relies on extracting sentences out of context, but the meaning of extracted material tends to depend on where in the text the extracted sentence was found. However, sentence selection still has the distinct advantage of robustness.

We think sentence selection could be improved substantially if the global rhetorical context of the extracted material was taken into account more. Marcu (1997) makes a similar point based on rhetorical relations as defined by Rhetorical Structure Theory (RST, (Mann and Thompson, 1987)). In contrast to this approach, we stress the importance of rhetorical moves which are *global* to the argumentation of the paper, as opposed to local RST–type moves. For example, sentences which describe weaknesses of previous approaches can provide a good characterization of the scientific articles in which they occur, since they are likely to also be a description of the problem that paper is intending to solve. Take a sentence like "*Unfortunately, this work does not solve problem X*": if X is a shortcoming in someone else's work, this usually means that the current paper ***will*** try to solve X. Sentence extraction methods can locate sentences like these, e.g. using a cue phrase method (Paice, 1990).

But a very similar-looking sentence can play a completely different argumentative role in a scientific text: when it occurs in the section "Future Work", it might refer to a minor weakness in the work presented in the source paper (i.e. of the author's *own* solution). In that case, the sentence is *not* a good characterization of the paper.

Our approach to automatic text summarization is to find important sentences in a source text by determining their most likely argumentative role. In order to create an automatic process to do so, either by symbolic or machine learning techniques, we need training material: a collection of texts (in this case, scientific articles) where each sentence is annotated with information about the argumentative role that sentence plays in the paper. Currently, no such resource is available. We developed an annotation scheme as a starting point for building up such a resource, which we will describe in section 2. In section 3, we use content analysis techniques to test the annotation scheme's reliability.

2 The annotation scheme

We wanted the scheme to cover one text type, namely research articles, but from different presentational traditions and subject matters, so that we can use it for text summarization in a range of fields. This means we cannot rely on similarities in external presentation, e.g. section structure and typical linguistic formulaic expressions.

Previous discourse-level annotation schemes (e.g. Liddy, 1991; Kircz, 1991) show that information retrieval can profit from added rhetorical information in scientific texts. However, the definitions of the categories in these schemes relies on domain dependent knowledge like typical research methodology, and are thus too specific for our purposes.

General frameworks of text structure and argumentation, like Cohen's (1984) theoretical framework for general argumentation and Rhetorical Structure Theory (Mann and Thompson, 1987), are theoretically applicable to many different kinds of text types. However, we believe that restricting ourselves to the text type of research articles will give us an advantage over such general schemes, because it will allow us to rely on communicative goals typically occurring within that text type.

Swales' (1990) CARS (Creating a Research Space) model provides a description at the right level for our purposes. Swales claims that the regularities in the argumentative structure of research article introductions follow from the authors' primary communicative goal: namely to convince their audience that they have provided a contribution to science. From this goal follow highly predictable subgoals which he calls *argumentative moves* ("recurring and regularized communicative events"). An example for such a move is "*Indication of a gap*", where the author argues that there is a weakness in an earlier approach which needs to be solved.

Swales' model has been used extensively by discourse analysts and researchers in the field of English for Specific Purposes, for tasks as varied as teaching English as a foreign language, human translation and citation analysis (Myers, 1992; Thompson and Ye, 1991; Duszak, 1994), but always for manual analysis by a single person. Our annotation scheme is based on Swales' model but we needed to modify it. Firstly, the CARS model only applies to introductions of research articles, so we needed new moves to cover the other paper sections; secondly, we needed more precise guidelines to make the scheme applicable to reliable annotation for several non-discourse analysts (and for potential automatic annotation).

For the development of our scheme, we used computational linguistics articles. The papers in our collection cover a challenging range of subject matters due to the interdisciplinarity of the field, such as logic programming, statistical language modelling, theoretical semantics and computational psycholinguistics. Because the research methodology and tradition of presentation is so different in these fields, we would expect the scheme to be equally applicable in a range of disciplines other than those named.

Our annotation scheme consists of the seven categories shown in Figure 1. There are two versions of the annotation scheme. The *basic* scheme provides a distinction between three textual segments which we think is a necessary precondition for argumentatively-justified summarization. This distinction is concerned with the attribution of *authorship* to scientific ideas and solutions described in the text. Authors need to make clear, and readers need to understand:

- which sections describe generally accepted statements (BACKGROUND);
- which ideas are attributed to some other, specific piece of research outside the given paper, including own previous work (OTHER);
- and which statements are the authors' own *new* contributions (OWN).

The *full* annotation scheme consists of the basic scheme plus four other categories, which are based on Swales' moves. The most important of these is AIM (Swales' move "*Explicit statements of research goal*"), as these moves are good characterizations of the entire paper. We are interested in how far humans can be trained to consistently annotate these sentences; similar experiments where subjects selected one or several 'most relevant' sentences from a paper have traditionally reported low agreement (Rath et al., 1961). There is also the category TEXTUAL (Swales' move "*Indicate structure*"), which provides helpful information about section structure, and two moves having to do with attitude towards previous research, namely BASIS and CONTRAST.

The relative simplicity of the scheme was a compromise between two demands: we wanted the scheme to contain enough information for automatic summarization, but still be practicable for hand coding.

Annotation proceeds sentence by sentence according to the decision tree given in Figure 2. No instructions about the use of cue phrases were given, although some of the example sentences given in the guidelines contained cue phrases. The categorisation task resembles the judgements performed e.g. in dialogue act coding (Carletta et al.,

BASIC SCHEME	Background	Sentences describing some (generally accepted) background knowledge	**FULL SCHEME**
	Other	Sentences describing aspects of some specific other research in a neutral way (excluding contrastive or Basis statements)	
	Own	Sentences describing any aspect of the own work presented in this paper – except what is covered by Aim or Textual, e.g. details of solution (methodology), limitations, and further work.	
	Aim	Sentences best portraying the particular (main) research goal of the article	
	Textual	Explicit statements about the textual section structure of the paper	
	Contrast	Sentences contrasting own work to other work; sentences pointing out weaknesses in other research; sentences stating that the research task of the current paper has never been done before; direct comparisons	
	Basis	Statements that the own work uses some other work as its basis or starting point, or gets support from this other work	

Figure 1: Overview of the annotation scheme

1997; Alexandersson et al., 1995; Jurafsky et al., 1997), but our task is more difficult since it requires more subjective interpretation.

3 Annotation experiment

Our annotation scheme is based on the intuition that its categories provide an adequate and intuitive description of scientific texts. But this intuition alone is not enough of a justification: we believe that our claims, like claims about any other descriptive account of textual interpretation, should be substantiated by demonstrating that other humans can apply this interpretation consistently to actual texts.

We did three studies. Study I and II were designed to find out if the two versions of the annotation scheme (basic vs. full) can be learned by human coders with a significant amount of training. We are interested in two formal properties of the annotation scheme: stability and reproducibility (Krippendorff, 1980). Stability, the extent to which one annotator will produce the same classifications at different times, is important because an instable annotation scheme can never be reproducible. Reproducibility, the extent to which different annotators will produce the same classifications, is important because it measures the consistency of shared understandings (or meaning) held between annotators.

We use the Kappa coefficient K (Siegel and Castellan, 1988) to measure stability and reproducibility among k annotators on N items. In our experiment, the items are sentences. Kappa is a better measurement of agreement than raw percentage agreement (Carletta, 1996) because it factors out the level of agreement which would be reached by random annotators using the same distribution of categories as the real coders. No matter how many items or annotators, or how the categories are distributed, K=0 when there is no agreement other than what would be expected by chance, and K=1 when agreement is perfect. We expect high random agreement for our annotation scheme because so many sentences fall into the Own category.

Studies I and II will determine how far we can trust in the human-annotated training material for both learning and evaluation of the automatic method. The outcome of Study II (full annotation scheme) is crucial to the task, as some of the categories specific to the full annotation scheme (particularly Aim) add considerable value to the information contained in the training material.

Study III tries to answer the question whether the considerable training effort used in Studies I and II can be reduced. If it were the case that coders with hardly any task-specific training can produce similar results to highly trained coders, the training material could be acquired in a more efficient way. A positive outcome of Study III would also strengthen claims about the intuitivity of the category definitions.

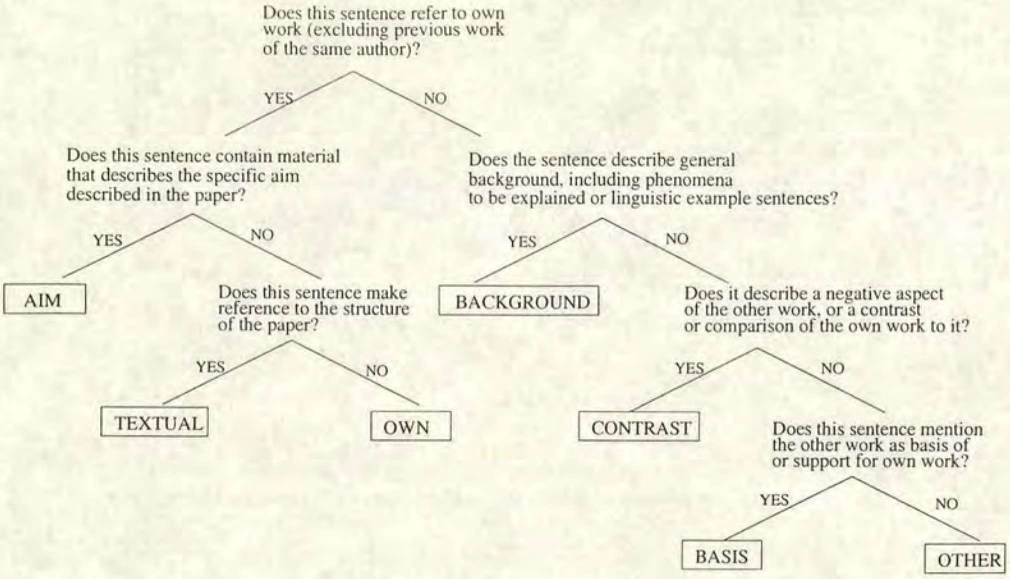

Does this sentence refer to own work (excluding previous work of the same author)?
YES
NO
Does this sentence contain material that describes the specific aim described in the paper?
YES
NO
AIM
Does this sentence make reference to the structure of the paper?
YES
NO
TEXTUAL
OWN
Does the sentence describe general background, including phenomena to be explained or linguistic example sentences?
YES
NO
BACKGROUND
Does it describe a negative aspect of the other work, or a contrast or comparison of the own work to it?
YES
NO
CONTRAST
Does this sentence mention the other work as basis of or support for own work?
YES
NO
BASIS
OTHER

Figure 2: Decision tree for annotation

Our materials consist of 48 computational linguistics papers (22 for Study I, 26 for Study II), taken from the Computation and Language E-Print Archive (`http://xxx.lanl.gov/cmp-lg/`). We chose papers that had been presented at COLING, ANLP or ACL conferences (including student sessions), or ACL-sponsored workshops, and been put onto the archive between April 1994 and April 1995.

3.1 Studies I and II

For Studies I and II, we used three highly trained annotators. The annotators (two graduate students and the first author) can be considered skilled at extracting information from scientific papers but they were not experts in all of the subdomains of the papers they annotated. The annotators went through a substantial amount of training, including the reading of coding instructions for the two versions of the scheme (6 pages for the basic scheme and 17 pages for the full scheme), four training papers and weekly discussions, in which previous annotations were discussed. However, annotators were not allowed to change any previous decisions. For the stability figures (intra-annotator agreement), annotators re-coded 6 randomly chosen papers 6 weeks after the end of the annotation experiment. Skim-reading and annotation of an average length paper (3800 words) typically took the annotators 20–30 minutes.

During the annotation phase, one of the papers turned out to be a review paper. This paper caused the annotators difficulty as the scheme was not intended to cover reviews. Thus, we discarded this paper from the analysis.

The results show that the basic annotation scheme is stable (K=.83, .79, .81; N=1248; k=2 for all three annotators) and reproducible (K=.78, N=4031, k=3). This reconfirms that trained annotators are capable of making the basic distinction between own work, specific other work, and general background. The full annotation scheme is stable (K=.82, .81, .76; N=1220; k=2 for all three annotators) and reproducible (K=.71, N=4261, k=3). Because of the increased cognitive difficulty of the task, the decrease in stability and reproducibility in comparison to Study I is acceptable. Leaving the coding developer out of the coder pool for Study II did not change the results (K=.71, N=4261, k=2), suggesting that the training conveyed her intentions fairly well.

We collected informal comments from our annotators about how natural the task felt, but did not conduct a formal evaluation of subjective perception of the difficulty of the task. As a general approach in our analysis, we wanted to look at the trends in the data as our main information source.

Figure 3 reports how well the four non-basic categories could be distinguished from all other categories, measured by Krippendorff's diagnostics for category distinctions (i.e. collapsing all *other* distinctions). When compared to the overall reproducibility of .71, we notice that the annotators were good at distinguishing AIM and TEX-

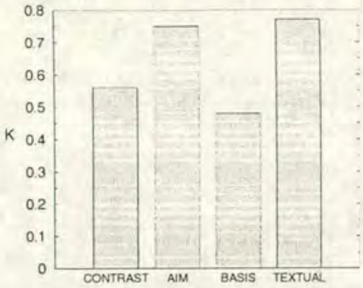


Figure 3: Reproducibility diagnostics: non-basic categories (Study II)

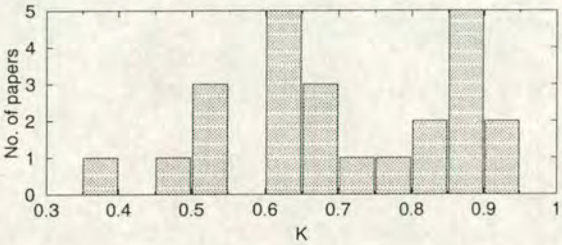


Figure 4: Distribution by reproducibility (Study II)

TUAL. This is an important result: as AIM sentences constitute the best characterization of the research paper for the summarization task we are particularly interested in having them annotated consistently in our training material. The annotators were less good at determining BASIS and CONTRAST. This might have to do with the location of those types of sentences in the paper: AIM and TEXTUAL are usually found at the beginning or end of the introduction section, whereas CONTRAST, and even more so BASIS, are usually interspersed within longer stretches of OWN. As a result, these categories are more exposed to lapses of attention during annotation.

If we blur the less important distinctions between CONTRAST, OTHER, and BACKGROUND, the reproducibility of the scheme increases to K=.75. Structuring our training set in this way seems to be a good compromise for our task, because with high reliability, it would still give us the crucial distinctions contained in the basic annotation scheme, plus the highly important AIM sentences, plus the useful TEXTUAL and BASIS sentences.

The variation in reproducibility across papers is large, both in Study I and Study II (cf. the quasi-bimodal distribution shown in Figure 4). Some hypotheses for why this might be so are the following:

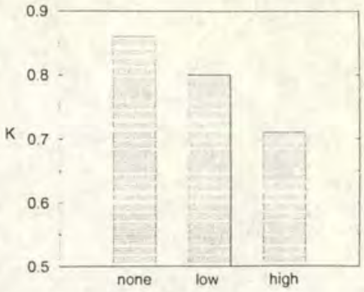


Figure 5: Effect of self-citation ratio on reproducibility (Study I)

- One problem our annotators reported was a difficulty in distinguishing OTHER work from OWN work, due to the fact that some authors did not express a clear distinction between *previous* own work (which, according to our instructions, had to be coded as OTHER) and *current, new* work. This was particularly the case where authors had published several papers about different aspects of one piece of research. We found a correlation with self citation ratio (ratio of self citations to all citations in running text): papers with many self citations are more difficult to annotate than papers that have few or no self citations (cf. Figure 5).

- Another persistent problematic distinction for our annotators was that between OWN and BACKGROUND. This could be a sign that some authors aimed their papers at an expert audience, and thus thought it unnecessary to signal clearly which statements are commonly agreed in the field, as opposed to their own new claims. If a paper is written in such a way, it can indeed only be understood with a considerable amount of domain knowledge, which our annotators did not have.

- There is also a difference in reproducibility between papers from different *conference types*, as Figure 6 suggests. Out of our 25 papers, 4 were presented in student sessions, 4 came from workshops, the remaining 16 ones were main conference papers. Student session papers are easiest to annotate, which might be due to the fact that they are shorter and have a simpler structure, with less mentions of previous research. Main conference papers dedicate more space to describing and

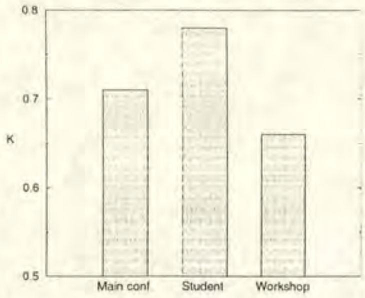


Figure 6: Effect of conference type on reproducibility (Study II)

criticising other people's work than student or workshop papers (on average about one fourth of the paper). They seem to be carefully prepared (and thus easy to annotate); conference authors must express themselves more clearly than workshop authors because they are reporting finished work to a wider audience.

3.2 Study III

For Study III, we used a different subject pool: 18 subjects with no prior annotation training. All of them had a graduate degree in Cognitive Science, with two exceptions: one was a graduate student in Sociology of Science, and one was a secretary. Subjects were given only minimal instructions (1 page A4), and the decision tree in Figure 2. Each annotator was randomly assigned to a group of six, all of whom independently annotated the same single paper. These three papers were randomly chosen from the set of papers for which our trained annotators had previously achieved good reproducibility in Study II (K=.65,N=205, k=3; K=.85,N=192,k=3; K=.87,N=144,k=3, respectively).

Reproducibility varied considerably between groups (K=.35, N=205, k=6; K=.49, N=192, k=6; K=.72, N=144, k=6). Kappa is designed to abstract over the number of coders. Lower reliablity for Study III as compared to Studies I and II is not an artefact of how K was calculated.

Some subjects in Group 1 and 2 did not understand the instructions as intended – we must conclude that our very short instructions did not provide enough information for consistent annotation. This is not surprising, given that human indexers (whose task is very similar to the task introduced here) are highly skilled professionals. However, part of this result can be attributed to the papers: Group 3, which annotated the paper found to be most reproducible in Study II, performed almost as well as trained annotators; Group 1, which performed worst, also happened to have the paper with the lowest reproducibility. In Groups 1 and 2, the most similar three annotators reached a respectable reproducibility (K=.5, N=205, k=3; K=.63, N=192, k=3). That, together with the good performance of Group 3, seems to show that the instructions did at least convey some of the meaning of the categories.

It is remarkable that the two subjects who had no training in computational linguistics performed reasonably well: they were not part of the circle of the three most similar subjects in their groups, but they were also not performing worse than the other two annotators.

4 Discussion

It is an interesting question how far shallow (human and automatic) information extraction methods, i.e. those using no domain knowledge, can be successful in a task such as ours. We believe that argumentative structure has so many reliable linguistic or non-linguistic correlates on the surface – physical layout being one of these correlates, others are linguistic indicators like *"to our knowledge"* and the relative order of the individual argumentative moves – that it should be possible to detect the line of argumentation of a text without much world knowledge. The two non-experts in the subject pool of Study III, who must have used some other information besides computational linguistics knowledge, performed satisfactorily – a fact that seems to confirm the promise of shallow methods.

Overall, reproducibility and stability for trained annotators does not quite reach the levels found for, for instance, the best dialogue act coding schemes (around K=.80). Our annotation requires more subjective judgments and is possibly more cognitively complex. Our reproducibility and stability results are in the range which Krippendorff (1980) describes as giving marginally significant results for reasonable size data sets when correlating two coded variables which would show a clear correlation if there were prefectly agreement. That is, the coding contains enough signal to be found among the noise of disagreement.

Of course, our requirements are rather less stringent than Krippendorff's because only one coded variable is involved, although coding is expensive enough that simply building larger data sets is not an attractive option. Overall, we find the level of agreement which we achieved acceptable. However, as with all coding schemes, its usefulness will only be clarified by the final application.

The single most surprising result of the experiments is the large variation in reproducibility between papers. Intuitively, the reason for this are qualitative differences in individual writing style – annotators reported that some papers are better structured and better written than others, and that some authors tend to write more clearly than others. It would be interesting to compare our reproducibility results to independent quality judgements of the papers, in order to determine if our experiments can indeed measure the clarity of scientific argumentation.

Most of the problems we identified in our studies have to do with a lack of distinction between own and other people's work (or own previous work). Because our scheme discriminates based on these properties, as well as being useful for summarizing research papers, it might be used for automatically detecting whether a paper is a review, a position paper, an evaluation paper or a 'pure' research article by looking at the relative frequencies of automatically annotated categories.

5 Conclusions

We have introduced an annotation scheme for research articles which marks the aims of the paper in relation to past literature. We have argued that this scheme is useful for building better abstracts, and have conducted some experiments which show that the annotation scheme can be learned by trained annotators and subsequently applied in a consistent way. Because the scheme is reliable, hand-annotated data can be used to train a system which applies the scheme automatically to unseen text.

The novel aspects of our scheme are that it applies to different kinds of scientific research articles, because it relies on the *form and meaning of argumentative aspects* found in the text type rather than on contents or physical format. As such, it should be independent of article length and article discipline. In the future, we plan to show this by applying our scheme to journal and conference articles from a range of disciplines. Practical reasons have kept us from using journal articles as data so far (namely the difficulty of corpus collection and the increased length and subsequent time effort of human experiments), but we are particularly interested in them as they can be expected to be of higher quality. As the basic argumentation is the same as in conference articles, our scheme should be applicable to journal articles at least as consistently as to the papers in our current collection.

6 Acknowledgements

We wish to thank our annotators, Vasilis Karaiskos and Ann Wilson, for their patience and diligence in this work, and for their insightful, critical, and very useful observations.

The first author is supported by an EPSRC studentship.

References

Jan Alexandersson, Elisabeth Maier, and Norbert Reithinger. 1995. A robust and efficient three-layered dialogue component for a speech-to-speech translation system. In *Proceedings of the Seventh European Meeting of the ACL*, pages 188–193.

Ronald Brandow, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Robin Cohen. 1984. A computational theory of the function of clue words in argument understanding. In *Proceedings of COLING-84*, pages 251–255.

Anna Duszak. 1994. Academic discourse and intellectual styles. *Journal of Pragmatics*, 21:291–313.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca, 1997. *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual.* University of Colorado, Institute of Cognitive Science. TR-97-02.

Joost G. Kircz. 1991. The rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of Documentation*, 47(4):354–372.

Klaus Krippendorff. 1980. *Content analysis: an introduction to its methodology.* Sage Commtext series; 5. Sage, Beverly Hills London.

Julian Kupiec, Jan O. Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference, Association for Computing Machinery, Special Interest Group Information Retrieval*, pages 68–73.

Elizabeth DuRoss Liddy. 1991. The discourse-level structure of empirical abstracts: an exploratory study. *Information Processing and Management*, 27(1):55–81.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: description and construction of text structures. In G. Kempen, editor, *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, pages 85–95, Dordrecht. Nijhoff.

Daniel Marcu. 1997. From discourse structures to text summaries. In Inderjeet Mani and Mark T. Maybury, editors, *Proceedings of the workshop on Intelligent Scalable Text Summarization, in association with ACL/EACL-97.*

Greg Myers. 1992. In this paper we report... – speech acts and scientific facts. *Journal of Pragmatics*, 17(4):295–313.

Chris D. Paice. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26:171–186.

G.J Rath, A. Resnick, and T. R. Savage. 1961. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2):139–143.

Lisa F. Rau, Paul S. Jacobs, and Uri Zernik. 1989. Information extraction and text processing using linguistic knowledge acquisition. *Information Processing and Management*, 25(4):419–428.

Sidney Siegel and N.J. Jr. Castellan. 1988. *Nonparametric statistics for the Behavioral Sciences.* McGraw-Hill, second edition edition.

Karen Spärck Jones. 1998. Automatic summarising: factors and directions. In *ACL/EACL-97 Workshop 'Intelligent Scalable Text Summarization'.*

John Swales. 1990. *Genre analysis: English in academic and research settings.* Cambridge University Press.

Geoff Thompson and Yiyun Ye. 1991. Evaluation in the reporting verbs used in academic papers. *Applied Linguistics*, 12(4):365–382.

Sheryl R. Young and Phillip J. Hayes. 1985. Automatic classification and summarization of banking telexes. In *Proceedings of the Second Conference on Artificial Intelligence Applications.*

E.5. Teufel and Moens (1999a)

Argumentative classification of extracted sentences as a first step towards flexible abstracting

Simone Teufel and Marc Moens
HCRC Language Technology Group
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
S.Teufel@ed.ac.uk M.Moens@ed.ac.uk

Abstract

Knowledge about the rhetorical structure of a text is useful for automatic abstraction. We are interested in the automatic extraction of rhetorical units from the source text, units such as PROBLEM STATEMENT, CONCLUSIONS and RESULTS. We want to use such extracts to generate high-compression abstracts of scientific articles. In this paper, we present an extension of Kupiec, Pedersen and Chen's (1995) methodology for trainable statistical sentence extraction. Our extension additionally classifies the extracted sentences according to their rhetorical role.

1 Introduction

1.1 Flexible abstracting

Until recently, the world of research publications was heavily paper-oriented. Journals, dissertations and other publications were available only in paper form. To keep researchers informed of publications in their area of interest, secondary publishers produced journals with abstracts of research material. The main role of these abstracts was to act as a *decision* tool: on the basis of the abstract a researcher could decide whether the source text was worth a visit to the library or a letter to the author requesting a copy of the full article.

For reasons of consistency (and copyright) these abstracts often were not the abstracts produced by the original authors, but by professional abstractors, and written according to agreed guidelines and recommendations (Borko and Chatman, 1963). These guidelines suggest that such abstracts should be aimed at the "partially informed reader"—someone who knows enough about the field to understand the basic methodology and general goals of the paper but does not necessarily have enough of an overview of previous work to assess where a certain article is situated in the field or how articles are related to each other (Kircz, 1991). For a novice reader, such an abstract would be too terse; for experienced researchers the abstract would provide unnecessary detail. In addition, because the abstract is a pointer to an article not immediately available, the abstract has to be self-contained: the reader should be able to grasp the main goals and achievements of the full article without needing the source text for clarification.

Over the past few years this picture has changed dramatically. Research articles are now increasingly being made available on-line. Indeed, the goal of automated summarization presupposes that the full article is available in machine-readable form. As a result, abstracts will have different or additional functions from the ones they used to have.

A typical scenario might be one where a user receives a large quantity of machine-readable articles, for example in reply to a search query, from a database of scientific articles or from the Internet. In such a context, abstracts can still be used as a decision tool, to help the user decide which articles to look at first. But in this context abstracts could also be used as a *navigation* tool, helping users find their way through the retrieved document collection. When abstracts are generated as needed, rather than stored in a fixed form, they could show how certain articles are related to other articles in logical and chronological respect, e.g. they could summarize similarities between articles, indicating which of the retrieved articles share the same research questions or methodologies. This type of navigation within a set of papers can support users in making a more informed decision on how well a paper fits their information needs.

Abstracts also don't need to be self-contained anymore. They can contain pointers (e.g. in the form of hyperlinks) to certain passages in the full article. And they can be "embedded" in the source text, highlighting in context the most relevant sentences, as has been demonstrated with commercial products such as Microsoft's "AutoSummarize" feature in Word97.

Abstracts can thus play an important role for the non-linear reading of textual material—the process whereby readers efficiently take in the content of a text by jumping in seemingly arbitrary fashion from conclusion to table of contents, section headers, captions, etc. Nonlinear reading is typical for scientists (Pirelli et al., 1984; Bazerman, 1988); it serves to efficiently build a model of the text's structure as well as to extract the main concepts of the paper. However, O'Hara and Sellen (1997) have shown that nonlinear reading is something people only do well with paper: the physical properties of paper allow readers to quickly scan the document and jump back and forth without losing their place in the document. On-line display mechanisms do not as yet have such facilities. Embedded or otherwise contextualized abstracts can facilitate this process of nonlinear reading by revealing the text's logical and semantic organization.

The old type of abstract was a fixed, long-lived, stand-alone text, targeted at one particular type of user. The new type of abstract is more dynamic and user-responsive, generated automatically when needed and thus less long-lived. Even though such abstracts will be of a lower quality when compared to human-crafted abstracts, we predict that they will be of more use in many situations. It is the flexible automatic generation of such abstracts which we see as our long-term goal.

1.2 Our approach

We would like to develop a summarization system which is not tied to a particular scientific *domain*. The processing robustness needed for this, as well as the speed with which we would like to be able to deliver abstracts, suggests that a deep semantic analysis of the source text is not a viable option.

Many robust summarization systems have opted for statistical sentence extraction: systems have been designed which extract "important" sentences from a text, where the importance of the sentence is inferred from low-level properties which can be more or less objectively calculated. Over the years there have been many suggestions as to which low-level features can help determine the importance of a sentence in the context of a source text, such as stochastic measurements for the significance of key words in the sentence (Luhn, 1958), its location in the source text (Baxendale, 1958; Edmundson,1969), connections with other sentences (Skorochod'ko, 1972; Salton et al., 1994), and the presence of cue or indicator phrases (Paice, 1981) or of title words (Edmundson, 1969) . The result of this process is an *extract*, i.e. a collection of sentences selected verbatim from the text.

These extracts are then used as the abstract of the text. But this has a number of disadvantages. For one thing, they are just a collection of sentences, possibly difficult to interpret because of phenomena like unresolved anaphora and unexpected topic shifts. Post-processing of the extracts can remove some of these shortcomings, e.g. by not using sentences in the extract which contain obviously anaphoric expressions or by including surrounding sentences into the extract which are likely to resolve the anaphora (Johnson et al., 1993). Of course, this may lead to extracts which are too long, or it might mean losing sentences which are crucial to the content of the source text, thereby reducing the value of the resulting extract.

But even if—after postprocessing—each individual sentence might be interpretable in isolation, that still does not mean that the extract as a whole will be easy to understand. Assuming that the text is coherent, people will try to fill in the semantics gaps between potentially unconnected sentences. In the act of doing so, they may introduce inappropriate semantics links and get the wrong idea about the content of the source text.

Another problem is that sentence extraction does not work very well for high compression summarization. Typical sentence extraction programs compress to about 10 or 15% of the original—for example, reducing a short newspaper article to a few sentences. Even if these sentences do not form a coherent text, that does not matter much: the extract is short enough to still make sense. But we are interested in summarizing longer texts, such as journal articles. Simple sentence extraction methods will reduce a 20-page article to a 2-page collection of unconnected sentences, a document surrogate which is not adequate as an abstract. Reducing the extract further to obtain a real abstract is difficult.

The reason for this difficulty is that once the abstract-worthy sentences have been extracted, the logical and rhetorical organization of the text is lost, and it becomes difficult to make sensible decisions on how to reduce the text further. To overcome this problem, we want to select abstract-worthy material from the source text, whilst at the same time keeping information about the overall rhetorical structure of the source text and of the role of each of the extract sentences in that rhetorical structure.

However, the full rhetorical structure of a paper (and the logical structure of the research it reports) is a very complex structure, and is difficult to model automatically. Although Marcu (1997) presents an approach for the automated rhetorical analysis of texts, these texts are considerably shorter than the ones we are interested in summarizing. Rather than attempting a full rhetorical analysis of the source text, we wanted to extract just enough rhetorical information so as to be able to determine the rhetorical contribution of all and only the abstract-worthy sentences, without modeling domain knowledge or performing domain-sensitive reasoning. We make use of meta-comments in the text, phrases like *"we have presented a method for"*, and *"however, to our knowledge there is no"* which signal rhetorical status.

The abstract we envisage is construed as an argumentative template, where the slots represent certain argumentative or rhetorical roles, such as GOAL, ACHIEVEMENT, BACKGROUND, METHOD, etc. Abstracting means analysing the argumentative structure of the source text and identifying textual extracts which constitute appropriate fillers for the template. For each slot in the template (i.e. each rhetorical role) the system identifies a number of plausible fillers (i.e. text excerpts), with different levels of confidence. We call this collection of meaningful sentences *together* with information about their rhetorical role in the full article a *rhetorically annotated extract*.

Our idea of an abstract is thus more related to the *structured abstracts* which have become prevalent in the medical domain in the past decade (Broer, 1971; Adhoc, 1987; Rennie and Glas, 1991). Hartley et al. (1996) and Hartley and Sydes (1997) show in user studies that these abstracts are easier to read and more efficient for information assessment than traditional summaries.

In a further step (the generation of the real abstract), some of this information can be added or suppressed, in order to allow abstracts of varying length to be generated. For example, the amount of BACKGROUND information supplied in the abstract can be varied depending on whether users have been identified as novices or experienced readers. Rhetorical roles for which only low-probability evidence was found in the source document can be pruned until an abstract of the required length is reached.

Two questions arise from this approach. The first question is how the building blocks of the abstract template, i.e. the rhetorical roles, should be defined. This is a particular problem for our approach because very little is known about what our new type of abstract should look like. Most of the information on good abstracts deals with the world of paper, not with the use of on-line research publications. That means that we cannot take existing guidelines on how to produce balanced, informative, concise abstracts at face value; we will need to fall back on a different set of intuitions as to what constitutes a good abstract. To answer this question, we take research on the argumentative structure of research articles and their abstracts as our starting point. This will be discussed in section 2.

The second question is how a system can be trained to find suitable fillers in a source text to complete such a template. In section 3 we report on our experiments to train a system to automatically detect meaningful sentences in the source text together with their rhetorical role.

2 The argumentative structure of research articles and their abstracts

2.1 Rhetorical divisions in research articles

Scholarly articles serve the process of communicating scientific information. The communicative function of a scientific research article is thus very well-defined: to present and refer to the results of specific research (Salager, 1992). In some scientific domains research follows predictable patterns of methodology and also of presentation. A rigid, highly structured building plan for research articles has evolved as a result, where rhetorical divisions are clearly marked in section headers (Kintsch and van Dijk, 1978). Prototypical rhetorical divisions include *Introduction*, *Purpose*, *Experimental Design*, *Results*, *Discussion*, and *Conclusions*. This is very efficient: researchers in psycholinguistics, for example, know with great accuracy where in any given article to find the information on the number of participants in an experiment.

The papers in our corpus do not show this pattern. This has undoubtedly to do with the fact that our corpus consists of articles in computational linguistics and cognitive science. The papers draw from many subdisciplines, and most papers in our collection cannot be uniquely classified by sub-discipline, because they report on truly interdisciplinary research coming from different sub-disciplines. As a rough estimate, about 45% of the articles in our collection are predominantly technical in style, describing implementations (i.e. engineering solutions); about 25% report on research in theoretical linguistics, with an argumentative tenet; the remaining 30% are empirical (psycholinguistic or psychological experiments or corpus studies). As a result, we found a heterogeneous mixture of methodologies and traditions of presentation, with fewer prototypical rhetorical divisions than expected. Even though most of our articles have an introduction and conclusions (sometimes occurring under headers with different names), and almost all of them cite previous work, the presentation of the problem and the methodology/solution are idiosyncratic to the domain and personal writing style. Figure 1 shows the headers with the highest frequency for 123 examined papers—surprisingly few of them correspond to prototypical rhetorical divisions; the rest contain content specific terminology.

vspace4m

Freq.	Header
104	Introduction
56	Conclusion
27	Conclusions
21	Acknowledgments
15	Discussion
14	Results
11	Experimental Results
8	Related Work
8	Implementation
8	Evaluation
7	Example
7	Background

Figure 1: Headers with highest frequency from our collection

Apart from not being easily identified in our corpus, distinctions as expressed in rhetorical divisions are also too coarse for our purposes, namely to analyze scientific articles with respect to document structure, in a way which is flexible enough to cover the variety found in our corpus. A rhetorical division like *Introduction* can contain a problem statement, a motivation, a description of previous relevant work, and other such units. These smaller units are the ones that we are interested in, units which Swales (1981) calls *moves*, where a move is defined as "a semantic unit related to the writer's purpose".

2.2 Author intentions and argumentation in research articles

Swales (1990) claims that the main communicative goal of an author, far from the unbiased reporting of research, is to convince readers of the validity and importance of the work, in order to have the paper reviewed positively and thus published. Argumentation is used to show that the presented research was a contribution to science: that the solution proposed in the paper either solves a *new* problem, or, if a *known* problem is addressed, that the presented solution is better than that proposed by other researchers.

Swales analyzed several hundred introduction sections of scientific research papers from two data collections: research articles in the physical sciences and a mixture of research articles from several science and engineering fields. This analysis led to his CARS model ("Create a Research Space") which is schematically depicted in Figure 2; the right hand side of the figure shows examples from our corpus. This model describes prototypical rhetorical building plans of introductions, based on the rhetorical moves that authors typically employ to fulfill the communicative goal of writing a paper. One such rhetorical move is to motivate the need for the research presented (Move 2), which can be done in different ways, e.g. by pointing out a weakness of a previous approach (Move 2A/B) or by explicitly stating the research question (Move 2C). Note that context plays an important role for the classification of a sentence in Swales' system: the example sentence for Move 2D (which characterizes the work actually reported in the article) would constitute a different move if it had appeared towards the end of the article, or under the heading *Future Work*.

Inspection of introduction sections in our corpus showed that the steps defined by Swales' CARS model describe the argumentation phenomena at the right level of abstraction for our purposes; the author's typical intentions, expressed as predictable textual moves, seem to generalize well to the domain of computational linguistics and cognitive science.

We also observed a wide range of meta-comments in our corpus (the underlined phrases in the right hand side of Figure 2). The source of our collection being an unmoderated medium, writing style in the articles varies from formal to quite informal. About a third of the articles were not written (or subsequently edited) by native speakers of English. Also, meta-comments need not be unambiguous with respect to the rhetorical move they signal. Nevertheless, we claim that overall, they are still good enough indicators of rhetorical status to be extremely useful in a practical, shallow kind of discourse analysis.

2.3 Argumentative structure of abstracts

Although we argued that guidelines for abstracts cannot be taken at face value when designing a high-level framework for on-line abstracts, there is ample information in the literature which can be used to inform decisions about a desirable argumentative structure for abstracts.

As is the case with the communicative function of the whole paper, the communicative function of an abstract is one of a narrow range of things: it can be an indicative abstract, reporting the topic of the full article, or an informative abstract, reporting the topic of the source article as well as its main findings and conclusions (Cremmins, 1996; Rowley, 1982). As in the case of research articles, the communicative function of abstracts has led to common expectations of their rhetorical building blocks, such as *General Background, Specific Problem* tackled by full article, *Main Results, Recommendations*, etc. Buxton and Meadows (1978) provide a comparative survey of the contents of abstracts in the physics domain. They studied which rhetorical section in the source text (*Introduction–Method–Result–Discussion*) corresponds to the information in the abstracts and found, for example, that abstracts tend not to report material from the *Method* section. There is similar research on medical abstracts (Salager-Meyer, 1992) and sociological and humanities abstracts (Milas-Bracovic, 1987).

There is a consensus about the content units of informative abstracts for such articles in the experimental sciences—the majority of information in the descriptive and prescriptive abstracting literature seems to have concentrated on experimental sciences. Most authors agree that informative abstracts should mention the following four information units (ANSI, 1979; ISO, 1976; Day, 1995; Rowley, 1982; Cremmins, 1996):

1. the PURPOSE or PROBLEM of the full article,
2. the SCOPE or METHODOLOGY,
3. the RESULTS,
4. and CONCLUSIONS or RECOMMENDATIONS

In line with these recommendations, Manning (1990) argues that informative abstracts are not a miniature version of the full article in the sense of offering "a paraphrase of every rhetorical section" of the source article.

There is more disagreement about "peripheral" content units, such as BACKGROUND, INCIDENTAL FINDINGS, FUTURE WORK, RELATED WORK, and DATA. Of particular interest to us is the content unit BACKGROUND. According to Alley (1996) , BACKGROUND is a useful content unit in an abstract if it is restricted to being the first sentence of the abstract. Other authors (Rowley, 1982; Cremmins, 1996) recommend not to include any background information at all. We believe that background information is potentially important, Move 1: Establishing a territory

1.1	**Claiming centrality**	• *The last decade has seen a growing interest in the application of machine learning to different kinds of linguistic domains ...*
1.2	**Making topic generalizations** (background knowledge) OR	• *The traditional approach has been to plot isoglosses, delineating regions where the same word is used for the same concept.*
	(description of phenomena)	• *In the Japanese language, the causative and the change of voice are realized by agglutinations of those auxiliary verbs at the tail of current verbs.*
1.3	**Reviewing previous research**	• *Brown et al. (1992) suggest a class-based n-gram model in which words with similar cooccurrence distributions are clustered in word classes.*

Move 2: Establishing a niche

2A	**Counter-claiming**	• *However, we argue that such formalisms offer little help to computational linguists in practice.*
or 2B	**Indicating a gap**	• *... no formal framework has been proposed, to our knowledge, to regulate the interaction between regular and exceptional grammatical resources.*
or 2C	**Question-Raising**	• *Can the restrictive power of a single constraint be estimated in a reliable way to allow an effective scheduling procedure being devised?*
or 2D	**Continuing a tradition**	• *The remaining issue is to find a way of better accounting for unsymmetrical accommodation.*

Move 3: Occupying a niche

3.1A	**Outlining purpose**	• *The aim of this paper is to examine the role that training plays in the tagging process ...*
or 3.1B	**Announcing present research**	• *In this paper, we argue that instead of applying the arbitration process to the discourse level, it should be applied to...*
3.2	**Announcing principle findings**	• *In our corpus study, we found that three types of utterances (prompts, repetitions and summaries) were consistently used to signal control shifts....*
3.3	**Indicating article structure**	• *This paper is organized as follows: We begin in Section [CREF] by examining the distribution of possessive pronouns...*

Figure 2: Swales' (1990) CARS model with illustrative examples from our corpus

especially for self-contained abstracts and for abstracts for novice readers.

There is similar disagreement over the content unit Related work. Cremmins (1996) states that it should not be included in an abstract unless the studies are replications or evaluations of earlier work. However, depending on the information need, previous work might actually have been central to the original information need of the user. Therefore, we want to preserve the possibility of including it in our modular abstract.

For the experiments reported in this paper, we chose the four generally accepted categories, but we had to redefine each class slightly in order to achieve higher domain-independence.

For example, we use the label Solution/Method instead of Methodology/Scope: unlike in purely experimental research, where methodologies are long-lived research tools that are agreed upon in the field and do not change often, the range of possible methodologies in computational linguistics is vast, and a new, short-lived methodology might be invented just for the given problem-solving task, in which case the label "solution" seems more appropriate.

We added the two controversial roles Related Work and Background. And we added the role Topic, as the name of the research area or of the most general problem in the field. Thus, we ended up with the seven argumentative units listed in Figure 3.

Note that the labels of our annotation scheme can be naturally defined by rhetorical moves, such as the

Rhetorical role	
Background	Back
Topic/Aboutness	Topi
Related Work	Rwrk
Purpose/Problem	Pu/Pr
Solution/Method	Solu
Result	Resu
Conclusion/Claim	Co/Cl

Figure 3: Rhetorical roles in our annotation scheme

ones in Swales' CARS model. For example, Move 1.1 ("claiming centrality") provides good fillers for the Topic slot, whereas Problem, i.e. the specific problem of the paper, is very likely to be found in Move 2A–D ("indicating a gap").

Our annotation scheme forms the basis of the manual and automatic classification which is reported in the next section.

3 Our experiment

3.1 Previous work

Kupiec et al. (1995) introduce the notion of corpus-based abstracting: they recast the problem of sentence extraction as statistical classification. More specifically, they use supervised learning to automatically adjust feature weights with a Naive Bayesian classifier, combining the features (heuristics) mentioned in the literature. They used a corpus of research articles and corresponding summaries. The new idea in Kupiec *et al.*'s work is how they defined their *gold standards*. Gold standards are the class of sentences that, by definition, constitute the correct set of answers, usually defined by an expert in the field. The gold standard has to be defined independently and before the experiment. In Kupiec *et al.*'s work, the gold standard sentences are defined as the set of sentences in the source text that "align" with a sentence in the summary—i.e. sentences that show sufficient semantic and syntactic similarity with a summary sentence. The underlying reason is that a sentence in the source text is abstract-worthy if professional abstractors used it or parts of it when producing their summary. In Kupiec *et al.*'s corpus of 188 engineering articles with summaries written by professional abstractors, 79% of sentences in the summary also occurred in the source text with at most minor modifications.

Kupiec *et al.* then try to determine the characteristic properties of abstract-worthy sentences according to a number of features, viz. presence of particular cue phrases, location in the text, sentence length, occurrence of thematic words, and occurrence of proper names. Each document sentence receives a score for each of the features, resulting in an estimate for the sentence's probability to also occur in the summary. This probability is calculated for each feature value as a combination of the probability of the feature-value pair occurring in a sentence which is in the summary (successful case) and the probability that the feature-value pair occurs unconditionally.

Evaluation of the training relies on cross-validation: the model is trained on a training set of documents, leaving all documents from one journal out at a time (the current test set). The model is then used to extract candidate sentences from all documents of the test set. Evaluation measures co-selection between the extracted sentences and the gold standard sentences in precision (number of sentences extracted correctly over total number of sentences selected) and recall (number of sentences extracted correctly over total number of gold standard sentences). Since from any given test text as many sentences are selected as there are gold standard sentences, numerical values for precision and recall are the same. The precision/recall values of the individual heuristics range between 20–33%; the highest cumulative result (44%) was achieved using paragraph, fixed phrases (indicators) and sentence length features.

3.2 Abstracting as stepwise classification

We decided to perform the automatic generation of rhetorically annotated extracts by a process of repeated classification, borrowing the classification methodology from Kupiec *et al.* The basic procedure for the sentence extraction and classification experiment is the following:

Step one: Extraction of abstract-worthy sentences. We try to separate sentences which carry *any* rhetorical roles (grey set of sentences in Figure 4) from irrelevant sentences, which are by far the larger part of the text (white set of sentences in Figure 4). The output of this step is called the *intermediate extract*. Errors in this task will lead to the inclusion of irrelevant material in the extracts (false positives), or the exclusion of relevant material from the extracts (false negatives).

Step two: Identification of the correct rhetorical role. Once good sentence candidates have been identified, we classify them according to one of the seven rhetorical roles (in Figure 4, this corresponds to the sub-classification of the grey sentences). The output of this step is called a *rhetorically annotated extract*.

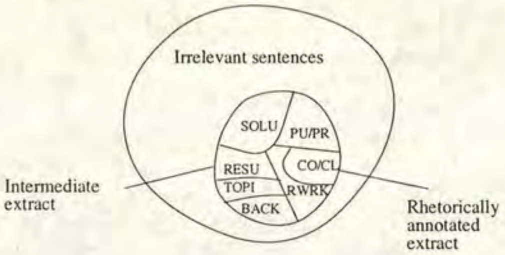


Figure 4: Abstracting as classification

We decided to split the task because we suspected that different heuristics would be more useful for the different tasks—a two-step process allows for the separation of these distinctions into two training processes.

Also, another motivation for the separation of the tasks stems from the fact that indicator phrases don't have to be unambiguous with respect to their argumentative status. For example, the phrase *"in this paper, we have"* is a very good overall relevance indicator, and it is quite likely that a sentence or paragraph starting with it will carry important global-level information. However, without an analysis of the following verb, we cannot be sure about the argumentative status of the extract. The sentence could continue with *"... used machine learning techniques for ... "*, in which case we have a solution instance; just as well, the sentence could be a conclusion (*"... argued that ... "*) or a problem statement (*"... attacked the hard problem of ... "*). Thus, the phrase *"in this paper we will"* is very useful for step one, but not useful for step two.

3.3 Corpus

Our corpus is a collection of 201 articles and their author-written summaries from different areas of computational linguistics and cognitive science, drawn from the computation and language archive (`http://xxx.lanl.gov/cmp-lg`). We assume that most of the articles had been accepted for publication in conference proceedings, although we have not verified this in each case. The documents were converted from LaTeX source into HTML in order to extract raw text and minimal structure automatically, then transformed into SGML format and manually corrected. We used all documents dated between 04/94 and 05/96 which we could semi-automatically retrieve with our conversion pipeline and which contained no less than 2,000 and no more than 10,000 words. The resulting corpus contains 568,000 word tokens; the average length of the documents is 187 sentences, the average length of the original summaries 4.7 sentences. In each text we marked up the following structural information: title, summary, headings, paragraph structure and sentences. We also removed tables, equations, figures, captions, references and cross references and replaced them by place holders (e.g. the symbol [REF] marks the place where a reference was cited in the text; [EQN] marks the place of equations).

We randomly divided our corpus into a training and test set of 123 documents which were further analyzed and annotated, and a remaining set of 78 documents which remain unseen. Only the first set was used for the experiments described here.

3.4 Annotation of gold standards

In line with Kupiec *et al.*'s method, we tried to use the summaries in our corpus for training and evaluation. However, the summaries of our articles were written by the authors themselves, and it is commonly assumed that author summaries are of a lower quality when compared to summaries by professional abstractors.

We first tested to which degree the authors' summaries reused sentences from the body of the document. In order to establish alignment between summary and document sentences, we used a semi-automatic method, assisted by a simple surface similarity measure which computed the longest common subsequence of non-stop-list words. Final alignment was decided by a human judge, where the criterion was similarity of semantic contents of the compared sentences. The following sentence pair illustrates a *direct match*:

> **Summary:** In understanding a reference, an agent determines his confidence in its adequacy as a means of identifying the referent.
>
> **Document:** An agent understands a reference once he is confident in the adequacy of its (inferred) plan as a means of identifying the referent.

Unlike Kupiec *et al.*'s professional annotators, our authors had not reused document sentences to a large degree—we had a low 31% alignment rate as compared to Kupiec *et al.*'s 79%.

In addition to this, the authors had obviously not used a prototypical scheme to write their summaries, in contrast to professional abstractors surveyed by Liddy (1991). When we inspected the rhetorical contents of the sentences in the author summaries by applying our annotation scheme to them, we found that argumentative structure varied widely, even though most summaries are understandable and many are well-written. Some summaries are extremely short, and many of them are not self-contained, and would thus be difficult to understand for the partially informed reader. This again confirms the claim that author summaries are less systematically constructed than summaries by professional abstractors.

Because of the low alignment and the heterogeneous rhetorical structure of the summaries, we decided not to use them directly for annotation and evaluation. Annotation of the training corpus had to proceed in the following three steps:

1. Alignment of summary and document sentences (semi-automatic);
2. Additional annotation of further relevant sentences (manual);
3. Annotation of the argumentative status of these sentences (manual).

A human judge annotated additional abstract-worthy sentences in the source text. We gave no restrictions as to how many additional sentences were to be selected. After this process, our texts had two gold standards of different origin: gold standard A, consisting of aligned sentences; and gold standard B, consisting of sentences selected by the human judge, 948 sentences in total.

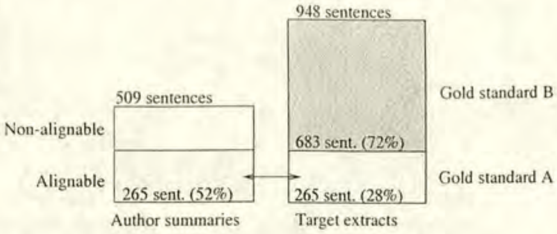


Figure 5: Composition of gold standards with respect to origin

Figure 5 shows the composition of gold standards: there are 2.5 times as many gold standard B sentences as there are gold standard A sentences. The alignment rate in our training and test set of 123 documents, which consists of the best-aligned documents, is 52% (the alignment rate of 31% refers to all 201 documents). With respect to compression (i.e. ratio of gold standard sentences to document sentences), our combined gold standards achieve 4.4% (as compared to Kupiec *et al.*'s 3.0% compression). Gold standard A had a compression of 1.2%, gold standard B 3.2%.

The second annotation step consisted of manually determining the argumentative roles for the abstract-worthy sentences (as defined in step one) for each article in the training set.

The following sentence with its rhetorical label illustrates this type of mark-up:

> Repeating the argument of Section 2, we conclude that a construction grammar that encodes the formal language [EQN] is at least an order of magnitude more compact than any lexicalized grammar that encodes this language. CONCLUSION/CLAIM

Difficulties encountered during annotation often concerned the status of a statement in the line of the argument, when the status was dependent on the context. For example, a weakness of the authors' solution might be classified as a limitation or as a local problem, depending on whether that problem will be solved later on in the given article. In cases of true ambiguity between two roles, we allowed for multiple annotation.

Another difficulty had to do with the fact that we annotated entire sentences: often, one sentence covers more than one role, as the following sentence illustrates:

> We also examined how utterance type related to topic shift and found that few interruptions introduced a new topic. PURPOSE/PROBLEM AND CONCLUSION/CLAIM

Figure 6 shows the composition of the gold standard sentences with respect to rhetorical roles. SOLUTION and PROBLEM are the most common rhetorical roles with about one third each of the judgements, the other roles sharing the last third. The least common role was RESULT.

There were 1172 instances of rhetorical roles in our 948 gold standard sentences. 232 sentences (24%) contained multiple mark-up (either ambiguous or concatenative). Figure 7 shows the distribution of *multiple* markup over the rhetorical roles, which is about proportional, except for a low involvement of BACKGROUND in multiple markup and a proportionally higher one for RELATED WORK and PROBLEM. We believe this is partly due to conceptual difficulties and partly due to concatenative markup: BACKGROUND sentences tend to contain nothing but background information, whereas the information units for PROBLEM statements and RELATED WORK tend to be smaller.

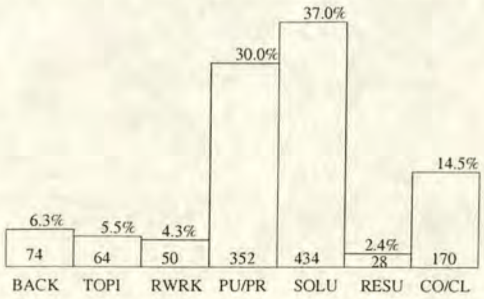


Figure 6: Composition of gold standard sentences with respect to rhetorical roles set

Rhetorical role	**Multiple annotation**	
BACKGROUND	16	(21%)
TOPIC/ABOUTNESS	25	(39%)
RELATED WORK	24	(48%)
PURPOSE/PROBLEM	168	(47%)
SOLUTION/METHOD	167	(38%)
RESULT	11	(39%)
CONCLUSION/CLAIM	64	(37%)

Figure 7: Percentages of judgements involving multiple annotation for the respective rhetorical roles

3.5 Heuristics Pool

We employed 7 heuristics in the two tasks: 4 of the heuristics used by Kupiec et al. (Indicator Quality Feature, Relative Location Feature, Sentence Length Feature and Thematic Word Feature), and 3 additional ones (Indicator Rhetorics Feature, Title Feature and Header Type Feature).

Indicator Quality Feature: The Indicator Quality Feature identifies meta-comments in a text, as opposed to subject matter. We use a list consisting of 1728 indicator phrases or formulaic expressions, such as communicative verbs and phrases related to argumentation and research activities. Our indicator phrase list was manually created by a cycle of inspection of extracted sentences and addition of indicator phrases to the list.

Figure 8 shows an extract from the indicator list: the first group of indicator phrases is centered around the concept *"argue"*, the second group uses the global indicator *"in this article"*, the third is centered around the concept *"attempt"*.

The largest part of these phrases is positive, but the last entry in Figure 8 illustrates a negative indicator phrase, typically occurring in the rhetorical division *Acknowledgements* (which is of no interest to content selection).

Indicator Phrase	Quality Score
we argued	2
we have argued	1
we have argued that	1
we will argue	1
what I have argued is	1
what we have argued is	1
This article	3
in this article	3
is an attempt to	1
I attempt to	2
I have attempted	2
I have attempted to	2
our work attempts	2
the present paper is an attempt	2
this paper is an attempt to	2
supported by grant	-1

Figure 8: An extract from the indicator list

Using the strings directly as values in a feature would result in a sparse distribution, and thus in an over-fitted feature, i.e. a feature that works well for the training data but not for different, but similar kinds of data. Thus, we classified the strings according to different criteria. For the Indicator Quality Feature, indicator phrases were manually classified into 5 quality classes according to their occurrence frequencies within the target extract sentences (cf. the column 'Quality Score' in Figure 8). The scores mirror the likelihood of a sentence containing the given indicator phrase to be included in the summary on a 5-valued scale from 'very likely to be included in a summary' to 'very unlikely'. For example, the likelihood of the phrase *"we argued"* to appear in the summary is higher than the likelihood of variations of this string in other tenses, a fact that is mirrored by its higher score of +2.

Indicator Rhetorics Feature: This feature tries to model the semantics (rhetorical contribution) of the phrases. Each indicator phrase was manually classified into one of 16 classes. Classes correspond to the 7 rhetorical roles (Back, Topi, Rwrk, Pu/Pr, Solu, Resu, Co/Cl), and 8 confusion classes, viz. Solu–Pu/Pr, Solu–Co/Cl, Pu/Pr–Co/Cl, Pu/Pr–Rwrk, Pu/Pr–Back, Co/Cl–Rwrk, Co/Cl–Resu, Back–Rwrk plus the value Zero for phrases that do not predict a specific rhetorical role. The first group of phrases in Figure 8 (*"argue"*), for example, was classified as a most likely indicator of the rhetorical class Conclusion/Claim, and the third group (*"attempt"*) was classified as an indicator of Purpose/Problem, whereas the second and fourth groups received the value Zero.

Relative Location Feature: This feature distinguishes peripheral sentences in the document and within each paragraph, assuming a hierarchical organization of documents and paragraphs. The algorithm is sensitive to prototypical headings (e.g. *Introduction*); if such headings cannot be found, it uses a fixed range of paragraphs (first 7 and last 3 paragraphs). Document final and initial areas receive different values, but paragraph initial and final sentences are collapsed into one group.

Sentence Length Feature: All sentences under a certain length (current threshold: 15 tokens including punctuation) receive a 0 score, all sentences above the threshold a 1 score.

Thematic Word Feature: This feature is a variation of the "Term-frequency times inverse document frequency" (tf.idf) feature, a document specific keyword weighing method which is commonly used in Information Retrieval (Salton and McGill, 1993). It tries to identify key words that are characteristic for the contents of the document, viz. those of a medium range frequency relative to the overall collection. The 10 top-scoring words according to the tf.idf method are chosen as thematic words; sentence scores are then computed as a weighted count of thematic words in a sentence, meaned by sentence length. The 40 top-rated sentences obtain score 1, all others 0.

Title Feature: Words occurring in the title are good candidates for document specific concepts. The Title Feature score of a sentence is the mean frequency of title word occurrences (excluding stop-list words). The 18 top-scoring sentences receive the value 1, all other sentences 0. We also experimented with taking words occurring in all headings into account (these words were scored according to the tf.idf method) but received better results for title words only.

Header Type Feature: The rhetorical division that a sentence appears in can be a good indication of its rhetorical status. The Header Type Feature uses a list of prototypical header key words like *discussion, introduction, concluding remarks, conclusions*. Each sentence is assigned one of 15 values, depending on the header it appears under. Headers are classified as one of 14 prototypical groups if they contain one or more of the header key words (or a morphological variant of it); otherwise (i.e. if they contain only domain-specific strings) they are classified as 'non-prototypical'.

3.6 Classifiers

As in Kupiec *et al.*'s (1995) experiment, each document sentence receives scores for each of the features, resulting in an estimate for the sentence's probability to also occur in the summary. This probability is calculated for each feature value as a combination of the probability of the feature-value pair occurring in a sentence which is in the summary (successful case) and the probability that the feature-value pair occurs unconditionally.

Kupiec *et al.*'s estimation for the probability that a given sentence is contained in the summary is:

$$P(s \in E|F_1, \ldots, F_k) \approx \frac{P(s{\in}E) \prod_{j=1}^{k} P(F_j|s{\in}E)}{\prod_{j=1}^{k} P(F_j)}$$

where

$P(s \in E|F_1, \ldots, F_k)$: Probability that sentence s in the source text is included in the intermediate extract E, given its feature values;
$P(s \in E)$: compression rate (constant);
$P(F_j| \; s \in E)$: probability of feature-value pair occurring in a sentence which is in the extract;
$P(F_j)$: probability that the feature-value pair occurs unconditionally;
k: number of feature-value pairs;
F_j: j-th feature-value pair.

For the second step, the probability that a certain sentence from the new base set (the intermediate extract) is associated with a rhetorical role is calculated analogously as follows:

$$P(e \in R_i|F_1, \ldots, F_k) \approx \frac{P(e{\in}R_i) \prod_{j=1}^{k} P(F_j|e{\in}R_i)}{\prod_{j=1}^{k} P(F_j)}$$

where

$P(e \in R_i|F_1, \ldots, F_k)$: Probability that sentence e in the intermediate extract is assigned the rhetorical role R_i, given its feature values;
$P(e \in R_i)$: probability of role R_i in extract (unconditional of feature values);
$P(F_j| \; e \in R_i)$: probability of feature-value pair occurring in an extract sentence which has rhetorical role R_i;
$P(F_j)$: probability that the feature-value pair occurs unconditionally in the extract;
k: number of feature-value pairs;
F_j: j-th feature-value pair.

Assuming statistical independence of the features, $P(F_j)$ (for the two different base sets), $P(F_j| \; s \in E)$ and $P(F_j|e \in R_i)$ can be estimated from the corpus for each F_j and each R_i. The second step returns a vector of probabilities for each sentence in a document (cf. Figure 9), with each cell in the vector corresponding to a rhetorical role. For each sentence, the role with the highest probability is chosen (cf. grey boxes).

BACK	TOPI	RWRK	PU/PR	SOLU	RESU	CO/CL	
0.9e-9	0.1e-9	0.2e-9	0.6e-10	0.3e-9	0.7e-10	0.9e-10	0
0.3e-12	0.9e-14	0.3e-14	0.6e-13	0.1e-17	0.7e-14	0.9e-12	1
0.6e-14	0.4e-10	0.9e-11	0.5e-10	0.3e-7	0.7e-8	0.1e-10	2
			. . .				
0.4e-8	0.9e-10	0.3e-9	0.5e-10	0.6e-8	0.7e-9	0.1e-10	235

Figure 9: Probability vectors for document sentences No. 0, 1, 2 and 235

3.7 Evaluation

The evaluation we report here is based on co-selection between the gold standard sentences (i.e. target extracts) and the automatic results. This kind of evaluation is useful in a corpus-based approach like ours to fine-tune the single heuristics, but in our opinion final evaluation should not be based on co-selection with target extracts. Co-selection measures might give a distorted picture of the quality of an extract, because there might be many good abstracts/extracts, but a comparison with a target can only ever measure how well it approximates *one* of these. Real evaluation should be task-based, i.e. measure how well a certain document surrogate supports a human in fulfilling a certain task.

In our experiments, co-selection measures were used as follows: for extraction, co-selection reports how many of the extracted sentences had independently been identified as relevant sentences by the human annotator. For classification, co-selection reports how often the rhetorical roles identified by the algorithm were indeed the roles the human annotator had assigned. The numerical results reported for classification refer to the intermediate extract as a base set (i.e. those sentences that have been correctly identified in the first step). Cross-validation is used: the model is trained on a training set of documents, leaving a single document out at a time (the current test document). We did not have an indication as to subject matter like Kupiec *et al.* did (by journal name), so we chose to use all other documents but the single test document for training. After training, the model is used to extract candidate sentences from the test document, and co-selection values are measured.

Numerical values in the tables always give precision and recall rates as percentages. Due to the setup of the experiment (there are always as many sentences chosen as there are gold standards), precision and recall values are identical for extraction and for the *overall* results of classification. However, it is possible that precision and recall values for the classification of a *specific* rhetorical role differ. This is because it is possible that the algorithm overestimates the frequency of one role X at the expense of another role Y, in which case the recall of X would increase, but the precision of X would decrease. For multiply-annotated gold standard sentences, a correct classification was scored when the algorithm identified *one* of the ambiguous roles correctly.

As a baseline for the first task we chose sentences from the beginning of the source text, which constituted a recall and precision of 28.0%. This "from-top" baseline is a more conservative baseline than random order: it is more difficult to beat, as prototypical document structure places a high percentage of relevant information in the beginning.

The baseline for the second task (classification) is computed by classifying each sentence as the most frequent role (SOLUTION); it stands at an amazing 40.1% which means that this task is statistically much easier than extraction.

3.8 Results

3.8.1 Extraction

Extraction	Indiv.	Cumul.
Indicator Quality Feature	54.4	54.4
Relative Location Feature	41.0	63.9
Sentence Length Feature	28.9	65.6
Title Feature	21.6	65.6
Header Type Feature	39.6	65.3
Thematic Word Feature	16.2	66.0
Indicator Rhetorics Feature	44.0	65.6
Baseline	28.0	

Figure 10: Impact of individual heuristics on extraction

Figure 10 summarizes the contribution of the features, individually and cumulatively. Precision and recall values for the features vary between 16.2% (Thematic Word Feature) and 54.4% (Indicator Quality Feature). The most successful combination of the 7 available heuristics at 66.0% actually excludes the Indicator Rhetorics Feature—including it would decrease the results slightly (by 0.4%). The fact that a subset of all heuristics achieves a better result than all heuristics taken together means that the combination of heuristics in our implementation is non-monotonic. Non-monotonicity would be an unfortunate property in a real world setting where there are no gold standards available, and where we have to rely on the fact that *each* heuristic in the pool contributes positively to the results. However, in the supervised experiments described here co-selection measures are used to fine-tune the heuristics, in order to identify weaknesses of features (or features that should be removed from the pool completely).

Also note that even such weak features as the Title Feature and Thematic Word Feature with precision and recall lower than the baseline can still contribute positively to the results, whereas the relatively strong Indicator Rhetorics Feature does not. This does not mean that the Indicator Rhetorics Feature is not a good feature, but only that it is not completely independent from the more successful features, contrary to assumption (in this case, it is probably very similar to the Indicator Quality Feature). Thus, how helpful a heuristic will be in combination with others cannot be judged from its individual performance alone, but also from its similarity to the other heuristics.

Overall, these results reconfirm the usefulness of Kupiec *et al.*'s method of heuristic combination. The method increases precision for the best feature by around 20%.

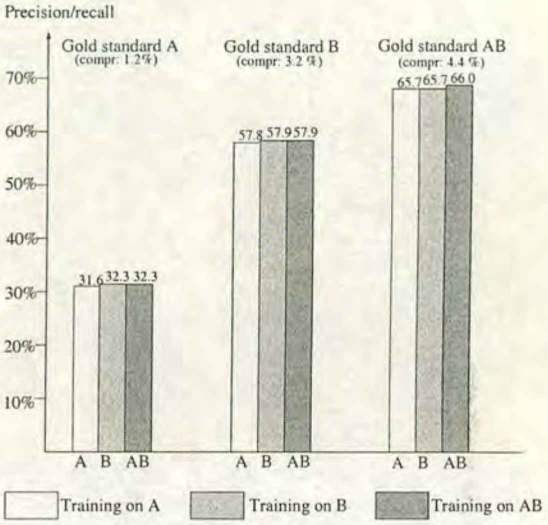


Figure 11: Influence of training material/gold standards

In order to see how the different origins of our gold standards contribute to the results, we trained three models (cf. Figure 11): one by training only on gold standard A sentences (light grey), one by training only on gold standards B (medium grey), and the third by training on both kinds of gold standards (dark grey). We then used the 3 models for 3 different tasks—first trying to identify A gold standards, then B gold standards and then both. Due to the higher compression of the task, extraction in the first task is statistically more difficult, which accounts for the much lower precision and recall values when compared to the other tasks. If we compare the values *within* extraction tasks, where the only difference is in *training*, the results show a surprising consistency: the distribution of heuristics values was almost identical between gold standards, no matter which gold standards we had trained our model on. The practical conclusion from this experiment is that we can get intermediate extracts of a similar quality (if we were to be content with these as end results) by training only on the relatively cheaply attainable gold standard A (alignment), rather than using the labor-intensive gold standard B (human judgement).

3.8.2 Classification

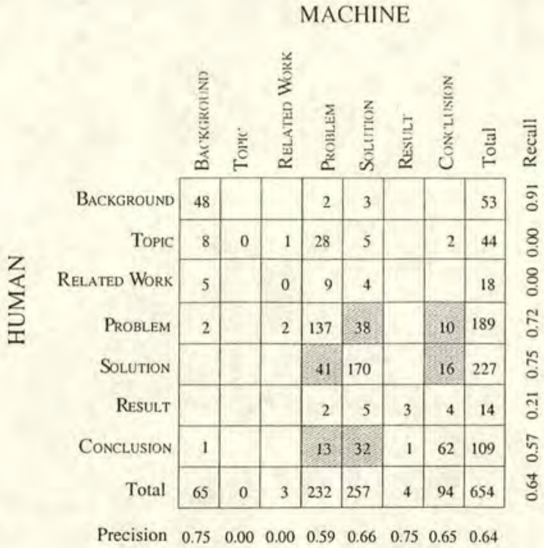
Classification	Indiv.	Cumul.
Indicator Rhetorics Feature	56.3	56.3
Relative Location Feature	46.5	63.8
Title Feature	40.0	64.2
Indicator Quality Feature	45.9	63.8
Sentence Length Feature	39.7	61.6
Thematic Word Feature	16.2	61.5
Header Type Feature	39.6	57.2
Baseline	40.1	

Figure 12: Impact of individual heuristics on classification

Figure 12 summarizes the contribution of the individual features for classification, taken individually and cumulatively. Precision and recall values for the features vary between 16.2% (Thematic Word Feature) and 56.3% (Indicator Rhetorics Feature). The most successful combination consisted of Indicator Rhetorics Feature, Relative Location Feature and Title Feature (with a combined precision/recall value of 64.2%). The combination is non-monotonic to a higher degree than in the extraction task: addition of the other 4 heuristics steadily decreased precision and recall to 57.2%.

Where does the system make errors? The confusion matrix in Figure 13 shows the distribution of machine and human classifications for the different roles (best heuristic combination), where the columns in the table refer to the roles assigned by our algorithm ("Machine") and the rows denote roles assigned in the gold standard sentences ("Human"). For example, out of the 227 SOLUTION gold standard sentences that the human judge identified, the system found 170 correctly; it misclassified 41 as PROBLEM and the remaining 16 as CONCLUSION. The grey boxes along the diagonal show the absolute numbers of successful machine classifications per role; also, precision and recall values of the automatic classification are given for each rhetorical role.

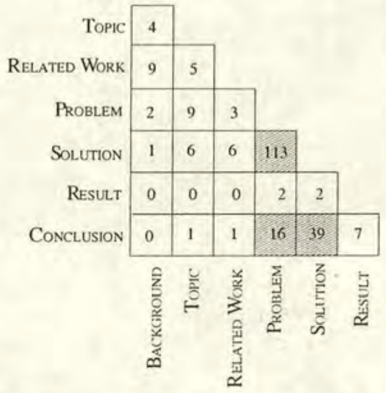
It is obvious that the system significantly underestimates low-frequency roles—there are only very few RELATED WORK and RESULT roles assigned by the system, and none at all for TOPIC. In comparison, the estimation of the frequency of the higher frequency roles is quite adequate.



MACHINE

HUMAN		BACKGROUND	TOPIC	RELATED WORK	PROBLEM	SOLUTION	RESULT	CONCLUSION	Total	Recall
	BACKGROUND	48			2	3			53	0.91
	TOPIC	8	0	1	28	5		2	44	0.00
	RELATED WORK	5		0	9	4			18	0.00
	PROBLEM	2		2	137	▨ 38		▨ 10	189	0.72
	SOLUTION				▨ 41	170		▨ 16	227	0.75
	RESULT				2	5	3	4	14	0.21
	CONCLUSION	1			▨ 13	▨ 32	1	62	109	0.57
	Total	65	0	3	232	257	4	94	654	0.64
	Precision	0.75	0.00	0.00	0.59	0.66	0.75	0.65	0.64	

Figure 13: Confusion matrix for argumentative classification by roles (machine)



	BACKGROUND	TOPIC	RELATED WORK	PROBLEM	SOLUTION	RESULT
TOPIC	4					
RELATED WORK	9	5				
PROBLEM	2	9	3			
SOLUTION	1	6	6	▨ 113		
RESULT	0	0	0	2	2	
CONCLUSION	0	1	1	▨ 16	▨ 39	7

Figure 14: Number of sentences involved in multiple markup (gold standards)

The confusion matrix illustrates that our system often misclassifies PROBLEMS as SOLUTIONS (38 times) and SOLUTIONS as PROBLEMS (41 times). But these roles are often co-classified by the human judge, as Figure 14 shows: 113 out of the 434 SOLUTION instances and the 352 PROBLEM instances were co-classifications "PROBLEM and/or SOLUTION". Apart from ambiguities between PROBLEM and SOLUTION, there were also many misclassifications including these roles and CONCLUSION (cf. the hatched boxes in Figure 14). These were exactly the ones where our algorithm had a high percentage of misclassifications (cf. the hatched boxes in Figure 13), which implies that the low performance of the system must be partly attributed to the inherent difficulty of the task. The distinction between these roles is conceptually difficult: conclusions are often statements *about* properties of the solution or *about* phenomena in the world (which are annotated as problems); problems and solutions co-occur often in the same sentence, and sometimes it is difficult to distinguish between a research goal and its solution, i.e. to find out if the sentence describes a goal in itself or a research step towards the main goal. This decision is particularly hard where the status of the sentence is not linguistically marked. In that case, only inference on the argumentation in the article as a whole might help a human judge disambiguate, a possibility obviously not open to our system.

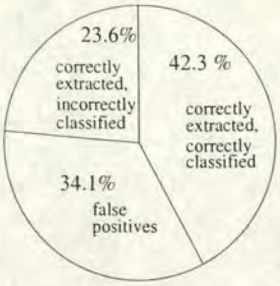


Figure 15: Overall results

Overall results for both tasks of the experiment (extraction and classification) are shown in Figure 15. At our high compression of 4.4%, 42.3% of all gold-standard sentences have been both correctly extracted and classified. This number includes the cases where one of several ambiguous roles has been identified correctly. A further 23.6% of the presented sentences can be counted as almost correct; they have been correctly extracted but have been assigned the wrong rhetorical role. 34.1% of all sentences are false positives, i.e. they should not have been extracted at all because the human annotator had not marked them.

Figure 16 shows a typical example of a rhetorically annotated extract. It is the output of our system after processing the first article in our collection, cmp_lg-9404003. Examples for correctly extracted and classified sentences are sentences 0 and 4, and also sentences 235, 236 and 238 (where one role was correctly identified). Correctly extracted, but incorrectly classified, are sentences 2 and 7. In our example, the only false positive is sentence 8.

The example also shows just how difficult rhetorical classification is. Consider sentence 7—a point could be made for the system's classification as well as for the human classification. Is "redefinition of synchronous TAG derivation" the Topic of the paper, or is it the Solution? Or is the Problem "How can synchronous TAG derivation be redefined?" One of these possibilities had to be chosen by objective criteria, which are documented in the coding manual for the annotation task.

4 Discussion

We find the results encouraging: with shallow processing, in a high-compression task, our algorithm finds 66% of all marked-up gold standard sentences in our training text and subsequently associates the right role for 64% of the correctly extracted sentences. Even though these results are only measurements of co-selection, they support our hypothesis that argumentative document structure can be approximated by low-level properties of the sentence. We see our prototype as a shallow document structure analyzer, specially designed for scientific text and geared towards the kinds of meta-

	Machine	Human
0 The formalism of synchronous tree-adjoining grammars [REF], a variant of standard tree-adjoining grammars (TAG), was intended to allow the use of TAGs for language transduction in addition to language specification.	Back	Back
2 This paper concerns the formal definitions underlying synchronous tree-adjoining grammars.	Solu	Topi
4 This sort of rewriting definition of derivation is problematic for several reasons.	Prob	Prob
7 In this paper, we describe how synchronous TAG derivation can be redefined so as to eliminate these problems.	Prob	Solu Topic
8 The redefinition relies on an independent redefinition of the notion of tree-adjoining derivation [REF] that was motivated completely independently of worries about the generative capacity of synchronous TAGs, but which happens to solve this problem in an elegant manner.	Prob	—
235 We have introduced a simple, natural definition of synchronous tree-adjoining derivation, based on isomorphisms between standard tree-adjoining derivations, that avoids the expressivity and implementability problems of the original rewriting definition.	Solu	Solu Prob
236 The decrease in expressivity, which would otherwise make the method unusable, is offset by the incorporation of an alternative definition of standard tree-adjoining derivation, previously proposed for completely separate reasons, that allows for multiple adjunctions at a single node in an elementary tree.	Prob	Solu Prob
238 Nonetheless, some remaining problematic cases call for yet more flexibility in the definition; the isomorphism requirement may have to be relaxed.	Solu	Solu Rwrk

Figure 16: Example of a rhetorically annotated extract, with gold standard judgement ("Human")

linguistic, argumentative constructs typically found in this text type.¡

However, our approach crucially depends on the quality of the indicator list. As our indicator list is handcrafted, (i.e. gained during the reading and annotation of the 123 papers in our training corpus), as opposed to automatically acquired, one might be suspicious of its performance—it might be over-fitted to the data, i.e. too dependent on phrases that occur only rarely rather than relying on generic phrases. As a result, it might not generalize well to other documents from the same source. The first question is thus how robust the indicator list is to different data of the same source.

In order to test the robustness of the list, we need *unseen* data, i.e. documents which were not taken into account when building the system or its knowledge sources, but for which gold standard judgements exist. As the process of annotation and indicator phrase addition happened simultaneously in our experiment, we do not have gold standards for the unseen part of our corpus. But we can simulate 'unseen' data as follows. We compare versions of our indicator phrase list before and after the annotation process for the last third of our training set (42 documents). Before the annotation process for that third, the indicator phrase list already contained 1501 indicator phrases; the annotation process for the last third only contributed another 262 phrases. When using the indicator list before the annotation process, the last third of the training data is practically treated as unseen: only indicator phrases are used that already occurred in the first two thirds of our training corpus. We report results only for the Indicator Features, because the performance of the other heuristics would not change by the analysis of more data. The results (Figure 17) show that there is only a minor decrease in performance if the first list is used (left column). This means that the indicator list, even though hand-crafted, is robust and general enough for our purposes; it generalizes reasonably well to texts of a similar kind, viz. research articles in computational linguistics of around 6 to 20 pages in length.

Another question is how well the list of phrases we collected might scale up to other domains. We make no claims about other *text types*, e.g. newspaper articles on scientific topics, or articles in *Scientific American*; our method depends on the explicitness of meta-linguistic information of scientific research articles which is not necessarily present in other text types.

We are interested in different domains, however, because we believe that the definition of rhetorical roles in our annotation scheme are generic rhetorical steps in scientific research papers. We are now planning to move to articles in the medical domain, in order to validate this hypothesis. With our corpus already consisting of articles from different sub-domains of computational linguistics, we are confident that performance should be similar in different domains as long as we have the right indicator phrases available. In the light of these considerations, the main challenge is to make the indicator features more adaptive to new text. What is needed is a method for the automatic and reliable acquisition of indicator phrases from corpus data, so that indicators get recognized even if the linguistic expression found is not identical, but only similar to one of the examples in the list.

Extraction		
	Last 42 files treated as	
Heuristics used	seen	unseen
Indicator Quality Feature	57.62	54.32
Indicator Rhetorics Feature	47.76	44.48
Indicator Quality, Title, Sentence Length, and Header Type Features	68.36	64.78
Baseline	25.67	

Classification		
	Last 42 files treated as	
Heuristics used	seen	unseen
Indicator Quality Feature	50.21	49.36
Indicator Rhetorics Feature	56.47	55.79
Indicator Rhetorics, Relative Location and Title Features	61.37	60.26
Baseline	45.26	

Figure 17: Difference between seen and unseen data

We have run some preliminary experiments in indicator list acquisition. We used a simple method: using the gold standard sentences as a base, we compiled frequency lists of strings of different length occurring under each rhetorical role. Because subject matter specific strings get automatically cancelled out during this procedure, we ended up with a proto-list of around 500 very frequently occurring indicator phrases. In the extraction/classification experiment, this list performed about 30% below our hand-crafted list, a drop in performance which we believe to be mostly due to the fact that the new list is very short compared to the manually created list. On the positive side, the automatically created list is very unlikely to be over-fitted to our data. Further research could aim at improving this baseline by taking more sophisticated criteria like statistical interaction between the words in phrases into account, and by using different similarity measures to cluster similar phrases together.

In our approach, the rhetorically annotated extracts are collections of *sentences*. Although sentences are a natural choice of information unit when the collection of sentences is itself the abstract, there are several reasons why sentences are *not* the ideal information units for the approach we take. One problem is that as sentences are rhetorically connected to previous ones, they might not mean the same thing in isolation. They certainly don't look the same: Salager (1992), who analyzed summaries in the medical domain for the use of hedging and their rhetorical structure, found that in summaries claims are stated boldly without explanations or comments, whereas in the full article a sentence conveying the same information tends to be formulated much more tentatively and with a higher level of reserve. Thus, it is unlikely that we will be able to use sentences extracted from the body of the text without change. The main problem is that sentences are too large a unit for rhetorical annotation and extraction, as became apparent during the human annotation phase: ideally, one would like to annotate and extract a unit that corresponds to a proposition, i.e. a clause. However, due to problems of ambiguity between sentential and phrasal coordination (and subordination), it is difficult to find clauses automatically with low-level tools like tokenizers. For now, we have to content ourselves with sentences as our selection unit for purely practical reasons. The sentence-based approach put forward here achieves good results, which might be improved later by a more sophisticated unit identification.

One of the main motivations behind our definition of rhetorical roles found in scientific articles is that this classification is intuitive to humans. This could be relevant for the procedure of how gold standards for training are gained. Typically, when human annotation is used to define gold standards for sentence extraction (Zechner, 1995; Marcu, 1997), the instructions to the annotators are vague and phrased in terms of importance ("annotate important sentences"). Due to the subjectivity and task dependence of the term 'important', such instructions usually result in individually varying annotations. If our claim that our annotation scheme defines relevance criteria in a more objective way is true, a definition of importance in terms of these rhetorical roles should make the task of annotating gold standards easier.

An experiment is currently underway to substantiate this claim. We have written a coding manual, i.e. an operational description of how the rhetorical roles are to be annotated, based on Swales' rhetorical moves, indicator phrases, and context. In the experiment, we compare the inter-annotator reliability of annotators who have read the annotation guidelines to that of two control groups: a second group who has been instructed to mark instances of the seven rhetorical roles without any further instructions, and a third group which had only been instructed to annotate important sentences. If our definitions of the rhetorical roles can be conveyed to other humans operationally, group 1 will have the highest inter-annotator reliability. If they are intuitive, group 2 will annotate similarly to group 1. Inter-annotator reliability should be higher in either group 1 or 2 than in group 3.

The usability of gold standards gained in an annotation based on our rhetorical roles will have to be established in an independent, task-based evaluation.

5 Related Work

Paice (1981) was probably the first attempt at implementing an extraction mechanism for physics articles that relied on pattern-matching operations, based on indicator phrases. Indicator phrases have been frequently used since then (Johnson et al., 1993). Paice and Jones (1993) made the method more flexible by supplying a finite state grammar for indicator phrases specific to the agriculture domain. However, we are the first to explicitly use the rhetorical status of indicator phrase for extraction and rhetorical classification.

There is a similar notion of *cue* phrases, typically used in discourse analysis, which is closely related to our notion of indicator phrases. Cohen (1987) defines cue words as all words and phrases used by the speaker to directly indicate the structure of the argument to the hearer. Cue phrases are typically short and come from a closed-class vocabulary (e.g. adverbials or sentence connectives (Litman, 1996)). As a result, the linguistic realization of the cue phrases between different authors tends to be invariant. Our indicator phrases, on the other hand, are longer and more variable; because they depend on the individual writing style, they are more difficult to identify automatically.

Rhetorical Structure Theory (RST) defines local rhetorical relations between sentences and clauses (Mann and Thompson, 1987), in order to build up a fixed rhetorically annotated tree structure through a complete rhetorical analysis of the text. There are automatic procedures for recognizing RST relations, either heuristically (Miike et al., 1994; Sumita et al:, 1992) or by full rhetorical parse (Marcu, 1997) .

There are some analogies between these approaches and the analysis proposed in this paper, even though they are not obvious. We, too, believe that the main discourse structure of a paper is a hierarchical, rhetorically annotated tree structure. The branches are annotated differently, but one could argue that our rhetorical roles are text-type specific realizations of RST relations.

We believe that the upper parts of the tree are more important for abstracting than the lower level parts. Unlike RST, we are not concerned with rhetorical relations between each sentence or clause, but we concentrate on the higher levels of the tree, what we call *global* rhetorical relations: relations of content units with respect to the content of the whole article. We use indicator phrases which mark global rhetorical moves, rather than those that mark rhetorical relations between sentences or clauses.

As a result, we can perform a robust rhetorical analysis without the need for a *full* analysis. Our two-step approach ensures that we find global fillers for this flat tree structure with a reasonably high confidence level, at the cost of some detail in the lower areas of the tree. Indeed, the annotation scheme described in this chapter only allows us to build rhetorical trees which are one level deep.

Of course the representation of the text's structure as a flat tree is a simplification. In principle, our annotation scheme could be extended to include these lower-level relations (e.g. the subproblem relationship between two problems), with intersegment relations holding at each level (e.g. the problem-solution relationship between a given problem and a solution when more than one problem is mentioned). This more detailed analysis may prove useful for the construction of longer and even more modular abstracts. But we believe that many of the local rhetorical relations between sentences and clauses are not immediately important for robust high-compression abstracting.

Our use of meta-linguistic information makes our approach different from methods which aim at representing the *contents* of the text. Lexical cohesion methods (Barzilay and Elahad, 1997), like statistical, keyword based methods, model main document concepts shallowly by using presumably content-specific lexical items observed in the text. Our method, in contrast, employs structural heuristics alone and uses *everything else* in the text but the content-specific lexical items.

Having said this, we can very well envisage our method cooperating with a complementary module that is based on an analysis of content rather than form. In a larger summarizing system, information from both types of module could flow together, in order to fulfill the tasks needed for generating abstracts from rhetorically annotated extracts: finding duplicate fillers, deciding on the best candidates for a filler, and resolving conflicts between fillers.

6 Conclusion

Robust, high-compression abstracting can be improved greatly if the discourse structure of the text is taken into account. We have argued that rhetorical classification of extracted material is a useful subtask for the production of a new kind of abstract that can be tailored in length and focus to users' expertise and specific information needs.

Our goal is to recognize abstract-worthy sentences with respect to global rhetorical structure, and to perform a subsequent classification of these sentences into a set of predefined rhetorical roles. We have presented a robust method which uses supervised learning techniques to deduce rhetorical roles from lower-level properties of sentences. This is technically feasible, because restrictions with respect to the task of the reader on the one hand, and knowledge about the typical argumentation of the writers on the other hand, can be exploited.

The results are encouraging; our algorithm determines 66% of all marked-up gold standards sentences in our training text and subsequently associates the right role for 64% of the correctly extracted sentences.

7 Acknowledgements

Data collection of our corpus took place collaboratively with Byron Georgantopolous. The first author is supported by an EPSRC studentship.

References

Ad Hoc Working Group For Critical Appraisal Of The Medical Literature. 1987. A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine.*

Alley, M. 1996. *The craft of scientific writing.* Englewood Cliffs, N.J.: Prentice-Hall.

American National Standards Institute, Inc. 1979. American national standard for writing abstracts. Technical report, American National Standards Institute, Inc., New York. ANSI Z39.14.1979.

Barzilay, R., and Elahad, M. 1997. Using lexical chains for text summarization. In Mani, I., and Maybury, M. T., eds., *Proceedings of the ACL/EACL-97 workshop on Intelligent Scalable Text Summarization.* Association for Computational Linguistics.

Baxendale, P. B. 1958. Man-made index for technical literature – an experiment. *IBM journal on research and development* 2(4):354–361.

Bazerman, C. 1988. *Shaping writing knowledge.* Madison: University of Wisconsin Press.

Borko, H., and Chatman, S. 1963. Criteria for acceptable abstracts: a survey of abstractors' instructions. *American Documentation* 14(2):149–160.

Broer, J. W. 1971. Abstracts in block diagram form. *IEEE Transactions on Engineering Writing and Speech* 14(2):64–67. ISA, 72-1626.

Buxton, A. B., and Meadows, A. J. 1978. Categorization of the information in experimental papers and their author abstracts. *Journal of Research in Communication Studies* 1:161–182.

Cohen, R. 1987. Analyzing the structure of argumentative discourse. *Computational Linguistics* 13:11–24.

Cremmins, E. T. 1996. *The art of abstracting.* Information Resources Press.

Day, R. A. 1995. *How to write and publish a scientific paper.* Cambridge: Cambridge University Press.

Edmundson, H. P. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery* 16(2):264–285.

Hartley, J., and Sydes, M. 1997. Are structured abstracts easier to read than traditional ones? *Journal of Research in Reading* 20(2):122–136.

Hartley, J.; Sydes, M.; and Blurton, A. 1996. Obtaining information accurately and quickly: are structured abstracts more efficient? *Journal of Information Science* 22(5):349–356.

International Organisation for Standardisation. 1976. Documentation – Abstracts for publication and documentation. Technical report, International Organisation for Standardisation. ISO 214-1976.

Johnson, F. C.; Paice, C. D.; Black, W. J.; and Neal, A. P. 1993. The application of linguistic processing to automatic abstract generation. *Journal of Document and Text Management* 1(3):215–42.

Kintsch, W., and van Dijk, T. A. 1978. Toward a model of text comprehension and production. *Psychological Review* 85(5):363–394.

Kircz, J. G. 1991. The rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of Documentation* 47(4):354–372.

Kupiec, J.; Pedersen, J. O.; and Chen, F. 1995. A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference*, 68–73.

Liddy, E. D. 1991. The discourse-level structure of empirical abstracts: an exploratory study. *Information Processing and Management* 27(1):55–81.

Litman, D. J. 1996. Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research* 5:53–94.

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2):159–165.

Mann, W. C., and Thompson, S. A. 1987. Rhetorical structure theory: A theory of text organisation. Technical report, Information Sciences Institute, U of South California. ISI/RS-87-190.

Manning, A. 1990. Abstracts in relation to larger and smaller discourse structures. *Journal of Technical Writing and Communication* 20(4):369–390.

Marcu, D. 1997. *The rhetorical parsing, summarization, and generation of natural language texts*. Ph.D. Dissertation, University of Toronto.

Miike, S.; Itoh, E.; Ono, K.; and Sumita, K. 1994. A full-text retrieval system with a dynamic abstract generation function. In *Proceedings of the 17th ACM-SIGIR Conference, Association for Computing Machinery, Special Interest Group Information Retrieval*, 152–163.

Milas-Bracovic, M. 1987. The structure of scientific papers and their author abstracts. *Informatologia Yugoslavica* 19(1-2):51–67.

O'Hara, K., and Sellen, A. 1997. A comparison of reading paper and on-line documents. In *Proceedings of CHI-97, Special Interest Group on Computer & Human Interaction.*

Paice, C. D., and Jones, A. P. 1993. The identification of important concepts in highly structured technical papers. In *Proceedings of the Sixteenth Annual International ACM-SIGIR conference on research and development in IR, Association for Computing Machinery, Special Interest Group Information Retrieval.*

Paice, C. D. 1981. The automatic generation of literary abstracts: an approach based on the identification of self-indicating phrases. In Oddy, R. N.; Robertson, S. E.; van Rijsbergen, C. J.; and Williams, P. W., eds., *Information Retrieval Research*. London: Butterworth. 172–191.

Pinelli, T. E.; Cordle, V. M.; and Vondran, R. F. 1984. The function of report components in the screening and reading of technical reports. *Journal of Technical Writing and Communication* 14(2):87–94.

Rennie, D., and Glass, R. M. 1991. Structuring abstracts to make them more informative. *Journal of the American Medical Association* 266(1).

Rowley, J. 1982. *Abstracting and indexing*. London: Bingley.

Salager-Meyer, F. 1992. A text-type and move analysis study of verb tense and modality distributions in medical English abstracts. *English for Specific Purposes* 11:93–113.

Salton, G., and McGill, M. J. 1983. *Introduction to Modern Information Retrieval.* Tokyo: McGraw-Hill.

Salton, G.; Allan, J.; Buckley, C.; and Singhal, A. 1994. Automatic analysis, theme generation, and summarisation of machine readable texts. *Science* 264:1421–1426.

Skorochod'ko, E. F. 1972. Adaptive method of automatic abstracting and indexing. In *Information Processing 71*, volume 2. North Holland Publishing company. 1179–1182.

Sumita, K.; Ono, K.; Chino, T.; Ukita, T.; and Amaro, S. 1992. A discourse structure analyzer for japanese text. In *Proceedings of the International Conference on Fifth Generation Computer Systems.*

Swales, J. 1981. Aspects of article introductions. Aston ESP Research Project No. 1. Technical report, The University of Aston, Birmingham, U.K.

Swales, J. 1990. *Genre analysis: English in academic and research settings.* Cambridge University Press.

Zechner, K. 1995. Automatic text abstracting by selecting relevant passages. Master's thesis, Centre for Cognitive Science, University of Edinburgh.

E.6. Teufel and Moens (1999b)

Discourse-level argumentation in scientific articles: human and automatic annotation

Simone Teufel and Marc Moens
HCRC Language Technology Group
Division of Informatics
University of Edinburgh
S.Teufel@ed.ac.uk, M.Moens@ed.ac.uk

Abstract

In this paper we present a rhetorically defined annotation scheme which is part of our corpus-based method for the summarisation of scientific articles. The annotation scheme consists of seven non-hierarchical labels which model prototypical academic argumentation and expected intentional 'moves'. In a large-scale experiments with three expert coders, we found the scheme stable and reproducible. We have built a resource consisting of 80 papers annotated by the scheme, and we show that this kind of resource can be used to train a system to automate the annotation work.

1 Introduction

Work on summarisation has suffered from a lack of appropriately annotated corpora that can be used for building, training and evaluating summarisation systems. Typically, corpus work in this area has taken as its starting point texts target summaries: abstracts written by the researchers, supplied by the original authors or provided by professional abstractors. Training a summarisation system then involves learning the properties of sentences in those abstracts and using this knowledge to extract similar abstract-worthy sentences from unseen texts. In this scenario, system performance or development progress can be evaluated by taking texts in a test sample and comparing the sentences extracted from these texts with the sentences in the target abstract.

But this approach has a number of shortcomings. First, sentence extraction on its own is a very general methodology, which can produce extracts that are incoherent or under-informative especially when used for high-compression summarisation (i.e. reducing a document to a small percentage of its original size). It is difficult to overcome this problem, because once sentences have been extracted from the source text, the context that is needed for their interpretation is not available anymore and cannot be used to produce more coherent abstracts (Spärck Jones, 1998).

Our proposed solution to this problem is to extract sentences but also to *classify* them into one of a small number of possible argumentative roles, reflecting whether the sentence expresses a main goal of the source text, a shortcoming in someone else's work, etc. The summarisation system can then use this information to generate template-like abstracts: Main goal of the text:...; Builds on work by:...; Contrasts with:...; etc.

Second, the question of what constitutes a useful gold standard has not yet been solved satisfactorily. Researchers developing corpus resources for summarisation work have often defined their own gold standard, relying on their own intuitions (see, e.g. Luhn, 1958; Edmundson, 1969) or have used abstracts supplied by authors or by professional abstractors as their gold standard (e.g. Kupiec et al., 1995; Mani and Bloedorn, 1998). Neither approach is very satisfactory. Relying only on your own intuitions inevitably creates a biased resource; indeed, Rath et al. (1961) report low agreement between human judges carrying out this kind of task. On the other hand, using abstracts as targets is not necessarily a *good* gold standard for comparison of the systems' results, although abstracts are the only kind of gold standard that comes for free with the papers. Even if the abstracts are written by professional abstractors, there are considerable differences in length, structure, and information content. This is due to differences in the common abstract presentation style in different disciplines and to the projected use of the abstracts (cf. Liddy, 1991). In the case of our corpus, an additional problem was the fact that the abstracts are written by the authors themselves and thus susceptible to differences in individual writing style.

For the task of summarisation and relevance decision between similar papers, however, it is essential that the information contained in the gold standard is *comparable* between papers. In our approach, the vehicle for comparability of information is similarity in argumentative roles of the associated sentences. We argue that it is more difficult to find the kind of information that preserves similarity of argumentative roles, and that it is not guaranteed that it will occur in the abstract.

A related problem concerns fair evaluation of the extraction methodology. The evaluation of extracted material necessarily consists of a comparison of *sentences*, whereas one would really want to compare the informational *content* of the extracted sentences and the target abstract. Thus it will often be the case that a system extracts a sentence which in that form does not appear in the supplied abstract (resulting in a low performance score) but which is nevertheless an abstract-worthy sentence. The mismatch often arises simply because a similar idea is expressed in the supplied abstract in a very different form. But comparison of content is difficult to perform: it would require sentences to be mapped into some underlying meaning representations and then comparing these to the representations of the sentences in the gold standard. As this is technically not feasible, system performance is typically performed against a fixed gold standard (e.g. the aforementioned abstracts), which is ultimately undesirable.

Our proposed solution to this problem is to build a corpus which details not only what the abstract-worthy sentences are but also what their argumentative role is. This corpus can then be used as a resource to build a system to similarly classify sentences in unseen texts, and to evaluate that system. This paper reports on the development of a set of such argumentative roles that we have been using in our work.

In particular, we employ human intuition to annotate argumentatively defined information. We ask our annotators to classify *every* sentence in the source text in terms of its argumentative role (e.g. that it expresses the main goal of the source text, or identifies open problems in earlier work, etc). Under this scenario, system evaluation is no longer a comparison of extracted sentences against a supplied abstract, or against a single sentence that was chosen as expressing (e.g.) the main goal of the source text. Instead, *every* sentence in the source text which expresses the main goal will have been identified, and the system's performance is evaluated against *that* classification.

Of course, having someone annotate text in this way may still lead to a biased or careless annotation. We therefore needed an annotation scheme which is simple enough to be usable in a stable and intuitive way for several annotators. This paper also reports on how we tested the stability of the annotation scheme we developed. A second design criterion for our annotation scheme was that we wanted the roles to be annotated automatically. This paper reports on preliminary results which show that the annotation process can indeed be automated.

To summarise, we have argued that discourse structure information will improve summarisation. Other researchers (Ono et al., 1994; Marcu, 1997) have argued similarly, although most previous work on discourse-based summarisation follows a different discourse model, namely Rhetorical Structure Theory (Mann and Thompson, 1987). In contrast to RST, we stress the importance of rhetorical moves which are *global* to the argumentation of the paper, as opposed to more local RST–type relations. Our categories are not hierarchical, and they are much less fine-grained than RST-relations. As mentioned above, we wanted them to a) provide context information for flexible summarisation, b) provide a higher degree of comparability between papers, and c) provide a fairer evaluation of superficially different sentences.

In the rest of this paper, we will first describe how we chose the categories (section 2). Second, we had to construct training and evaluation material such that we could be sure that the proposed categorisation yielded a reliable resource of annotated text to train a system against, a gold standard. The human annotation experiments are reported in section 3. Finally, in section 4, we describe some of the automated annotation work which we have started recently and which uses a corpus annotated according to our scheme as its training material.

2 The annotation scheme

The domain in which we work is that of scientific research articles, in particular computational linguistics articles. We settled on this domain for a number of reasons. One reason is that it is a domain we are familiar with, which helps for intermediate evaluation of the annotation work. The other reason is that computational linguistics is also a rather heterogeneous domain: the papers in our collection cover a wide range of subject matters, such as logic programming, statistical language modelling, theoretical semantics and computational psycholinguistics. This makes it a challenging test bed for our

BASIC SCHEME	BACKGROUND	Sentences describing some (generally accepted) background knowledge	**FULL SCHEME**
	OTHER	Sentences describing aspects of some specific other research in a neutral way (excluding contrastive or BASIS statements)	
	OWN	Sentences describing any aspect of the own work presented in this paper – except what is covered by AIM or TEXTUAL, e.g. details of solution (methodology), limitations, and further work.	
	AIM	Sentences best portraying the particular (main) research goal of the article	
	TEXTUAL	Explicit statements about the textual section structure of the paper	
	CONTRAST	Sentences contrasting own work to other work; sentences pointing out weaknesses in other research; sentences stating that the research task of the current paper has never been done before; direct comparisons	
	BASIS	Statements that the own work uses some other work as its basis or starting point, or gets support from this other work	

Figure 1: Overview of the annotation scheme

scheme which we hope to be applicable in a range of disciplines.

Despite its heterogeneity, our collection of papers does exhibit predictable rhetorical patterns of scientific argumentation. To analyse these patterns we used Swales' (1990) CARS (Creating a Research space) model as our starting point.

The annotation scheme we designed is summarised in Figure 1. The seven categories describe argumentative roles with respect to the overall communicative act of the paper. They are to be read as mutually exclusive labels, one of which is attributed to each sentence in a text. There are two kinds of categories in this scheme: *basic* categories and *non-basic* categories. Basic categories are defined by attribution of intellectual ownership; they distinguish between:

- statements which are presented as generally accepted (BACKGROUND);
- statements which are attributed to other, specific pieces of research outside the given paper, including the authors' own previous work (OTHER);
- statements which describe the authors' own *new* contributions (OWN).

The four additional (non-basic) categories are more directly based on Swales' theory. The most important of these is AIM, as this move on its own is already a good characterisation of the entire paper, and thus very useful for the generation of abstracts. The other categories are TEXTUAL, which provides information about section structure that might prove helpful for subsequent search steps. There are two moves having to do with the author's attitude towards previous research, namely BASIS and CONTRAST. We expect this kind of information to be useful for the creation of typed links for bibliometric search tools and for the automatic determination of rival approaches in the field and intellectual ancestry of methodologies (cf. Garfield's (1979) classification of the function of citation within researchers' papers).

The structure in Figure 2, for example, displays a common rhetorical pattern of scientific argumentation which we found in many introductions. A BACKGROUND segment, in which the history and the importance of the task is discussed, is followed by a longer sequence of OTHER sentences, in which specific prior work is described in a neutral way. This discussion usually terminates in a criticism of the prior work, thus giving a motivation for the own work presented in the paper. The next sentence typically states the specific goal or contribution of the paper, often in a formulaic way (Myers, 1992).

Such regularities, where the segments are contiguous, non-overlapping and non-hierarchical, can be

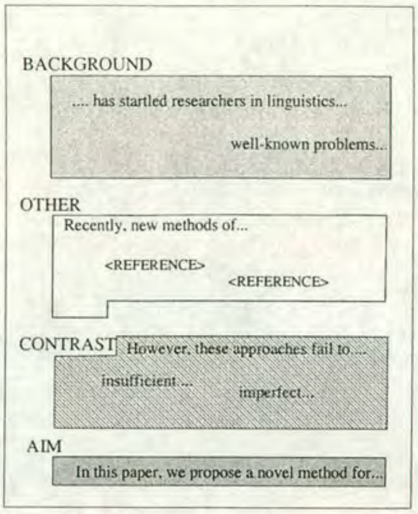

BACKGROUND
.... has startled researchers in linguistics...
well-known problems...
OTHER
Recently, new methods of...
<REFERENCE>
<REFERENCE>
CONTRAST
However, these approaches fail to....
insufficient ...
imperfect...
AIM
In this paper, we propose a novel method for...

Figure 2: Typical rhetorical pattern in a research paper introduction

expressed well with our category labels. Whereas non-basic categories are typically short segments of one or two sentences, the basic categories form much larger segments of sentences with the same rhetorical role.

3 Human Annotation

3.1 Annotating full texts

To ensure that our coding scheme leads to less biased annotation than some of the other resources available for building summarisation systems, and to ensure that other researchers besides ourselves can use it to replicate our results on different types of texts, we wanted to examine two properties of our scheme: stability and reproducibility (Krippendorff, 1980). Stability is the extent to which an annotator will produce the same classifications at different times. Reproducibility is the extent to which different annotators will produce the same classification. We use the Kappa coefficient (Siegel and Castellan, 1988) to measure stability and reproducibility. The rationale for using Kappa is explained in (Carletta, 1996).

The studies used to evaluate stability and reproducibility we describe in more detail in (Teufel et al., To Appear). In brief, 48 papers were annotated by three extensively trained annotators. The training period was four weeks consisting of 5 hours of annotation per week. There were written instructions (guidelines) of 17 pages. Skim-reading and annotation of an average length (3800 word) paper typically took 20-30 minutes. The studies show that the training material is reliable. In particular, the basic annotation scheme is stable (K=.82, .81, .76; N=1220; k=2 for all three annotators) and reproducible (K=.71, N=4261, k=3), where k denotes the number of annotators, N the number of sentences annotated, and K gives the Kappa value. The full annotation scheme is stable (K=.83, .79, .81; N=1248; k=2 for all three annotators) and reproducible (K=.78, N=4031, k=3). Overall, reproducibility and stability for trained annotators does not quite reach the levels found for, for instance, the best dialogue act coding schemes, which typically reach Kappa values of around K=.80 (Carletta et al., 1997; Jurafsky et al., 1997). Our annotation requires more subjective judgements and is possibly more cognitively complex. Our reproducibility and stability results are in the range which Krippendorff (1980) describes as giving marginally significant results for reasonable size data sets when correlating two coded variables which would show a clear correlation if there were perfect agreement. As our requirements are less stringent than Krippendorff's, we find the level of agreement which we achieved acceptable.

Category	Percentage
OWN	69.4%
OTHER	15.8%
BACKGROUND	5.7%
CONTRAST	4.4%
AIM	2.4%
BASIS	1.4%
TEXTUAL	0.9%

Figure 3: Distribution of categories

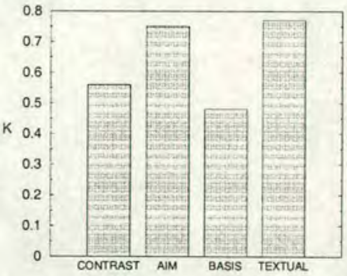


Figure 4: Reproducibility diagnostics: non-basic categories

Figure 3, which gives the overall distribution of categories, shows that OWN is by far the most frequent category. Figure 4 reports how well the four non-basic categories could be distinguished from all other categories, measured by Krippendorff's diagnostics for category distinctions (i.e. collapsing all *other* distinctions). When compared to the overall reproducibility of .71, we notice that the annotators were good at distinguishing AIM and TEXTUAL, and less good at determining BASIS and CONTRAST. This might have to do with the location of those types of sentences in the paper: AIM and TEXTUAL are usually found at the beginning or end of the introduction section, whereas CONTRAST, and even more so BASIS, are usually interspersed within longer stretches of OWN. As a result, these categories are more exposed to lapses of attention during annotation.

The fact that the annotators are good at determining AIM sentences is an important result: as AIM sentences constitute the best characterisation of the research paper for the summarisation task at a very high compression to 1.8% of the original text length, we are particularly interested in having them annotated consistently in our training material. This result is clearly in contrast to studies which conclude that humans are not very reliable at this kind of task (Rath et al., 1961). We attribute this difference to a difference in our instructions. Whereas the subjects in Rath et al.'s experiment were asked to look for the most *relevant* sentences, our annotators had to look for specific argumentative roles which seems to have eased the task. In addition, our guidelines give very specific instructions for ambiguous cases.

These reproducibility values are important because they can act as a good evaluation measure as it factors random agreement out, unlike percentage agreement. It also provides a realistic upper bound on performance: if the machine is treated as another coder, and if reproducibity does not decrease then the machine has reached the theoretically best result, considering the cognitive difficulty of the task.

3.2 Annotating parts of texts

Annotating texts with our scheme is time-consuming, so we wanted to determine if there was a more efficient way of obtaining hand-coded training material, namely by annotating only parts of the source texts. For example, the abstract, introductions and conclusions of source texts are often like "condensed" versions of the contents of the entire paper and might be good areas to restrict annotation to. Alternatively, it might be a good idea to restrict annotation to the first 20% or the last 10% of any given text. Yet another possibility for restricting the range of sentences to be annotated is based on the 'alignment' idea introduced in (Kupiec et al., 1995): a simple surface measure determines sentences in the document that are maximally similar to sentences in the abstract.

Obviously, any of these strategies of area restriction would give us fewer gold standard sentences per paper, so we would have to make sure that we still had enough candidate sentences for all seven categories. On the other hand, because these areas could well be the most clearly written and informationally rich sections, it might be the case that the quality of the resulting gold standard is higher. In this case we would expect the reliability of the coding in these areas to be higher in comparison to the reliability achieved overall, which in turn would result in higher accuracy when this task is done automatically.

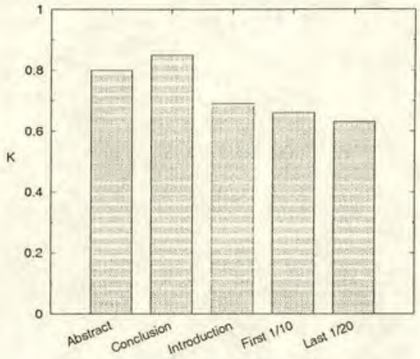


Figure 5: Reproducibility by annotated area

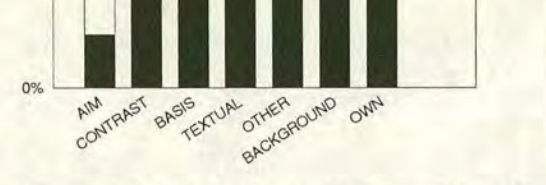


Figure 6: Label distribution by annotated area

We did extensive experiments on this. Figure 5 shows reliability values for each of the annotated portions of text, and Figure 6 shows the composition in terms of our labels for each of the annotated portions of text. The implications for corpus preparation for abstract generation experiments can be summarised as follows. If one wants to avoid manually annotating entire papers but still make all argumentative distinctions, one can restrict the annotation to sentences appearing in the introduction section, even though annotators will find them slightly harder to classify (K=.69), or to all alignable abstract sentences, even if there are not many alignable abstract sentences detectable overall (around 50% of the sentences in the abstract), or to conclusion sentences, even if the coverage of argumentative categories is very restricted in the conclusions (mostly AIM and OWN sentences).

We also examined a fall-back option of just annotating the first 10% or last 5% of a paper (as not all papers in our collection have an explicitly marked introduction and conclusion section), but the reliability results of this were far less good (K=.66 and K=.63, respectively).

4 Automatic annotation

All the annotation work is obviously in aid of development work, in particular for the training of a system. We will provide a brief description of training results so as to show the practical viability of the proposed corpus preparation method.

4.1 Data

Our training material is a collection of 80 conference papers and their summaries, taken from the Computation and Language E-Print Archive (`http://xxx.lanl.gov/cmp-lg/`). The training material contains 330,000 word tokens.

The data is automatically preprocessed into xml format, and the following structural information is marked up: title, summary, headings, paragraph structure and sentences, citations in running text, and reference list at the end of the paper. If one of the paper's authors also appears on the author list of a cited paper, then that citation is marked as self citation. Tables, equations, figures, captions, cross references are removed and replaced by place holders. Sentence boundaries are automatically detected, and the text is POS-tagged according to the UPenn tagset.

Annotation of rhetorical roles for all 80 papers (around 12,000 sentences) was provided by one of our human judges during the annotation study mentioned above.

4.2 The method

(Kupiec et al., 1995) use supervised learning to automatically adjust feature weights. Each document sentence receives scores for each of the features, resulting in an estimate for the sentence's probability to also occur in the summary. This probability is calculated for each feature value as a combination of the probability of the feature-value pair occurring in a sentence which is in the summary (successful case) and the probability that the feature-value pair occurs unconditionally.

We extend Kupiec et al.'s estimation of the probability that a sentence is contained in the abstract, to the probability that it has rhetorical role R (cf. Figure 7).

$$P(s \in R|F_1, \ldots, F_k) \approx \frac{P(s \in R) \prod_{j=1}^{k} P(F_j|s \in R)}{\prod_{j=1}^{k} P(F_j)}$$

where

$P(s \in R\|F_1, \ldots, F_k)$:	Probability that sentence s in the source text has rhetorical role R, given its feature values;
$P(s \in R)$:	relative frequency of role R (constant);
$P(F_j\| s \in R)$:	probability of feature-value pair occurring in a sentence which is in rhetorical class R;
$P(F_j)$:	probability that the feature-value pair occurs unconditionally;
k:	number of feature-value pairs;
F_j:	j-th feature-value pair.

Figure 7: Naive Bayesian classifier

Evaluation of the method relies on cross-validation: the model is trained on a training set of documents, leaving one document out at a time (the test document). The model is then used to assign each sentence a probability for each category R, and the category with the highest probability is chosen as answer for the sentence.

4.3 Features

The features we use in training (see Figure 8) are different from Kupiec et al.'s because we do not estimate overall importance in one step, but instead guess argumentative status first and determine importance later.

Many of our features can be read off directly from the way the corpus is encoded: our preprocessors determine sentence-boundaries and parse the reference list at the end. This gives us a good handle on structural and locational features, as well as on features related to citations.

Type of feature	Name	Feature description	Feature values
Explicit structure	Struct-1	Type of Headline of current section	8 prototypical headlines or 'non-prototypical'
	Struct-2	Relative position of sentence within paragraph	initial, medial, final
	Struct-3	Relative position of sentence within section	first, second or last third
Relative location	Loc	Paper is segmented into 10 equally-sized segments	1-10
Citations	Cit-1	Does the sentence contain a citation or the name of an author contained in the reference list?	Full Citation, Author Name or None
	Cit-2	Does the sentence contain a *self* citation?	Yes or No
Syntactic features	Syn-1	Tense (associated with first finite verb in sentence)	Present, Past, Present Perfect, Past Perfect, Future or Nothing
	Syn-2	Modal Auxiliaries	Present or Not
	Syn-3	Voice	Active or Passive
	Syn-4	Negation	Present or Not
Semantic features	Sem-1	Action type of first verb in sentence	20 different Action Types (cf. Figure 9) or Nothing
	Sem-2	Type of Agent	Authors or Others or Nothing
	Sem-3	Type of formulaic expression occurring in sentence	18 different types of Formulaic Expressions (cf. Figure 9) or Nothing
Content Features	Cont-1	Does the sentence contain keywords as determined by the tf/idf measure?	Yes or No
	Cont-2	Does the sentence contain words also occurring in the title or headlines?	Yes or No

Figure 8: Features for supervised learning

The syntactic features rely on determining the first finite verb in the sentence, which is done symbolically using POS-information. Heuristics are used to determine the tense and possible negation.

The semantic features rely on template matching. In the feature Sem-1, a hand-crafted lexicon is used to classify the verb into one of 20 Action Classes (cf. Figure 9, left half), if it is one of the 388 verbs contained in the lexicon. The feature Sem-2 encodes whether the agent of the action is most likely to refer to the authors, or to other agents, e.g. other researchers (177 templates). Heuristic rules determine that the agent is the subject in an active sentence, or the head of the by-phrase (if present) in a passive sentence. Sem-3 encodes various other formulaic expressions (indicator phrases (Paice, 1981), meta-comments (Zukerman, 1991)) in order to exploit explicit rhetoric phrases the authors might have used, cf. Figure 9, right half (414 templates).

The content features use the tf/idf method and title and header information for finding contentful words or phrases. In contrast to all other features they do not attempt to model the form or meta-discourse contained in the sentences but instead model their domain (object-level) contents.

4.4 Results

When the Naive Bayesian Model is added to the pool of coders, the reproducibility drops from K=.71 to K=.55. This reproducibility value is equivalent to the value achieved by 6 human annotators with no prior training, as found in an earlier experiment (Teufel et al., To Appear). Compared to one of the annotators, Kappa is K=.37, which corresponds to percentage accuracy of 71.2%. This number cannot be directly compared to experiments like Kupiec et al.'s because in their experiment a compression of around 3% was achieved whereas we classify each sentence into one of the categories.

Further analysis of our results shows the system performs well on the frequent category Own, cf. the confusion matrix in Fig. reftab:confusion. Indeed, as Figure 3 shows, Own is so frequent that choosing Own all the time gives us a seemingly hard-to-beat baseline with a high percentage agreement of 69% (Baseline 1). However, the Kappa statistic, which controls for expected random agreement, reveals just how bad that baseline really is: Kappa is K=-.12 (machine vs. one annotator). Random choice of categories according to the distribution of categories (Baseline 2) is a better baseline; Kappa

Action Types		Formulaic Expression Types	
AFFECT	*we hope to improve these results*	GENERAL_AGENT	*linguists*
ARGUMENTATION	*we argue against an application of*	SPECIFIC_AGENT	*according to <REF>*
AWARENESS	*we know of no other attempts...*	GAP_INTRODUCTION	*to our knowledge*
BETTER_SOLUTION	*our system outperforms that of ...*	AIM	*main contribution of this*
CHANGE	*we extend <CITE/>'s algorithm*	TEXTSTRUCTURE	*in section <CREF/>*
COMPARISON	*we tested our system against...*	DEIXIS	*in this paper*
CONTINUATION	*we follow X in postulating that*	CONTINUATION	*following the argument in*
CONTRAST	*our approach differs from X's ...*	SIMILARITY	*bears similarity to*
FUTURE_INTEREST	*we intend to improve our results...*	COMPARISON	*when compared to our*
INTEREST	*we are concerned with ...*	CONTRAST	*however*
NEED	*this approach, however, lacks...*	METHOD	*a novel method for XX-ing*
PRESENTATION	*we present here a method for...*	PREVIOUS_CONTEXT	*elsewhere, we have*
PROBLEM	*this raises the problem of how to...*	FUTURE	*avenue for improvement*
RESEARCH	*we collected our data from...*	AFFECT	*hopefully*
SIMILAR	*our approach resembles that of X...*	PROBLEM	*drawback*
SOLUTION	*we solve this problem by...*	SOLUTION	*insight*
TEXTSTRUCTURE	*the paper is organized as follows...*	POSITIVE_ADJECTIVE	*appealing*
USE	*we employ X's method...*	NEGATIVE_ADJECTIVE	*unsatisfactory*
COPULA	*our goal is to...*		
POSSESSION	*our approach has three advantages...*		

Figure 9: Types of actions and formulaic expressions

		MACHINE							
		AIM	CONTRAST	TEXTUAL	OWN	BACKGROUND	BASIS	OTHER	Total
HUMAN	AIM	115	4	10	46	15	13	4	207
	CONTRAST	11	79	5	280	92	40	89	596
	TEXTUAL	13	4	115	71	5	3	12	223
	OWN	75	61	61	7666	168	125	279	8435
	BACKGROUND	11	20	3	286	295	21	84	720
	BASIS	10	10	5	40	4	102	55	226
	OTHER	7	35	10	1120	203	173	466	2014
	Total	242	213	209	9509	782	477	989	12421

Figure 10: Confusion matrix: human vs. automatic annotation

for this baseline is K=0.

AIM categories can be determined with a precision of 48% and a recall of 56% (cf. Figure 11). These values are more directly comparable to Kupiec et al.'s results of 44% co-selection of extracted sentences with alignable summary sentences. We assume that most of the sentences extracted by their method would have fallen into the AIM category. The other easily determinable category for the automatic method is TEXTUAL (p=55%; r=52%), whereas the results for the other non-basic categories are relatively lower – mirroring the results for humans.

As far as the individual features are concerned, we found the strongest heuristics to be location, type of header, citations, and the semantic classes (indicator phrases, agents and actions); syntactic and content-based heuristics are the weakest. The first column in Figure 12 gives the predictiveness of the feature on its own, in terms of kappa between machine and one annotator. Some of the weaker features are not predictive enough on their own to break the dominance of the prior; in that case, they behave just like Baseline 1 (K=–.12).

Category	Precision	Recall
AIM	48%	56%
CONTRAST	37%	13%
TEXTUAL	55%	52%
OWN	81%	91%
BACKGROUND	38%	41%
BASIS	21%	45%
OTHER	47%	23%

Figure 11: Precision and recall per category

The second column gives kappa for experiments using all features except the given feature, i.e. the results if this feature is left out of the pool of fea-

Feature Code	Alone	Left out
Struct–1	–.12	.37
Struct–2	–.12	.36
Struct–3	.16	.36
Struct–1–3	.18	.34
Loc	.17	.34
Cit–1	.18	.37
Cit–2	.13	.37
Cit–1–2	.18	.36
Syn–1	–.12	.37
Syn–2	–.12	.37
Syn–3	–.12	.37
Syn–4	–.12	.37
Syn–1–4	–.12	.37
Sem–1	–.12	.36
Sem–2	.07	.35
Sem–3	–.03	.36
Sem–1–3	.13	.31
Cont–1	–.12	.37
Cont–2	–.12	.37
Cont–1–2	–.12	.37
Baseline 1 (all OWN): K=–.12 Baseline 2 (random by distr.): K=0		

Figure 12: Disambiguation potential of individual heuristics

tures. These numbers show that some of the weaker features contribute some predictive power in combination with others.

While not entirely satisfactory, these results might be taken as an indication that we have indeed managed to identify the right kinds of features for argumentative sentence classification. Taking the context into account should further increase results, as preliminary experiments with n-gram modelling have shown. In these experiments, we replaced the prior $P(s \in R)$ in Figure 7 with a n-gram based probability of that role occurring in the given context.

5 Conclusions

In this paper we have presented an annotation scheme for corpus based summarisation. In tests, we have found this annotation scheme to be stable and reproducible. On the basis of this scheme, we have created a new kind of resource for training summarisation systems: a corpus annotated with labels which indicate the argumentative role of each sentence in the text. Results of our training work show that the annotation work can be automated.

References

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

H. P. Edmundson. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285.

E. Garfield. 1979. *Citation indexing: its theory and application in science, thechnology and humanities*. Wiley, New York.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca, 1997. *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual*. University of Colorado, Institute of Cognitive Science. TR-97-02.

Klaus Krippendorff. 1980. *Content analysis: an introduction to its methodology*. Sage Commtext series; 5. Sage, Beverly Hills London.

Julian Kupiec, Jan O. Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference*, pages 68–73.

Elizabeth DuRoss Liddy. 1991. The discourse-level structure of empirical abstracts: an exploratory study. *Information Processing and Management*, 27(1):55–81.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Inderjeet Mani and Eric Bloedorn. 1998. Machine learning of generic and user-focused summarization. In *Proceedings of the Fifteenth National Conference on AI (AAAI-98)*, pages 821–826.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: description and construction of text structures. In G. Kempen, editor, *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, pages 85–95, Dordrecht. Nijhoff.

Daniel Marcu. 1997. From discourse structures to text summaries. In *Proceedings of the ACL/EACL workshop on Intelligent Scalable Text Summarization*.

Greg Myers. 1992. In this paper we report... – speech acts and scientific facts. *Journal of Pragmatics*, 17(4):295–313.

Kenji Ono, Kazuo Sumita, and Seijii Miike. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th International conference on Computational Linguistics (COLING-94)*.

Chris D. Paice. 1981. The automatic generation of literary abstracts: an approach based on the identification of self-indicating phrases. In Robert Norman Oddy, S. E. Robertson, C. J. van Rijsbergen, and P. W. Williams, editors, *Information Retrieval Research*, pages 172–191. Butterworth, London.

G.J Rath, A. Resnick, and T. R. Savage. 1961. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2):139–143.

Sidney Siegel and N.J. Jr. Castellan. 1988. *Nonparametric statistics for the Behavioral Sciences.* McGraw-Hill, second edition.

Karen Spärck Jones. 1998. Automatic summarising: factors and directions. In *AAAI Spring Symposium on Intelligent Text Summarization.*

John Swales. 1990. *Genre analysis: English in academic and research settings.* Cambridge University Press.

Simone Teufel, Jean Carletta, and Marc Moens. To Appear. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-99).*

Ingrid Zukerman. 1991. Using meta-comments to generate fluent text in a technical domain. *Computational Intelligence: Special Issue on Natural Language Generation*, 7(4):276.

Index of Citations