Discourse Oriented Summarization

WAUTER BOSMA

Samenstelling van de promotiecommissie:

Prof. dr. E. André, Universität Augsburg
Prof. dr. E. H. Hovy, USC Information Sciences Institute
Prof. dr. F. M. G. de Jong, Universiteit Twente
Prof. dr. E. J. Krahmer, Universiteit van Tilburg
Prof. dr. ir. A. J. Mouthaan, Universiteit Twente (voorzitter)
Prof. dr. ir. A. Nijholt, Universiteit Twente (promotor)
Prof. dr. M. de Rijke, Universiteit van Amsterdam
Prof. dr. M. F. Steehouder, Universiteit Twente
Dr. M. Theune, Universiteit Twente (co-promotor)


CTIT Dissertation Series No. 08-112
Center for Telematics and Information Technology (CTIT)
P.O. Box 217 – 7500AE Enschede – the Netherlands
ISSN: 1381-3617


NWO IMIX/IMOGEN
The research reported in this thesis has been carried out in the IMOGEN (Interactive Multimodal Output Generation) project. IMOGEN is a project within the Netherlands Organisation for Scientific Research (NWO) research program on Interactive Multimodal Information eXtraction (IMIX).


SIKS Dissertation Series No. 2008-10
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Resources used in this thesis were made available free of charge by Spectrum, Merck, and others. I also wish to express my gratitude to the people at HMI, IMIX, and elsewhere (most notably Mariët Theune), for their support and their contributions to this thesis.

DISCOURSE ORIENTED SUMMARIZATION

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. W. H. M. Zijm,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op donderdag 27 maart om 16.45 uur

door

Wauter Eduard Bosma

geboren op 2 juni 1979

te Zuidlaren

Promotor: prof. dr. A. Nijholt

Co-promotor: dr. M. Theune

ISBN 978-90-365-2649-4

Contents

1

Introduction

Nothing is so valuable as the right information at the right time. A diversifying range of applications of natural language processing is dedicated to delivering that. A wide spread application is the traditional web search (using information retrieval), but other methods are gaining ground. An example is the question answering feature in modern search engines. Today[1], a query such as *what is the population of Brussels* gives a direct answer in addition to a list of documents.

Answering specific types of trivia style (so-called 'factoid') questions such as the one on Brussels' population is the focus of *question answering* research. In question answering, questions are typically categorized by their answer types – e.g. a date, a name, a number, etc. Questions that fall outside these categories are not addressed in major evaluation programs (c.f. Voorhees, 2003). Furthermore, it is not trivial that a questions has only one possible answer type. For instance, a *who is* question may be used to ask for a name or to request a biography. Strzalkowski et al. (2000) showed that even if there is an unambiguous query, users appreciate more information than a direct answer. Someone querying a system for the population of Brussels may also be interested in aspects other than its size, such as ethnicity, cultural characteristics, etc. Bates (1990) helps explaining the findings of Strzalkowski et al. by viewing an information search as a 'berry picking' process. Consulting an information system is only part of a user's attempt to fulfill an information need. It's not the end point, but just one step whose result may motivate a follow-up step. The user may not only be interested in the answer to the question, but also in related information. The 'factoid answer approach' fails to show leads to related information that might also be of interest. Bakshi et al. (2003b) show that when answering questions, increasing the amount of text returned to users significantly reduces the number of queries that they pose to the system, suggesting that users utilize related information from surrounding text.

[1] Google, Yahoo, MSN, dd. November 23, 2007.

Query-based summarization is a way to return more information than just a direct answer to a question. Throughout this thesis, query-based summarization refers to presenting an answer in response to a user-specified query by means of a paragraph-sized text and (possibly) images. The answer's content is drawn from a set of documents (the *source*) providing an answer but not necessarily written to answer the query. A generic summarization system intends to distill the author's main points from a document. The objective of query-based summarization is not to find what is presented as important in the source, but what is of interest to the user. The user expresses his/her interest in the form of a query. The term *query* is more general than a *question*: a query is a request for information. A query is a question if it is expressed as an interrogative sentence.

A collection of query-based text summaries created by professional abstractors is produced in the context of the yearly DUC summarization evaluation event (Dang, 2006). An informal review of query-based summaries created for the 2006 edition of DUC reveals that human summarizers present answers in context. This context may provide general background knowledge or other information to make the actual answer more understandable or to make the reader more receptive to the answer. For instance, in response to the question which measures have been taken to improve automobile safety, three of the human summaries mentioned laws enforcing seat belt use. Two out of these three summaries first mentioned the reasons why these steps are deemed necessary. The fact that human summarizers include answer-related information is in line with the study of Bakshi et al. (2003b) on answer presentation mentioned earlier.

A deep analysis of both the query and the source would be required to 'understand' the interests of the user and respond adequately. A deep analysis of unrestricted text is not feasible with current technology. As an alternative, cues for recognizing the structure of the source may be derived from surface features of the text (c.f. Morris and Hirst, 1991; Marcu and Echihabi, 2002). Given a sentence which responds to the query (a 'direct answer'), text structure may help directing a summarization system toward related content. This related content may be of interest to the user as well, and at the same time, is likely to cohere with the answer.

The focus of this thesis is on using discourse structure for query-based summarization as an attempt to find more information than just a direct answer. I developed and evaluated models for discourse oriented summarization of text documents and multimedia documents which contain text and pictures. Summarization methods are evaluated by means of automatic algorithms and two user studies. Automatic methods are useful for determining how well a summary resembles an 'ideal' reference summary. User studies are useful to determine how well the summaries respond to information

Table 1.1: Applications of natural language processing.

application	information need	response unit	purpose
generic summarization	derived from content	paragraph	inform or indicate
question answering	user-specified	phrase/list	inform
query-based summarization	user-specified	paragraph	inform or indicate
information retrieval	user-specified	document list	indicate

needs, addressing both text and multimedia summarization. For testing the significance of differences between the measured quality of summaries, I propose a novel, non-standard method which is more likely to detect significant differences than existing methods. Apart from algorithms for summarization, attention is paid to establishing relations between content elements. Most notably, I address detecting entailment between two pieces of text. I also present new methods for measuring performance of entailment systems, which has quantifiable advantages over existing methods.

1.1 Summarization

Discourse structure has been proposed as a means for generic summarization by Marcu (1997a). This thesis focuses on discourse oriented methods for *query-based* summarization. Query-based summarization is related to generic summarization and other categories of natural language processing applications, listed in Table 1.1. Each of the applications serves to satisfy a user's need for information. The mentioned categories are distinct in three ways. First, who specifies the information need? Generic summarization aims to produce a concise version of a document (or a number of documents). The summary is not tailored to the user or any expressed information need. The other applications listed use some form of query or question to specify a need for information. Second, the type of answer ranges from a precise and short answer (question answering) to a list of documents which may contain the information needed (information retrieval). In between these two extremes is summarization, typically returning a paragraph-sized answer. Third, the intended result: is the system supposed to provide information as such (e.g. an answer or the tenor of a document), or to direct the user toward relevant information? Summaries can be written to *indicate* what the source document is about in order to help the user assessing the relevance of that particular document (indicative summaries), or summaries can be written to *inform* the user of its content (informative summaries). Summaries of multiple documents are typically informative: because multi-document summaries may contain information from

a number of sources, they are little suitable for indicating the relevance of a particular source. In this thesis, summarization refers to informative summarization unless stated otherwise.

Practical applications are often a combination of the applications of Table 1.1. For instance, information retrieval systems typically present more than a list of pointers to documents. In addition, they produce a brief summary of each document, to help the user determine its relevance. Also a combination of query-based summarization and question answering is conceivable, e.g. for presenting answers in context – question answering techniques are used to find an answer, and that answer is sent as a query to a summarization system in order to provide some background in addition to a precise answer.

1.2 IMIX

The work described in this thesis is done within the context of IMIX, a program for research on *Interactive Mutimodal Information Extraction* sponsored by the Netherlands Organization for Scientific Research (NWO). In addition to promote research in its focus area, one of the goals of IMIX was to produce a system for demonstration purposes which integrates and applies results of research, including the work presented in this thesis. The IMIX system is an application of this work embedded in a greater whole (c.f. Theune et al., 2007; Boves and den Os, 2005).

The IMIX system answers questions for medical information from a general audience of non-expert adult users. The purpose of the system is to answer 'encyclopaedic' questions to which answers can be typically found in an encyclopedia. Questions can be typed or spoken (in Dutch), and answers are presented in the form of speech, text and pictures. Questions can be asked in isolation, but the system is also capable of engaging in dialogs and answer follow-up questions.

Other projects of IMIX were responsible for question answering (van den Bosch et al., 2004; Tjong Kim Sang et al., 2005; Bouma et al., 2007), dialog and action management (op den Akker et al., 2005), speech synthesis (Marsi, 2004), and speech recognition (Hämäläinen et al., 2007). Work in this thesis contributed to the *answer presentation* module of IMIX. In the IMIX system, questions are pre-processed by the dialog manager and forwarded to *question answering*, which is responsible for searching for answers in a corpus of encyclopedia and web documents. The answer presentation module presumes a ranked list of pointers to sentences containing poten-

tial answers. These pointers are used for discourse oriented summarization of the text containing the answers, using summarization algorithms presented in this thesis. If the answer presentation module also illustrates answers with a picture, if sufficiently confident that an appropriate picture is available. The lingual component of the answer is presented in speech and text.

1.3 Research questions

This thesis aims to answer the question, *how can query-based summarization systems exploit discourse structure to produce better summaries?*

This question can be divided in several more specific research questions whose answers contribute to the main question above. My starting point is to use coherence analyses for query-based summarization. A specific coherence model, Rhetorical Structure Theory, has previously been used for summarization (Marcu, 1997a), but not for query-based summarization, and not in an extendible way. The ideal summarization system is extendible in the sense that it is capable of using coherence along with other aspects of discourse structure. The derived research question is: (1) *how can manual analyses of coherence be used in a query-based summarization system?*

Creating coherence analyses is laborious, but automatic statistical features of text may provide a less accurate but scalable alternative. The next question addresses this issue: (2) *how can automatic features replace manual coherence analyses in query-based summarization?*

Coherence explains internal text structure, but not how passages from different documents relate to each other. Nevertheless, a summarization system should be aware of the difference between *entailment* and the more general notion of *relatedness*, e.g. to avoid including redundant content in a summary. Therefore, a summarization system would benefit from the answer to the research question: (3) *how can entailment between arbitrary text passages be automatically detected?*

The previous subquestions address text summarization, but the added value of media items should not be neglected. Hence the fourth and last subquestion: (4) *how can discourse oriented text summarization techniques be generalized to multimedia summarization?*

1.4 Thesis outline

A general introduction to the theory of **discourse structure** in chapter 2 provides background information required to interpret the rest of the thesis. Discourse structure can be analyzed on several levels relevant for summarization. In text analysis, a distinction is made between structural relations between textual elements – such as a reference by the word *that* to a concept in a preceding sentence – and relations between ideas conveyed by the text, such as one part of a text providing a background for interpreting another part of the text. Similar issues play a role in multimedia documents. For instance, a media item may provide context to understand the text, or a textual element may be used to refer to (part of) a media item by means of a symbol (e.g., a reference such as *Table 1.1*) or a lingual description of part of the item (e.g., *the left figure*).

A third level of discourse analysis is that of relating text from different documents. While there are established methods for measuring similarity between text passages, these measures do not distinguish *relatedness* from *redundancy*. Being able to do so would particularly be a virtue if multiple documents are used as a source for summarization. Chapter 3 zooms in on **textual entailment** – a type of inter-document relations which implies redundancy. If one sentence is known to entail another, a summarization system can respond appropriately, e.g. by not including both sentences in one summary as to avoid redundancy. This chapter aims to answer question 3. I propose to decompose the task of recognizing entailment into *representation* and *matching*. Based on this decomposition, a systematic comparison is made of incrementally more complex methods of representation and matching. This chapter describes novel methods for recognizing entailment: a method based on syntactic patterns (described earlier in Marsi et al., 2006) and a method which employs paraphrase substitution (described in Bosma and Callison-Burch, 2007).

A literature review on **automatic summarization** is given in chapter 4. This chapter discusses summarization by humans, issues in evaluating summaries, and methods for the automatic generation of summaries and query-based summaries in particular.

When responding to a query, there are several reasons why returning more content may be preferable, even if a direct answer is readily available. As mentioned previously, information *related* to the answer may also be of interest to the user. Furthermore, since computer output cannot be expected to be free of errors, secondary information in the response may be used by the user as an implicit verification that the query was correctly interpreted. Chapter 5 answers question 1 by presenting a **dis-**

course oriented summarization method as well as the results of a user study, in which discourse oriented summarization and layout-based summarization are compared with respect to the relevance of presented content and the verifiability thereof (based on Bosma, 2005c). The discourse oriented summarization method is based on the method of Marcu (1999), adapted for query-based summarization (presented in Bosma, 2005a).

Chapter 6 is dedicated to answering question 2 by evaluating existing and novel **algorithms and features for query-based text summarization**. While the user experiments of the previous chapter used annotated text, the summarization system used for these experiments is fully automatic. First, I present a modular framework for discourse oriented summarization, dividing the summarization process in four phases which can be implemented independently. This framework is compatible with the summarization methods used in chapter 5. An extensive comparison of implementations of this framework is made using Rouge, varying the type of information used for content selection. Rouge is a package for automatic summarization evaluation (Lin, 2004). One implementation of the summarization system used underwent a full evaluation within the context of DUC 2006 (described earlier in Bosma, 2006). For all experiments in this chapter, the data of DUC 2006 (i.e. queries and reference summaries) were used for evaluation (Dang, 2006).

A specific instance of the summarization framework in chapter 6 is a system which automatically **illustrates answers** to medical questions. Such a system is presented in chapter 7 (research question 4). Given a textual answer to a medical question and a corpus of annotated pictures, a presentation is generated which contains the text and a picture. This is a specific case of query-based summarization: given an information need and a set of potential source documents, a concise presentation is generated answering that information need. The candidate pictures and their annotation are automatically extracted from medical literature. Two picture selection algorithms based on Bosma (2005b) were evaluated by means of a user study following the experimental design of van Hooijdonk et al. (2007a).

Chapter 8 reviews issues addressed in this thesis, summarizes the findings presented in this thesis. Chapter 8 highlights the main contributions of this thesis and gives pointers to promising directions of research.

2

Modelling discourse structure

In this chapter, I review literature on three levels of discourse structure in text and multimedia, and their potential use in summarization. The three levels of interest are cohesion (relations between textual or media elements), coherence (relations between ideas expressed in the text or multimedia realization), and cross-document relations. For various types of relations, attempts have been made to detect them automatically. Automatic means of detecting such relations can be exploited in automatic summarization.

If you visit an online store to buy a book, the book store suggests other books which may be of interest to you. If an information system is asked a question, why not provide more information than explicitly asked for? Humans tend to do this by default. When I asked a receptionist where to complain about a vending machine which takes money but does not give anything in return, he answered: "Report this to the canteen, but it is closed now." This is obviously more information than asked for.

Providing information not explicitly asked for may be rewarding because the answering side may have more knowledge about which information is needed than the person asking (it saved me a walk to the canteen; the book store visitor may find a valuable book s/he would not find otherwise). Providing this information is also a challenge. A book store may use meta-information such as sales statistics, names of authors, etc. When relating documents or parts of documents, meta-information may be unavailable or insufficient.

Relating text passages (or media items in general) in a meaningful way involves 'understanding' the text, or at least to understand it to the level necessary for detecting relations between passages. The whole of relations between passages that consitutes the structure of a text, I call *discourse structure*. A *passage*, in this thesis, refers to a contiguous part of a document; it may be a paragraph, a sentence or a clause, but also a picture if the document is a multimedia document.

The question is, what is discourse structure and how is it manifested in language? Within a sentence, structural constraints are imposed by grammar. However, grammaticality is not sufficient to constitute meaning. The interpretation of *he* in sentence 1A below probably relies on the meaning of another textual element, presuming the sentence is part of a larger whole.

1A It was he who rewrote history.

The reference established by *he* in 1A is an instance of a *cohesive tie* (Halliday and Hasan, 1976). Although cohesive ties may be bound by syntax (e.g. agreement in number, gender), they are not part of the grammatical structure of a sentence and they may cross sentence boundaries. Language provides a number of ways to refer to linguistic elements independent of the grammatical structure. Together, these references consitute *cohesion* in text. However, as the following passage shows, there is more to discourse than cohesion.

2A I'll have to cancel dinner tonight.

2B I lost my car keys.

This passage contains two statements and an implicit relation between them. Sentence 2B can be interpreted as providing a background or a justification of what is said in 2A. Nonetheless, no grammatical relations between the sentences can be identified and cohesion does not fully explain the relation between the sentences; the mere juxtaposition of the sentences adds information which is not in either sentence as such. Apparently, something happens while interpreting this text which causes the reader to relate pieces of information in a way depending not only on the content itself, but also on the organization of the text. Text organization on this level of understanding – concerning relations between ideas – has been termed *coherence*. Relations such as cause, temporality and contrast contribute to the coherence of text.

What distinguishes coherent from incoherent text? Text is a medium to transfer a message from its writer to a receiver. Coherence is what enables a writer to send a message of more than one sentence, i.e. what makes the difference between a message

and sequence of messages (Hobbs, 1985; Mann and Thompson, 2000a). Theories of coherence explain relations between the 'ideas' that contribute to the author's message – the ideational structure of discourse. Cohesion pertains to the textual realization of the message.

Cohesion and coherence are aspects of discourse organization, but do not explain or describe relations between documents. A document rarely stands alone. A document may cite (e.g. scientific articles), interpret (e.g. parodies), contain partly the same information as another document (e.g. a news article on the same topic) or be related to another document in some other way. Documents are embedded in a larger context in which cross-document relations appear (Radev, 2000). An essential difference between coherence and cross-document relations is that coherence can be presumed for well-written documents: the structure of a document corresponds with the line of argumentation followed by the author. A collection of documents written by different authors does not necessarily have a consistent or coherent line of argumentation. Radev found types of relations between (parts of) documents which do not appear within a well written document. When summarizing news articles, the most critical cross-document relation is *paraphrasing*: two passages express the same information.

The remainder of this chapter reviews three levels of discourse analysis: cohesion (section 2.1), coherence (section 2.2) and cross-document relations (section 2.3).

2.1 Cohesion

Skorochod'ko (1981) related cohesion to coherence. He viewed coherence as a derivative of cohesion: a semantic relation between two sentences can be established if the number and strength of relations between their words exceeds a certain threshold. Skorochod'ko defined measures for 'relatedness' between sentences, based on coreferences and repetition of words.

Skorochod'ko (1981) quantified certain aspects of cohesion from a computational perspective. To measure 'relatedness' between words, Skorochod'ko assigned a *type*, a *direction* and a *strength* to semantic relations. The strength of a semantic relation is the inverse of the 'semantic distance'. Examples of relation types are SUBJECT/ACTION (e.g. calculator/calculate) and ACTION/RESULT (e.g. calculate/calculation).

While Skorochod'ko was interested in creating a computational model of text structure, Halliday and Hasan (1976) described cohesion and its realization in text from a linguistic perspective. Halliday and Hasan introduced the term *cohesive tie* to refer to

3A Both [the shaggy man]$^{\diamondsuit}$ [and]$^{\clubsuit}$ Dorothy looked grave [and]$^{\clubsuit}$ anxious, [for]$^{\clubsuit}$ [they]$^{\diamondsuit}$ were sorrowful that [such a misfortune]$^{\diamondsuit}$ had overtaken [[their]$^{\diamondsuit}$ little companion]$^{\diamondsuit}$.

3B Toto barked at [the fox-boy]$^{\diamondsuit}$ once or twice, not realizing [it]$^{\diamondsuit}$ was [[his]$^{\diamondsuit}$ former friend]$^{\diamondsuit}$ [who]$^{\diamondsuit}$ now wore [the animal [head]$^{\heartsuit}$]$^{\diamondsuit}$; [but]$^{\clubsuit}$ Dorothy cuffed [the dog]$^{\diamondsuit}$ [and]$^{\clubsuit}$ made [him]$^{\diamondsuit}$ stop $\emptyset^{\text{barking}}$.

3C As for [the foxes]$^{\diamondsuit}$, [they]$^{\diamondsuit}$ all seemed to think Button-Bright's new [head]$^{\heartsuit}$ very becoming [and]$^{\clubsuit}$ that [their]$^{\diamondsuit}$ King had conferred a great honor on [this little stranger]$^{\diamondsuit}$.

3D It was funny to see [the boy]$^{\diamondsuit}$ reach up to feel [his]$^{\diamondsuit}$ sharp [nose]$^{\heartsuit}$ [and]$^{\clubsuit}$ wide [mouth]$^{\heartsuit}$, [and]$^{\clubsuit}$ wail afresh with grief.

3E [He]$^{\diamondsuit}$ wagged [his]$^{\diamondsuit}$ [ears]$^{\heartsuit}$ in a comical manner [and]$^{\clubsuit}$ tears were in [his]$^{\diamondsuit}$ little black [eyes]$^{\heartsuit}$.

3F [But]$^{\clubsuit}$ Dorothy couldn't laugh at [[her]$^{\diamondsuit}$ friend]$^{\diamondsuit}$ just yet, [because]$^{\clubsuit}$ [she]$^{\diamondsuit}$ felt so sorry.

Figure 2.1: Text annotated with cohesive ties. Excerpt from *L. Frank Baum, The road to Oz*, p. 10. Annotated cohesive ties are: [reference]$^{\diamondsuit}$, [conjunction]$^{\clubsuit}$, $\emptyset^{\text{elipsis}}$, [lexical cohesion]$^{\heartsuit}$.

the dependence of the interpretation of one element by reference to another (Halliday and Hasan, 1976, p.11). Halliday and Hasan distinguish five forms of cohesion, called reference, substitution, ellipsis, conjunction and lexical cohesion. Each of these will be discussed later in more detail.

Cohesion has also been related to information structure (Grosz et al., 1995; Kruijff and Kruijff-Korbayová, 2001). Theorists of information structure aim to explain how the textual context evolves while the text progresses. This is essential for determining the salience of information units at a particular point in the text. The discussion here will be restricted to describing cohesive features of text, i.e. how textual elements are referenced from elsewhere in the text, without going into too much detail on the semantical processes behind it.

2.1.1 Reference

The class of cohesive ties called *reference* is subdivided into situational and textual coreferences to a specific item. The first are references to extra-textual entities; the latter elements within a text. The difference is a matter of interpretation rather than appearance. Examples of references are pronouns (*they, she*), demonstratives (*that, these*), and specific uses of definite noun phrases. Instances of *reference* in Figure 2.1 are marked like [this]$^{\diamondsuit}$.

Abundance of references makes it rewarding to automate their detection. Hobbs (1986) focuses on automatic resolution of the pronouns *he*, *she*, *it* and *they*. Hobbs designed an algorithm for finding their antecedents, based on their grammatical form. This algorithm searches for eligible antecedents in the syntactic parse tree of the sentence containing the pronoun, and preceding sentences if necessary. With this algorithm he achieved an accuracy as high as 88.3 percent. On the other hand, he also recognized that references are constrained not only by grammar, but also by semantic validity and the reader's expectations, as the following example illustrates:

4A If the baby does not thrive on raw milk, boil it.

Does *it* refer to the baby or to raw milk? Such ambiguities are difficult to resolve without extensive knowledge of the domain. Hobbs proposes to use logical inferencing for knowledge intensive coreference resolution, but the extensive knowledge required for this task prevented him from creating a system which is useful in practice.

Perhaps the most well-known algorithm for resolving pronouns is the knowledge poor algorithm developed by Lappin and Leass (1994). Of all potential antecedents, Lappin and Leass first rule out ties that would be ungrammatical. Among the remaining options, the algorithm uses a set of heuristics to choose the most likely antecedent. Lappin and Leass model the reader's attentional state (c.f. Grosz and Sidner, 1986) to decide which potential antecedent is most salient. Lappin and Leass (1994) claim their algorithm outperforms the algorithm of Hobbs by a few percent.

2.1.2 Substitution and ellipsis

Substitution allows referring by using a place holder, such as [one]♠ in:

5A I hate hospitals.
5B My grandfather went into [one]♠, and when he came out, he was dead.

The substitute *one* refers to the class of hospitals. *Substitution* is distinguished from *reference* because a referential tie presupposes a specific item, whereas *substitution* is used to refer back to a class of items (i.e., a hospital, rather than a specific one). Ellipsis (marked $\emptyset^{\text{antecedent}}$ in Figure 2.1) is the specific type of substitution where an empty place holder is used.

2.1.3 Conjunction

Conjunctions ([marked]♣ in Figure 2.1) are used to indicate that two pieces of information are related to each other. The relation is indicated by a conjunctive adjunct.

Conjunctive adjuncts may be adverbs (*but, so, nevertheless*) or prepositional expressions (e.g. *on the contrary*), sometimes using a reference (e.g. *because of that*). In computational linguistics, they are often referred to as *cue phrases* or *discourse markers*.

Conjunctions are specifically interesting as a cohesive device, because they are on the borderline between cohesion and coherence. Halliday and Hasan (1976) classified conjunctions into four categories: additive (e.g., *and*), adversative (e.g., *yet*), causal (e.g., *so*) and temporal (e.g., *then*). It is not a coincidence that the terms Halliday and Hasan use to describe these categories are similar to relation types in theories of coherence, such as Rhetorical Structure Theory (Mann and Thompson, 1988). Theune et al. (2006) used the same classification of conjunctions as Halliday and Hasan for realizing coherence relations in a natural language generation system. Knott and Dale (1995) derived a taxonomy of coherence relations from cue phrases they encountered in text. Marcu and Echihabi (2002) used cue phrases to bootstrap a machine learning approach to automatic recognition of coherence relations.

Cohesion (and thus conjunction) is part of the realization of discourse, while coherence refers to the ideational structure of discourse. An author may or may not make use of conjunction to indicate a coherence relation. For instance, the author could have chosen to omit the adjunct *but* in sentence 3F, if s/he deemed it unnecessary as an explicit marker of the argumentative structure.

2.1.4 Lexical cohesion

Some words refer back to a preceding word just by the particular choice of words. Unlike the other types of cohesion, lexical cohesion is not reflected in grammar. The idea behind lexical ties is that words may need to be interpreted in the light of the context shaped by preceding related words. There is no restriction to what kind of relation this might be, and there is no restriction to the classes of related words. Halliday and Hasan (1976) write: *Text provides context within which the item will be incarnated on this particular occasion. This environment determines the 'instantial meaning', or text meaning, of the item, a meaning which is unique to each specific instance.*

One word affects the interpretation of the other by their co-occurrence in text. Examples of lexical ties that might appear in text are ⟨ *garden, digging* ⟩ and ⟨ *construction site, digging* ⟩. The interpretation of *digging* in relation to a garden would be different from an interpretation of *digging* in the context of a construction site.

Interpretation of a word is often not affected by a single preceding word, but by a chain of words which share a 'lexical environment'. These chains are called *cohesive chains* or *lexical chains*. The [marked]$^{\heartsuit}$ words in Figure 2.1 can be viewed as part of the same lexical chain. The definition of *lexical tie* imposes no restriction to how words participating in a tie are related, or how long lexical chains can be. This leaves room for ambiguity. (If *tears* in sentence 3E belongs to the same chain as *eyes*, does it automatically belong to the chain that started with *head*?)

Morris and Hirst (1991) explored the possibility to recognize lexical chains automatically, and they designed an algorithm that uses a thesaurus to extract lexical chains from text. To do so, they came up with a more precise definition of what their algorithm regards a lexical chain. The algorithm scans the text from left to right; each word (except high frequency and closed class words) is considered for inclusion in an existing chain. If no chain applies, a new chain is created. In their algorithm, they introduced the concepts of linear distance and the level of transitivity. The word is added to a chain if it relates to the first word of the chain and the linear distance between the last word of the chain and the candidate word (the number of sentences in between) is not more than 3. The level of transitivity of a relation between two words is expressed in number of transitive links connecting the two words within a chain. For example, if word *a* is related to *b*, and *b* is related to *c*, then the level of transitivity of the relation between *a* and *b* is 0; the level of transitivity of the relation between *a* and *c* is 1 (given that *a* and *b* are members of the same chain). For a word to be added to a chain, it must be related to the first word in the chain with a transitive distance of at most 1.

Morris and Hirst (1991) were not able to extract lexical chains automatically because they did not have access to a suitable thesaurus in machine-readable form. To evaluate their algorithm, they used manually extracted lexical chains for conducting user experiments to show that algorithmically extracted lexical chains largely correspond with an intuitive notion of lexical cohesion. Later, Teich and Fankhauser (2004) designed a new algorithm for computing lexical chains which uses WordNet (Miller et al., 1990) as a resource for discovering lexical relations. They report that missing links in the thesaurus pose a considerable problem to the possibility of automatic lexical chain extraction.

Manabu and Hajime (2000) abandoned the idea of using a thesaurus for finding related words. Instead, they used cosine similarity to calculate the similarity of a word pair in a set of documents. Cosine similarity is widely used as a measure of similarity of two documents, but can also be used to measure similarity of *terms*. To do so, each term is represented as a vector of documents $[d_1..d_n]$, where d_i is the number of

occurrences of the term in document *i*. Two terms can be compared by measuring the similarity of their vector representations. This is typically done by measuring the cosine of the angle between the two vectors.

In a corpus of *m* documents, a term can be written as a vector of length *m*. Given terms *A* and *B* and their respective vector representations $[a_1..a_m]$ and $[b_1..b_m]$, the cosine similarity of those terms is their angle in *m*-dimensional space, calculated as follows (Salton, 1988):

$$\begin{aligned} cosim(A,B) &= \frac{A \cdot B}{\|A\| \cdot \|B\|} \\ &= \frac{\sum_{i=1}^{m} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{m} a_i^2} \cdot \sqrt{\sum_{i=1}^{m} b_i^2}} \end{aligned} \tag{2.1}$$

2.1.5 Cross-modal references

Research on cohesion in multimedia is not addressed by Skorochod'ko (1981) or Halliday and Hasan (1976), who focus on phenomena of cohesion in text. A significant amount of work in this respect has been done in input processing for interactive multimodal systems. The first multimodal system was the *put-that-there* system of Bolt (1980). It allowed the user to issue commands to the computer in order to manipulate a virtual world. The commands (such as *put that there*) could consist of simultaneous text and gestures. Later research in this area concentrated on integrating parallel input in multiple modes. Integration is converting all input into a single, system-internal representation, and detecting cross-modal cohesion (Vergo et al., 2000). The nature of cross-modal cohesion is as diverse as applications of multimedia. Examples are cooperative references to a physical item using text and gestures (e.g. *that*), and synchronization of speech and lip movements.

2.2 Coherence

2.2.1 Coherence relations

Coherence is what makes the difference between a message and sequence of messages (Hobbs, 1985; Mann and Thompson, 2000a). What this means in practice can be illustrated by an example:

6A By lacking an erosive atmosphere and geologically active outer layers,
6B the moon has preserved a record of early events in the history of the solar system.[1]

The text 6A–6B contains three assertions: 6A the moon lacks an erosive atmosphere and geologically active outer layers; 6B the moon has preserved a record of early events in the history of the solar system; and an implicit causal relation, i.e. that 6B is a consequence of 6A. The causal relation is conveyed by the juxtaposition of passages and the cohesive conjunction indicated by *by*, and is part of the *coherence* of the text. According to Mann and Thompson (2000a), the presence of a coherence relation between passages implies an additional message which is not conveyed by any of the participants of the relations.

7A Of course, I'd have paid you back.
7B Unfortunately, I lost my wallet.

In the text of 6A–6B, the relation is in this case indicated by the conjunctive adjunct *by*. This is not necessarily the case, as demonstrated by text 7A–7B. The sentences 7A–7B are only related implicitly, as they do not contain connectives and they do not refer to one another explicitly. According to Hobbs (1985), a reader or listener hypothesizes coherence (e.g. a causal relation) and uses prior knowledge and inference to test the validity of the hypothesis. In case of text 7A–7B, a reader may recognize coherence by hypothesizing a causal relation between a lost wallet and the lack of money. A plausible interpretation is that the writer's intention is to convince the listener that paying is not possible because the wallet is lost, supposedly to generate understanding.

2.2.1.1 Discourse units

The smallest unit of text to participate in a coherence relation has been termed *discourse constituent unit* (Polanyi, 1988) or *elementary discourse unit* (Mann and Thompson, 1988). In order to participate in a coherence relation, a text passage must convey

[1] Example from Mann and Thompson (2000a).

meaning. Therefore, elementary discourse units are generally considered the smallest unit to have meaning. Polanyi (1988) and Mann and Thompson (1988) propose to use clauses as elementary discourse units; in the annotated corpus of Carlson et al. (2003), even smaller units are used.

The discussion on information-carrying units appears in various areas of natural language processing, such as machine translation and automatic summarization. In summarization evaluation, Nenkova and Passonneau (2004) introduced the *semantic content unit*, which they defined as an 'atomic fact'. From sentence 8A below, Nenkova and Passonneau derive two semantic content units: (1) *Pinochet was arrested*, and (2) *the arrest took place in Britain*. Analysis of information structure addresses the relative salience of these facts by examining their context (Kruijff and Kruijff-Korbayová, 2001). For instance, if sentence 8A was preceded by the question "who was arrested," fact (1) is the more salient. By contrast, if the question "where was Pinochet arrested?" was asked, fact (2) is salient. Recognizing coherence relations may require text analysis at this level of granularity, but this is not addressed by theories of coherence. Mann and Thompson would consider sentence 8A a single discourse unit.

8A Pinochet was arrested in the UK.

2.2.1.2 Intention and coordination

Coherence allows a writer to formulate complex messages. Coherence relations are often asymmetrical: if two sentences cohere, one passage may be more central to the writer's purpose than the other. In text 6A–6B, if the author's intention is to inform on the history of the solar system, sentence 6A is subordinate to 6B in the sense that it serves to elaborate on or enhance credibility of the other passage (Hobbs, 1985; Grosz and Sidner, 1986; Polanyi, 1988; Mann and Thompson, 1988). This interpretation renders the second sentence dominant, as the interpretation of the first relies on its relation to the second. If two passages cohere but they are of equal importance to the writer's intention, the relation is coordinate. Mann and Thompson (1988) call a superordinate participant of a relation the *nucleus*, while its subordinate counterpart is the *satellite*. The satellite's sole purpose is to increase the reader's understanding or belief of what is said in the nucleus. If related passages are of equal importance to the author's intention, both are nuclei and the relation is *multinuclear*. For instance, elements of a temporal sequence (e.g. *first ...; then ....*) are of equal importance and participate in a multinuclear relation.

2.2.1.3 Hierarchy

If a coherence relation holds between two elementary discourse units, together they constitute another discourse element. Composed elements may in turn participate in a coherence relation as if it were an elementary discourse unit (Hobbs, 1985; Grosz and Sidner, 1986; Polanyi, 1988; Mann and Thompson, 1988). Under a complete analysis, a coherent text is structured hierarchically, as a tree, in which the top nodes are the most representative of the writer's message.

The hierarchical nature of coherence was recently challenged by Wolf and Gibson (2005). They argue that the presence of crossed dependencies and nodes with multiple parents render the tree representation of discourse structure inappropriate. If the hierarchical constraint is maintained, passages are forced into unintuitive discourse relations in order to avoid illegal structures. Wolf and Gibson supported their argument with a study on a corpus of naturally occurring text in which the occurrence frequency was measured of relations violating the tree constraint. The corpus of 135 texts was manually annotated by their guidelines, similar to the Rhetorical Structure Theory (RST) corpus of Carlson et al. (2001), but without enforcing the tree constraint. Wolf and Gibson report very high frequencies of tree-violating relations, which could present a significant problem for the tree representation of discourse. However, their results also show that this phenomenon is primarily local. Combined with the fact that they use a fine grained segmentation, this may alleviate the problem, as the ratio of tree-violating relations may be related to the size of the segments. Moreover, Mann and Thompson (1988) identified a number of shortcomings of present discourse models, which may provide an alternative explanation to the findings of Wolf and Gibson. First of all, ambiguity may lead to multiple valid interpretations, in which case a distinct trees can be used for each interpretation. Mann and Thompson also report simultaneous analyses, i.e. multiple compatible trees representing 'parallel' interpretations. Ambiguity and simultaneous analyses are not discussed in Wolf and Gibson (2005).

2.2.1.4 Taxonomy

Theories of discourse organization categorized coherence relations into a finite (discrete) set of relation types. Much less than on the hierarchical character of text, there is consensus on the taxonomy of relation types. Hobbs (1985) proposed 8 relation types. Grosz and Sidner (1986) identified two types of functional relations between passages: dominance and satisfaction-precedence, where satisfaction-precedence applies when the purpose of one passage must be satisfied before the other. Mann and

Thompson (1988) argued for two broad classes of relation types: presentational and subject-matter, each of which is subdivided into several sub types. Subject-matter relations include causality and temporality. Rather than to inform, presentational relations are typically used when the writer intends to increase the reader's belief of something or to change the reader's attitude. In total, Mann and Thompson proposed a set of 24 relations types, which was later extended to 32. Similar binary classifications were proposed by Redeker (1990) (ideational/pragmatic) and Sanders and van Wijk (1996) (semantic/pragmatic). A more fine grained taxonomy has been developed by Carlson et al. (2001) (78 relations in 16 classes). Marcu and Echihabi (2002) used a unified taxonomy of four relations, based on relations proposed by others. Mann and Thompson (1988) remark that no one taxonomy may be generally appropriate for all genres. For this reason, Grosz and Sidner (1986) strongly argue against the use of fine grained taxonomies: the range of possible purposes of passages in discourse is open-ended.

Although the way text coheres is (largely) independent from its realization, Knott and Dale (1995) argued that its realization may well provide evidence of the existence of coherence relations. They designed a protocol to extract cue phrases from text, and to cluster them by function. Each function corresponds to a coherence relation.

2.2.2 Rhetorical Structure Theory

Of the theories discussed, the Rhetorical Structure Theory (RST) of Mann and Thompson (1988) is currently the most influential. Although RST was intended for use in text generation (Mann and Thompson, 1988), it is applied in many applications, including automatic summarization (Marcu, 1999). The use of RST was encouraged by the availability of an extensive annotated corpus of English news articles (Carlson et al., 2001). Good levels of agreement have been reported between human annotators of RST, which indicates that RST is well defined (Mann and Thompson, 1988; den Ouden, 2004).

RST aims at describing coherence in monolog text. Other theories focus on specific genres, such as instructional text (Sanders and van Wijk, 1996), or generalize to dialog (Polanyi, 1988). As various theories address different issues, their applicability has to be weighed for each application and genre individually. For summarization, RST has significant advantages over other theories: mainly the availability of annotated corpora and past research on RST-based summarization makes RST attractive. Therefore, RST will be used as a starting point for discussing manual and automatic annotation of coherence relations.

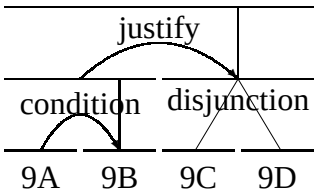

Figure 2.2: An example of a rhetorical structure analysis.

RST is a method for analyzing the intentional structure of text in an hierarchical manner. RST originally described a set of 24 subordinating (directed) and coordinating (multi-nuclear) relations. An example of an RST analysis is shown in Figure 2.2. The discourse units of this analysis are 9A, 9B, 9C and 9D. I use the notation introduced by Mann and Thompson (1988), in which the arrows represent subordinating relations with the arrow pointing to the dominant participant (nucleus); *disjunction* is a coordinating relation. Thus, according to this analysis, 9C and 9D are the most central to the writer's purpose, as they are not subordinate to any other discourse unit. If 'importance' of a sentence is measured by the number of subordinating relations that separate the sentence from these discourse units, the next-most important is 9B, followed by the least important, 9A.

2.2.3 Manual annotation

There is no correct or incorrect theory of discourse organization, only more and less useful theories, depending on the application (Mann and Thompson, 2000b). Arguably the most important criterion for the usefulness of a theory of discourse organization is the possibility of consistent and reproducible manual annotation in accordance with the theory. If a text can be annotated manually with high inter-annotator agreement, it is possible to annotate automatically as well, given the availability of sufficiently sophisticated machines. Therefore, annotation procedures are a central issue in discourse analysis. Discourse analysis can be divided into three (interdependent) sub tasks:

1. identifying discourse elements;
2. identifying the organizational structure of discourse;
3. identifying (labelling) structural relations.

Carlson et al. (2003) created a corpus of RST analyses of newspaper articles, and developed a corresponding annotation procedure for their interpretation of RST. They

used a finer grained segmentation, more discourse relations and more restricted tree structures than 'classic' RST as defined by Mann and Thompson (1988). In order to avoid circular dependencies, segmentation was done prior to identifying relations. Carlson et al. used a bottom-up approach to structure annotation: the first step is to identify a relation and its label between two segments. Once two segments are related, they act as a newly created segment which may in turn be in relation with another segment. The analysis is complete when the analysis tree is fully connected. In contrast, Hobbs (1985) used the reverse procedure. The intuition is that the sharpest topic break should be identified first. This results in two related segments, which can be further divided until the desired segmentation level is reached. The bottom-up approach leaves the order in which relations are marked open to the annotator. Since decisions in RST analysis are restricted by earlier decisions, the particular order may affect the final outcome of the analysis. Lascarides and Asher (1993) advocate a left-to-right approach, where the left-most segments are connected first. Others abandoned this idea (e.g. Stede and Heintze, 2004), claiming that the full picture often cannot be determined when reading the text until a certain point. Instead, their annotators first marked the most salient (signalled) relations before moving on to marking relations which require deeper understanding of the text.

2.2.4 Automatic annotation

Research on automatic annotation of coherence relations has concentrated mostly on RST. Automatic annotation involves the same three steps as manual annotation: segmentation, relation identification and combining those relations into a coherence analysis.

2.2.4.1 Segmentation

Marcu (1997b) devised a segmentation algorithm for detecting boundaries of elementary units in English text for RST analysis, based on a number of hand-crafted rules. The algorithm uses punctuation and cue phrases (*for example*, *but*, etc) to identify boundaries. He reports over 80% recall and 90% precision of detected boundaries in a small corpus (344 sentences; 643 elementary units). A limitation of Marcu's approach is that different uses of cue phrases are not distinguished. For instance, the algorithm anticipates the use of *but* as a conjunction, and unjustly segments the following sentence (c.f. Hirschberg and Litman, 1993).

10A The U.S. has
10B *but* a slight change to win a medal in Atlanta.

To increase accuracy, Corston-Oliver (1998) used a combination of syntax and cue phrases for boundary detection. It is unclear if this leads to improvement.

2.2.4.2 Relation identification

Apart from as indicators for segmentation, cue phrases are used for identifying discourse relations. Considering the number of cue phrases text exhibits, their role in identifying relations is significant. However, the relation identification task confronts the automatic analyzer with a number of additional issues.

As follows from the previous section, the discourse organization is functional (and thus semantic) of nature. Cues for recognizing discourse structure automatically rely on the way coherence reflects in realization. The signalling of coherence relations in spoken film descriptions was studied by Redeker (1990). In her corpus of 3,585 clauses (of which 1,897 from dialogs and 1,688 from monologs), approximately half was signalled by connectives such as conjunctions (e.g. *because, and*), relative pronouns (e.g. *that, who*), temporal expressions (e.g. *then, after that*), and discourse markers (e.g. *okay, well*). Although the relative use of specific categories of connectives varied between different classes of discourse, their total number was roughly the same for all subjected texts. A study on German newspaper text showed a smaller number (35%) of signalled coherence relations (Stede and Heintze, 2004). Schauer and Hahn (2001) included types of coreference relations (definite noun phrases, bridging) as an indicator of coherence that were excluded from previous studies. They concluded that in their corpus, up to 75% of coherence relations can be identified using a combination of cue phrases and coreference. However, it should be noted that identification of coreference relations is by no means trivial.

First, cue phrase disambiguation for relation identification is harder than for segmentation. In segmentation, it suffices to distinguish discourse markers from non-discourse markers. When identifying relations, one must be able to recognize not only the presence of a relation, but also the relation's type and scope. A cue phrase may be an indicator of more than one relation. For instance, *but* may indicate the relations of CONTRAST as well as CONCESSION or ANTITHESIS. RST imposes no restrictions to the scope of a relation: a relation may hold between clauses or, on the higher levels of the discourse tree, between sequences of sentences or paragraphs. For the purpose of his automatic RST annotation system, Marcu (1997b) derived information from a

corpus of manual annotations as to how the relations are used. Marcu found differences in the relation type, the satellite/nucleus order and the scope of relations, which correlated with the use of particular cue phrases.

Secondly, the use of cue phrases is not sufficient to derive a full RST tree. In his search for alternative indicators, Marcu (1997b) measured co-occurrence statistics. Inspired by lexical cohesion and lexical chains (c.f. Halliday and Hasan, 1976; Morris and Hirst, 1991), Marcu interpreted a low word concurrence between adjacent passages as a topic shift. Thus, these passages are less related than passages with a higher word concurrence.

Seemingly, the most obvious cue for relation identification Marcu (1997b) used is information about the layout of the text. Paragraphs and sentences are used by the author to convey information on the discourse structure. The boundaries between them signify topic shifts and, if marked, can be used to constrain the annotation process. The annotation system of Marcu related sentences or paragraphs only as a whole; a relation between part of a paragraph and sentences of other paragraphs was not allowed. There is nothing in RST which prevents a clause of a sentence to be related to another sentence, but Marcu found that such relations rarely occurred in his corpus.

More recently, Marcu and Echihabi (2002) hypothesized that there are certain words which by themselves do not provide much information about the presence of a relation, but when they occur together, they do. Consider the following example:

11A Yesterday, the sky was blue.

11B Today, the sky was grey.

There is no explicit link or signalled relation between the two sentences, but there are various instances of lexical cohesion. For instance, a *contrast* is conveyed by the use of the words *yesterday* and *today*. Marcu and Echihabi (2002) applies machine learning on a large corpus of raw (unannotated) text in order to derive *rhetorical* relations (such as *contrast*) from pairs of words in different sentences (such as *yesterday*/*today*). The machine learning method of Marcu and Echihabi consists of two steps. In step 1, the corpus is segmented and potential relations are marked, based on cue phrases. In step 2, concurrence frequencies of word pairs are extracted for each relation type. Once the database of word pair frequencies is constructed, these data can be applied on unseen text to identify relations. In the case of the above example, a high frequency of the triple (rained,sun,*contrast*) would indicate the presence of a *contrast* relation between the two sentences.

2.2.4.3 Building a tree structure

Once relations between arbitrary spans of text are identified, Marcu (1997b) derives a full parse for a text by combining those relations into a single tree. To this end, he uses the confidence values of recognized relations to assign a confidence value to the tree as a whole.

The summarization algorithm of Marcu (1997a) requires a single coherence hierarchy for summarization. Others suggest individual relations are useful as such (Blair-Goldensohn and McKeown, 2006). If a full hierarchy is not a requirement for the application at hand, it may be preferable to use the recognized relations and their confidence values directly, as information (such as confidence values and incompatible relations) is lost during construction of the RST tree.

2.2.5 Multimedia

Coherence plays a role on an intentional level rather than on the level of realization. Although RST is developed for describing coherence in text, André (1995) argued that RST largely abstracts from realization of information in particular media. André applied RST to multimedia documents containing text and images, after adding a few relations to the relation set of Mann and Thompson (1988) that do not appear in text-only documents. André used RST for generating coherent multimedia. Delin et al. (2002) included RST in their multi-layered multimedia annotation scheme. Other multimedia annotation schemes have been developed (see Geurts et al. (2005) for an overview), but they typically aim at describing the multimedia content itself, and fail to capture semantic interrelationships between modalities.

While André and Delin et al. use the same set of RST relations to annotate image-text relations as to annotate text-text relations, Mann and Thompson acknowledged that specific applications may call for specific relation sets. Levin (1981) studied image-text relations in educational documents for children.

While the coherence model of Mann and Thompson (1988) describes the argumentative structure and understanding of text, Levin focus on the role of images in learning and memorizing. Levin discovered eight relations, ranging from *decorative* to *organizational* (i.e., the image helps integrating information) and *interpretative* (i.e., the image helps comprehension). Marsh and White (2003) created a model specifically for image-text relations, but applicable in any domain. They analyzed documents from a variety of sources and invented a hierarchical taxonomy of image-text relations. On the

highest level, they used three relations: weak image-text relations (e.g. decorative images), strong image-text relations (e.g. images which concretize the text), and images which add entirely new information. Each of these three broad categories of relations are narrowed down to a total of 35 relation types.

Martinec and Salway (2005) proposed a multi-layered annotation scheme for text-image relations. The first layer of annotation is what they call *status*: either the image or the text is subordinate to the other, or if they are not, the image and the text are complementary or independent. That amounts to four possible relations. The *status* of Martinec and Salway is comparable to what is annotated by Levin and Marsh and White. On the other hand, Martinec and Salway also recognized the need for annotation of image-text relations of the level of rhetorical relations. The *status* layer is complemented by a layer of what they call *logico-semantic* relations, which resemble close similarities with subject matter relations of RST.

2.3 Cross-document relations

A document is designed to have structure. That is what makes it a document rather than just a collection of sentences. When searching information, we typically have to deal with a number of documents which may (or may not) provide some of the information we seek. Should we regard these documents as a coincidental collection of documents, or as a cluster with an internal structure, whose documents share certain properties or are in some other way related? Exploiting cross-document relations has been successful in information retrieval. Brin and Page (1998) indexed web pages not only by their contents but also by the labels of links referring to them. In the generation of summaries of multiple news articles, a major concern is identification of redundant sections, as to avoid providing the same information twice (e.g. Mani and Bloedorn, 1999). For creating 'update summaries' — summaries to provide the new information in an article with respect to a number of earlier publications — the publication date provides helpful clues to how documents relate.

Trigg and Weiser (1986) devised a framework for relating and structuring scientific papers in various ways. Although Trigg and Weiser go beyond the level of citations, scientific papers (and also web pages, c.f. Brin and Page, 1998) have the advantage of containing explicit links between documents. Radev (2000) designed a coding scheme for cross-document relations (Cross-document Structure Theory, CST) aimed at general applicability, but the application he has in mind is multi-document summarization

of news articles. His work was inspired by the work by Mann and Thompson (1988) on coherence (RST), but unlike coherence analysis, analysis of cross-document relations cannot rely on an author-intended structure. This forced Radev to deviate from RST in a number of ways. For instance, he dropped nuclearity in relations. More importantly, he created a taxonomy for cross-document relations from scratch. The taxonomy includes information-level relations (e.g. equivalence, subsumption, contradiction), relations regarding the perspective or opinion of the author or changes in the state of affairs (e.g. agreement, judgment, follow-up, change of perspective), and relations indicating differences in the level of detail (e.g. attribution, refinement, elaboration). CST was applied in summarization by Zhang et al. (2002) in a study using manually identified CST-relations, but practical application requires automatic recognition of relations. An attempt to do this was made by Zhang et al. (2003). They claim to have achieved promising results, but also report on a number of problems with hand-coding CST as well as automatic relation recognition.

Two CST-relations in particular received attention in multi-document summarization: subsumption and equivalence. Equivalence is established by paraphrasing: paraphrases are ways to express the same meaning. A special case of paraphrasing is synonym detection. Synonyms, but also other useful word-to-word relations such as generalization (hypernymy), can be looked up in a thesaurus, if available (e.g. Wordnet, Miller, 1995). However, thesauri such as Wordnet face a number of problems. First, thesauri are constructed manually for each language, which is a laborious and expensive process. Second, as thesauri are expensive to build, they are available in few languages with limited coverage in most. Even in WordNet, a large thesaurus for English, not all domains are equally covered. Third, the use of language varies with the domain and perspective. Words may be used interchangeably in one situation and differ in meaning in another. As a result, one may find synonyms which do not apply in the particular context of interest. These problems may be alleviated by automatic synonym mining, e.g. by means of matrix decomposition methods such as singular value decomposition (Deerwester et al., 1990). These methods are used to detect that certain terms often co-occur or appear in a similar context. This is used as evidence that the words are synonyms.

Appearance of different lingual expressions in a similar context is also the basis of the approach to sentence-level paraphrasing of Barzilay and Lee (2003). Their intent is to extract paraphrase lattices from a corpus of comparable (not necessarily parallel) corpus. For example, given the paraphrases *killing two other people and wounding 27*, and *killing himself and injuring seven people*, if we can recognize the similar structure,

we could derive a pair of templates, *killing X and wounding Y*, and *killing X and injuring Y*. This idea led to the construction of the DIRT paraphrase corpus (Lin and Pantel, 2001), although Lin and Pantel used a more simple representation of paraphrases. They represented a phrase as a path in a dependency tree between two nouns, connected by a verb. If two nouns are found to be connected by the same path in multiple occasions, the paths are taken as paraphrases. An example of a pair of paraphrases in the DIRT corpus is *X produces Y* and *X manufactures Y*. Since paths in DIRT are relatively short and contain exactly one verb, DIRT concentrates on paraphrasing verbs. Marsi et al. (2007) applied the DIRT corpus for detecting textual entailment.

Parallel corpora – useful for training machine translation systems – are also a useful resource for learning paraphrases (Bannard and Callison-Burch, 2005). Bannard and Callison-Burch mine paraphrases from a parallel corpus by searching for differences in translation of the same phrase. For instance, if phrase *a* is translated to *b* in one instance and to *c* in another, phrases *b* and *c* are taken as paraphrases. The paraphrasing method of Bannard and Callison-Burch is discussed in greater detail in section 3.3.2.2.

Paraphrasing is quite similar to the problem of recognizing textual entailment. Recognizing textual entailment between two passages is the task of determining whether the truth of a passage can be inferred from another passage (Monz and de Rijke, 2001; Dagan et al., 2006). Recognizing textual entailment as a natural language processing task is discussed in greater detail in chapter 3.

2.4 Conclusion

Cohesion and coherence are relevant for interpreting individual sentences, and identifying their function in text. Cohesion allows to stage a context necessary for understanding. Coherence allows a message to span more than one sentence and explains information-level differences between a text and its parts. Although cohesion, coherence and grammar are distinctly different phenomena, there is interaction between them that might be helpful for getting a more complete (and useful) model of discourse. For instance, grammatical restrictions to the use of cohesion (e.g. pronoun agreement in gender or number) help resolving cohesive ties, and conjunctions may help recognizing coherence relations in text.

3

Entailment recognition

In many applications, relating text from different documents is just as important as relating sentences within a document. This chapter examines the problem of recognizing textual entailment (RTE), a specific type of textual relations across documents. This chapter contributes to three aspects of recognizing entailment: representation, matching and evaluation. A new evaluation method is proposed which has quantifiable advantages over existing methods. Furthermore, I compare various new and existing representation and matching methods in an RTE system on PASCAL-RTE and CLEF-AVE corpora. No method consistently outperformed another method on all of the subjected corpora.

Progress in summarization is hindered by the lack of a model to describe the interrelationships between documents in detail. When the source text of a summary comprises multiple documents, redundancy becomes an important issue. Redundancy detection has received attention in summarization, but is usually just used for reranking candidate sentences, in order to reduce the probability of including redundant content (e.g. Carbonell and Goldstein, 1998). Redundancy detection could facilitate summarization better if it was a fundamental part of the summarization process. An example of how redundancy can be exploited in summarization is the work of Mani and Bloedorn (1999), who tried to find the intersection of the content of documents for summary generation. A step in the direction of an integral representation of inter-document relations is made with the RST-inspired Cross-document Structure Theory (CST) of

Radev (2000) (see chapter 2.3). CST includes an elaborate coding scheme of cross-document relations, including not only redundancy but also relations such as citation, contradiction, etc. I take one step back and focus on one particular CST relation between passages of different documents, subsumption, and a similar relation which is not included in the CST relation set, entailment. Although there is arguably a subtle difference between subsumption and entailment, for practical reasons, I consider them synonymous.

The reason to focus on entailment is that the relation is not specific for any domain but generally applicable in multi-document summarization (CST is specifically designed for news summarization). Applications of RTE include question answering systems which exploit identification of entailment to verify whether a hypothesized answer is entailed by a text document. A summarization system may use knowledge of entailment relations between documents to decide which information is redundant and which information to include in the summary. Other potential text processing applications of RTE are information extraction and information retrieval (Dagan et al., 2006).

The process of recognizing textual entailment consists of two stages. First, both passages are converted to a system-internal representation. Then, the passages are compared in order to decide whether an entailment relation holds between them. In the context of this work, an entailment system was implemented which uses one of three possible text representations: syntactic tree, sequence of words, or bag of words. For aligning the passages, the system uses paraphrasing or lexical matching, optionally with IDF-weighting (Spärck Jones, 1972). We presented a variant of the tree representation algorithm with lexical weighting in Marsi et al. (2006). The paraphrasing algorithm was presented in Bosma and Callison-Burch (2007). In this chapter, the relative performance of each of the representation methods and alignment methods is measured.

3.1 Related work

Vanderwende and Dolan (2005) analyzed the PASCAL-RTE test set (Dagan et al., 2006) to investigate what percentage of entailment relations could be recognized by using lexical-syntactic information. They concluded that a respectable 34 percent of the samples could be solved using syntax alone. Bar-Haim et al. (2005) followed up on this work with a more elaborate analysis, evaluating not only the use of lexical-syntactic in-

formation, but also other inference mechanisms such as synonymy, meronymy, syntactic transformations and paraphrases. They show that paraphrases have large potential of increasing an entailment system's performance on the PASCAL-RTE corpus.

Although hybrid approaches are possible, most RTE systems use one of four representations of text: the multiset representation ('bag of words'), a sequence, a tree, or a logical representation. Each of these representations can be combined with different inference algorithms: altering the relative weight of words, measuring overlap or edit distance, word stemming, synonymy, etc.

While Vanderwende and Dolan (2005) and Bar-Haim et al. (2005) use manual annotations, this study seeks to compare performance and analyze the problems of using various representations and matching algorithms for fully automatic textual entailment recognition.

3.2 The task

A proper definition of the task is essential for a good understanding, for devising algorithms and for evaluation of systems for automatic recognition of RTE. We could also reverse the argument: agreement on evaluation methods is required to define and understand the task, because evaluation methods essentially define the task. For practical reasons, I follow the definition of RTE as implied de facto by annotated entailment corpora.

3.2.1 Corpora and evaluation platforms

The most notable annotated entailment corpus is the corpus used in the PASCAL-RTE evaluation program, which started in 2005. One of the corpora used for performance measurements in this chapter is the corpus of the second PASCAL-RTE challenge (2006), henceforth referred to as the RTE2 corpus (Bar-Haim et al., 2006). The PASCAL-RTE program intends to bring together researchers from different areas of NLP who are interested in recognizing entailment. Specifically, the 2006 challenge focuses on information retrieval (IR), information extraction (IE), question answering (QA) and summarization (SUM).

The corpus consists of a set of passage pairs and a corresponding entailment value (positive or negative). One of the passages is labeled the *text*; the other is the *hypothesis*. If the hypothesis is entailed by the text, the pair is a positive example of entailment. The passages are usually single sentences, but multiple sentences are allowed.

Pair: RTE2/IR pair 1 (entailment: negative)

Text: As a result of these weaknesses, computer systems and the operations that rely on the systems were highly vulnerable to tampering, disruption, and misuse from both internal and external sources.

Hypothesis: Non-authorized personnel illegally entered into computer networks.

Pair: RTE2/SUM pair 2 (entailment: positive)

Text: Prime Minister Mahmoud Abbas has offered 'the hand of peace' to Israel after his landslide victory in Sunday's presidential election.

Hypothesis: Mahmoud Abbas has claimed victory in the presidential elections.

Pair: RTE2/IE pair 3 (entailment: negative)

Text: ECB spokeswoman, Regina Schueller, declined to comment on a report in Italy's La Repubblica newspaper that the ECB council will discuss Mr. Fazio's role in the takeover fight at its Sept. 15 meeting.

Hypothesis: Regina Shueller works for Italy's La Repubblica newspaper.

Pair: RTE2/QA pair 6 (entailment: positive)

Text: Muslim fundamentalists such as the Islamic Resistance Movement, also known as Hamas, and the smaller Islamic Jihad are determined to torpedo the peace process.

Hypothesis: The Islamic Resistance Movement is also known as Hamas.

Pair: CLEF-AVE 2006 pair 1 (entailment: negative)

Text: We will do this on Monday when Atlantis returns to Earth with the satellite German physicist Klaus Grossmann said .

Hypothesis: Atlantis is returns to Earth. [sic]

Figure 3.1: Samples from PASCAL-RTE and CLEF-AVE corpora. Note that CLEF-AVE may contain ungrammatical hypotheses because the corpus is based on machine output.

Because different applications of natural language processing cope with entailment in a different way, the RTE2 corpus is divided into four sub corpora, each corresponding to a different application (IR, IE, QA, and SUM). One of the differences between the corpus section is that the sentences to be entailed in information extraction are typically

Table 3.1: Performance matrix: a text/hypothesis pair has a manually assigned entailment value (positive: A_Y; or negative: A_N) and a system-labeled entailment value (positive: L_Y; or negative: L_N).

	A_Y	A_N
L_Y	$X_{Y,Y}$	$X_{N,Y}$
L_N	$X_{Y,N}$	$X_{N,N}$

quite short and uniform, while in summarization, entailment of longer sentences may need to be established. Examples from the RTE2 data are given in Figure 3.1.

In addition to the RTE2 corpus, I experimented with the English section of the CLEF 2006 Answer Validation Exercise (CLEF-AVE) corpus (Peñas et al., 2007). This corpus is based on results of QA systems participating in the CLEF QA evaluation program. A set of questions was composed for QA systems to return (1) an answer to the question, and (2) the text snippet in which the answer was found. For a subset of questions, a template was created to transform the QA system's answer into a RTE-style entailment candidate, i.e. a full sentence. The answers were manually judged for correctness, and this judgment was used to determine the entailment value. The result was another entailment corpus, similar in form to the PASCAL-RTE corpora, although the type of problems that need to be addressed is slightly different, mainly because (1) the pairs were created by QA engines; and (2) the entailment value is determined indirectly by judging correctness. Annotators may take different decisions when asked to judge correctness of answers or when asked to determine entailment between the resulting pairs.

3.2.2 Measuring performance

The goal of a performance metric is to rank systems, algorithms or methods by performance. If algorithm *A* is 'better' than *B* — by some definition — a metric should be able to detect this difference by assigning a significantly better score to *A* than to *B*. RTE systems typically produce a *confidence* value for each text/hypothesis pair, representing the confidence of the claim that the hypothesis is entailed. If the confidence exceeds a certain threshold, the system decides favorably to labeling the hypothesis 'entailed' (positive). Otherwise, the pair is labeled 'not entailed' (negative).

While it has been suggested that outcomes of an entailment test other than positive and negative are possible, the primitives of most performance metrics are the num-

ber of pairs in the four quadrants of the performance matrix in Table 3.1. Possible extensions are discussed in section 3.2.2.3. Alternatively, ranked lists can be used to derive performance measures such as *average precision*.[1] Average precision requires a list of text/hypothesis pairs ranked from confident of entailment to confident of non-entailment. As average precision approaches RTE as a ranking problem, it can only be calculated for RTE systems which produce a ranking rather than a binary decision.

In PASCAL-RTE and CLEF-AVE, recognizing entailment is understood as a positive/negative classification problem Dagan et al. (2006); Peñas et al. (2007). Pairs are manually annotated as belonging to either of the two sub sections of the corpus: positive (A_Y) or negative (A_N); and pairs are labeled by the system as positive (L_Y) or negative (L_N). RTE systems face a trade-off between recognizing positive pairs and recognizing negative pairs. Therefore, to measure performance, metrics such as accuracy and F-measure take both aspects of performance into account. The ability to recognize positive pairs can also be measured separately as *recall*, i.e. the ratio of correctly recognized positive pairs (eq. 3.1 below). Similarly, the ability to recognize negative pairs can be measured as the ratio of correctly recognized negative pairs (eq. 3.2).

$$R_Y = \frac{X_{Y,Y}}{X_{Y,Y} + X_{Y,N}} \tag{3.1}$$

$$R_N = \frac{X_{N,N}}{X_{N,N} + X_{N,Y}} \tag{3.2}$$

where R_Y is the recall of positive pairs; and R_N is the recall of negative pairs.

Accuracy and F-measure can be written as a function of R_Y and R_N, but they also depend on the generality of positive pairs in the data set. The *generality* G is the ratio of pairs in the corpus that are in A_Y:

$$G = \frac{A_Y}{A_Y + A_N} = \frac{X_{Y,Y} + X_{Y,N}}{X_{Y,Y} + X_{Y,N} + X_{N,Y} + X_{N,N}} \tag{3.3}$$

If a metric depends on the generality, the relative effect on the final performance score of system's ability to recognize positive and negative pairs changes with the generality. Hence, the generality affects the measured performance of the system and potentially the system ranking. In anticipation, a system may bias its judgment toward labeling pairs as positive or negative, depending on the expected generality, as there is a trade-off between recognizing positive and negative pairs. As mentioned earlier, the goal of a metric is to detect whether one system is better than another. If the generality

[1] Average precision was used in an optional evaluation in the third PASCAL-RTE challenge.

affects the system ranking, estimating generality is part of the task as presented by the combination of metric and corpus. Because RTE systems are typically embedded in a larger system to be functional to a user, it is not clear how RTE systems should deal with a variation in generality. Therefore, it seems plausible to not include generality as a factor of system performance and use a generality-independent metric instead.

In PASCAL-RTE, this issue is solved by normalizing the corpus to a generality of (approximately) fifty per cent. CLEF-AVE corpora are constructed from the output of question answering systems. Given a question, a question answering system produces an answer and a text snippet from which the answer was extracted. The question and answer are then rewritten to a hypothesis, of which an RTE system should be able to recognize whether its truth follows from the text snippet. This is a cost-effective way of corpus construction, but the corpus is constructed by machines of which the behavior is not known to the entailment system. As a result, the measured performance of an entailment system depends on the 'coincidental' set of machines which created the corpus.

Specificities of the machines involved should be predictable or taken out of the equation. Otherwise the resulting corpus poses an ill-defined task. An important part of the dependence of the corpus on individual question answering systems is the (unpredictable) variation in generality, which depends on the performance of question answering systems. CLEF-AVE corpora are not normalized.

3.2.2.1 Accuracy, F-measure, and generality

Accuracy is used to measure performance in the PASCAL-RTE challenges. Accuracy is calculated as the number of correctly labeled pairs (eq. 3.4). Accuracy can also be written as a function of R_Y, R_N and G (eq. 3.5).

$$S_{acc} = \frac{X_{Y,Y} + X_{N,N}}{X_{Y,Y} + X_{Y,N} + X_{N,Y} + X_{N,N}} \tag{3.4}$$

$$= R_Y G + R_N(1-G) \tag{3.5}$$

where S_{acc} is the system performance measured as accuracy.

Eq. 3.5 shows that accuracy is essentially an average of R_Y and R_N, weighted by G. As a (possibly undesired) result, a system evaluated by accuracy would focus more on optimizing R_Y at the cost of R_N if the corpus generality is known to be greater than fifty per cent, or vice versa if the generality is smaller than fifty per cent. For instance, if the generality is ten per cent, a system can achieve an accuracy of ninety

per cent by labeling all pairs as negative. State-of-the-art RTE systems may not beat that baseline, as any attempt to recognize positive pairs at the cost of a slight decrease in recognizing negative pairs is excessively penalized. In CLEF-AVE corpora, the generality is typically ten to twenty-five per cent.

There are two ways to remove dependence of accuracy on generality. The first is to normalize the corpus to a fixed generality, e.g. (near) fifty per cent, as is done in PASCAL-RTE. As an alternative, I propose to normalize the accuracy measure:

$$S_{Ga} = \frac{1}{2}(R_Y + R_N) \tag{3.6}$$

where S_{Ga} is the system performance measured as *G-accuracy*, a generality-safe alternative for accuracy. While accuracy is the generality-weighted mean of R_Y and R_N, G-accuracy is the arithmetic mean of R_Y and R_N. Note that eq. 3.5 and eq. 3.6 are equivalent if $G = \frac{1}{2}$.

While accuracy simply measures the ratio of correct judgments, the F-measure acknowledges that recognizing positive pairs is only useful at reasonable performance on negative pairs, and vice versa. F-measure is designed to measure how well both factors are balanced by taking the harmonic mean of *recall* and *precision*. Recall is R_Y, the ratio of positive pairs recognized (eq. 3.1); precision P measures the pollution by negative pairs of the set of pairs labeled positive (eq. 3.7, 3.8). The F-measure score S_F is the harmonic mean of R_Y and P (eq. 3.9).

$$P = \frac{X_{Y,Y}}{X_{Y,Y} + X_{N,Y}} \tag{3.7}$$

$$= \frac{R_Y G}{R_Y G + (1 - R_N)(1 - G)} \tag{3.8}$$

$$S_F = \frac{2 R_Y P}{R_Y + P} \tag{3.9}$$

As eq. 3.8 shows, precision is a function of generality. In essence, precision measures the system's ability to recognize negative pairs, but this is done as a complex function of R_Y, R_N and G. The F-measure depends on R_Y directly as well as indirectly, through P. If both R_Y and G are factored out as primitives of P, this leaves R_N, the ratio of correctly labeled negative pairs. Both R_N and P measure the system's ability to recognize negative pairs, but from a different perspective: R_N measures the ratio of correctly recognized negative samples; P measures the pollution of the set of samples labeled positive. Nevertheless, an increase of R_N results in an increase of P and

Table 3.2: The discriminativity of performance metrics, measured as the smallest observed system performance which is significantly ($p < 0.05$) better than a baseline system with an observed performance of $R_Y = R_N = 0.7$. Smaller numbers indicate the metric is more discriminative.

	PASCAL3		CLEF1	
performance metric	$R_Y{}^{\heartsuit}$	$R_N{}^{\diamondsuit}$	$R_Y{}^{\heartsuit}$	$R_N{}^{\diamondsuit}$
recall of positive samples (R_Y)	0.763	–	0.787	–
recall of negative samples (R_N)	–	0.763	–	0.729
accuracy (S_{acc})	0.787	0.792	0.993	0.731
G-accuracy (S_{Ga})	0.787	0.787	0.791	0.795
F-measure (S_F)	0.777	0.838	0.857	0.777
G-measure (S_G)	0.794	0.794	0.796	0.808

$^{\heartsuit}$ Smallest observed improvement of R_Y (with R_N unchanged) to cause a significant difference in performance.

$^{\diamondsuit}$ Smallest observed improvement of R_N (with R_Y unchanged) to cause a significant difference in performance.

vice versa. If R_N is used as a generality-safe alternative to precision, a generality-safe alternative to F-measure is the following:

$$S_G = \frac{2R_Y R_N}{R_Y + R_N} \tag{3.10}$$

where S_G is G-measure, the harmonic mean of R_Y and R_N.

3.2.2.2 Discrimination

The first requirement of a performance metric is to fit the task. The second is how well it succeeds to detect performance differences between systems, based on observed results. One metric is better than another if it is able to reliably detect smaller performance differences. In order to measure the quality of a metric, I gradually improve the results of a (non-existing) RTE system. The better the metric, the sooner the difference between the improved results and a baseline are significant. Whether one system can be shown to outperform another depends on (a) the observed performance of both systems; and (b) the corpus. An observation consists of the four values of the performance matrix in Table 3.1. This can be reduced to R_Y and R_N, the observed performance on the two sub sections of the data, A_Y and A_N respectively. Corpus variables of interest when determining significance are size and generality. For these experiments, the

variables of the test set of the third PASCAL-RTE challenge (800 samples; 51 per cent positive; henceforth called PASCAL3) and the test set of the first CLEF-AVE challenge (2088 samples; 9.5 per cent positive; henceforth called CLEF1) are used with different performance metrics.

The baseline observation I started with is $R_Y = 0.7$ and $R_N = 0.7$. These values are comparable to performance of state-of-the-art RTE systems Peñas et al. (2007); Giampiccolo et al. (2007). For each metric and the corpus variables of PASCAL3 and CLEF1, the performance is calculated. Then, one of R_Y and R_N is gradually increased (while leaving the other unchanged), until the observed increase of R_Y or R_N constitutes a significant difference in the performance metric used. Significance is determined by means of approximate randomization Noreen (1989). The increase of observed R_Y or R_N at which the performance difference is significant can be read from Table 3.2. For instance, when using accuracy (S_{acc}) to measure performance on the CLEF1 corpus, a system producing $R_Y = 0.7$ and $R_N \geq 0.731$ is significantly better than a system producing $R_Y = 0.7$ and $R_N = 0.7$.

Table 3.2 shows that accuracy is not suitable for measuring performance on the CLEF1 corpus, as it poorly detects differences in R_Y, although differences in R_N are well detected. This is not surprising: the effect of variations by chance of R_N on accuracy are greater than the effect of slight improvements of R_Y on accuracy. As a result, these improvements of R_Y cannot be measured with significant reliability. In general, accuracy is less discriminative of system improvements of R_Y if the generality is smaller than fifty per cent, and accuracy is less discriminative of system improvements of R_N if the generality is greater than fifty per cent.

F-measure poorly detects differences in R_N on PASCAL3 as well as differences in R_Y on CLEF1. In general, similarly to accuracy, F-measure is less discriminative of differences in R_N at a high generality (such as the 51 per cent of PASCAL3). As can be read from eq. 3.8, *precision* approaches 100 per cent as the generality increases to 100 per cent, regardless of R_Y or R_N. This reduces the effect of differences in R_N on F-measure. On the other hand, as the generality decreases, the effect of differences in R_N on precision increases, and precision itself decreases gradually, approaching zero as the generality approaches zero. Because, in F-measure, recall and precision are averaged as the harmonic mean, significant improvements of R_Y may be attributed to variation by chance in P (and indirectly, R_N). Thus, F-measure is less sensitive to improvements of R_Y at low generality.

G-accuracy and G-measure perform reasonably well under all circumstances. G-accuracy is slightly more discriminative than G-measure. Therefore, G-accuracy is

the preferred metric, unless the task demands a balance of R_Y and R_N, in which case G-measure is preferred.

3.2.2.3 Application

Two metrics are proposed as generality-safe alternatives to accuracy and F-measure, i.e. G-accuracy (eq. 3.6) and G-measure (eq. 3.10). G-accuracy is the arithmetic mean of R_Y and R_N; G-measure is the harmonic mean of R_Y and R_N. The harmonic mean is preferable if a balance of R_Y and R_N is required: a system which focuses on optimizing either R_Y or R_N scores low on G-measure but may still perform reasonably well on G-accuracy.

An advantage of G-accuracy and G-measure with respect to their generality-dependent counterparts is their applicability on all corpora. This opens up opportunities to harvest text/hypothesis pairs in a less restricted manner. The authors of CLEF-AVE corpora show that innovative ideas can lead to new ways to create corpora, but current performance metrics may be unsuitable for these corpora. G-accuracy and G-measure apply to all RTE corpora, regardless their generality.

So far, we assumed two possible outcomes of an entailment test: positive ($T \models H$) and negative ($T \not\models H$). *Contradiction* ($T \models \neg H$) has been suggested as a third possible outcome as a special case of non-entailment Giampiccolo et al. (2007). How well do the proposed metrics scale to interpretations of the RTE task involving more than two classes? If an arbitrary number of classes (e.g. positive, negative, ...) is used, recall can be calculated for each class individually. Recall R_c of samples of class c is calculated as follows:

$$R_c = \frac{c_c}{c_*} \tag{3.11}$$

where c_c is the number of correctly identified samples of class c; and c_* is the total number of samples of class c. Now, we can calculate for any number of classes the G-accuracy score (the arithmetic mean recall, eq. 3.12) and the G-measure score (the harmonic mean recall, eq. 3.14). If some classes are more important than others, e.g. if recognizing positive pairs is more important than recognizing negative pairs, a weighted average can be used for G-accuracy (eq. 3.13) and G-measure (eq. 3.15). The G-accuracy and G-measure values as calculated in eq. 3.6 and eq. 3.10 respectively, are the special case in which there are two equally weighted classes.

$$S_{Ga} = \frac{1}{\|C\|} \cdot \sum_{c \in \mathrm{C}} R_c \tag{3.12}$$

$$S_{Ga,w} = \left(\sum_{c \in \mathrm{C}} w_c \right)^{-1} \cdot \sum_{c \in \mathrm{C}} w_c R_c \tag{3.13}$$

$$S_G = \|C\| \cdot \left(\sum_{c \in \mathrm{C}} \frac{1}{R_c} \right)^{-1} \tag{3.14}$$

$$S_{G,w} = \left(\sum_{c \in \mathrm{C}} w_c \right) \cdot \left(\sum_{c \in \mathrm{C}} \frac{w_c}{R_c} \right)^{-1} \tag{3.15}$$

where $\|C\|$ is the number of classes; R_c is the recall of class c (eq. 3.11); and $w_c \geq 0$ is the relative weight of R_c in the performance measure.

3.3 Entailment experiments

3.3.1 Representation: tree, sequence or bag of words

Texts are modeled as a bag of words, a sequence of words or as a dependency tree:

- a *multiset* (or *bag*) is constructed by simply taking all words from the text snippet — a multiset is a set which may contain multiple instances of the same word;
- a *sequence* is a linearly ordered list of words;
- a labeled *tree* representation of a text snippet is derived from dependency trees generated by Minipar (Lin, 1998). In order to retain comparability between structural representations, only the terms and lemmas and the tree structure of the dependency tree are used; not part-of-speech tags or dependency relations. If a text snippet consists of multiple sentences, a new root node is created to which each of the sentences is attached.

Figure 3.2: The tree on the right is a valid subtree of the tree on the left.

3.3.1.1 Multisets and sequences

Although the same techniques (e.g. IDF, stemming) may be applied to each of these representations, matching algorithms are representation specific. In the case of a multiset, the entailment score $f_{multiset}(T,H)$ of a text T and a hypothesis H is calculated as the number of shared words divided by the number of words of the hypothesis:

$$f_{multiset}(T,H) = \frac{\|T \cap H\|}{\|H\|} \tag{3.16}$$

The longest common subsequence (LCS) is used as a measure of similarity between sequence representations of text snippets. In contrast to *substrings*, a subsequence does not require adjacency. For instance, $\langle 1,3 \rangle$ is a subsequence but not a substring of $\langle 1,2,3 \rangle$. LCS is also used by the ROUGE (Lin, 2004) summarization evaluation package to measure recall of a system summary with respect to a model summary. Here it is used to approximate the ratio of information in the hypothesis which is also in the text. A longest common subsequence of a text T and a hypothesis H is defined as a longest possible sequence which is a subsequence of both T and H. The entailment score of a text T and an hypothesis H is formally defined as follows, where the symbol $\sqsubseteq$ indicates the subsequence relation between two sequences.

$$f_{sequence}(T,H) = \max \left\{ \frac{\|Q\|}{\|H\|} \mid Q \sqsubseteq T;\, Q \sqsubseteq H \right\} \tag{3.17}$$

3.3.1.2 Trees

The definition for subsequence matching is extended to trees by measuring the largest common *subtree* rather than the longest common *subsequence*. The trees for text and hypothesis are dependency parses obtained from Minipar (Lin, 1998). If the text consists of multiple sentences, the dependency trees are joined by adding a new root node as the parent node of the root nodes of individual sentences.

For a tree to be a subtree of another tree, nodes may be 'skipped' in the hierarchy. For example, the tree on the right in Figure 3.2 is a subtree of the left tree although node B is missing in the hierarchical relation between A and C. However, the hierarchical relation must not be changed: if two nodes have a hierarchical relation in the subtree, they have a (direct or indirect) hierarchical relation in the supertree.

The algorithm used for aligning trees is presented in Marsi et al. (2006). Here, a modified version of the algorithm is used in order to enable different term weighting methods. The largest common sub tree of a text tree T and hypothesis tree H is defined recursively, but it ultimately relies on lexical similarity. For lexical matching, a function $sim(H,T)$ is used to return the lexical similarity of the words associated with the root nodes of H and T.

$$\begin{aligned} sim(H,T) &= 1, \quad \text{if } term(T) = term(H) \\ &= 1, \quad \text{if } lemma(T) = lemma(H) \\ &= 0, \quad \text{otherwise} \end{aligned} \tag{3.18}$$

$$f_{subtree}(H,T) = max\{align(H,T), skip_h(H,T), skip_t(H,T)\} \tag{3.19}$$

$$align(H,T) = sum\left\{ \frac{\|h_i\|}{\|H\|} \cdot skip_t(h_i,T) \mid h_i \in H \right\} + \frac{1}{\|H\|} \cdot sim(H,T) \Bigg\} \tag{3.20}$$

$$skip_h(H,T) = max\left\{ \frac{\|h_i\|}{\|H\|} \cdot f_{subtree}(h_i,T) \mid h_i \in H \right\} \tag{3.21}$$

$$skip_t(H,T) = max\left\{ f_{subtree}(H,t_j) \mid t_j \in T \right\} \tag{3.22}$$

The function $f_{subtree}(H,T)$ finds the largest common subtree of H and T by aligning nodes of H with nodes of T. The parent/child relation between nodes N and n_i respectively, is denoted by $n_i \in N$. The size of a tree of which N is the root node is denoted by $\|N\|$. Size depends on the number of nodes, but the size of individual nodes may vary, depending on the term weighting scheme used. The algorithm starts with the root of the text and hypothesis trees, and then traverses down the trees while aligning nodes to obtain the largest common subtree. At each step, the algorithm described by $f_{subtree}$ makes one of three possible decisions:

1. It adds the current node to the candidate-largest common subtree (eq. 3.20). Then, each of the hypothesis child nodes are aligned with the best possible text node.

2. It 'skips' the current hypothesis node and enters either of its child nodes (eq. 3.21). Consequently, any other child node is excluded from the alignment.

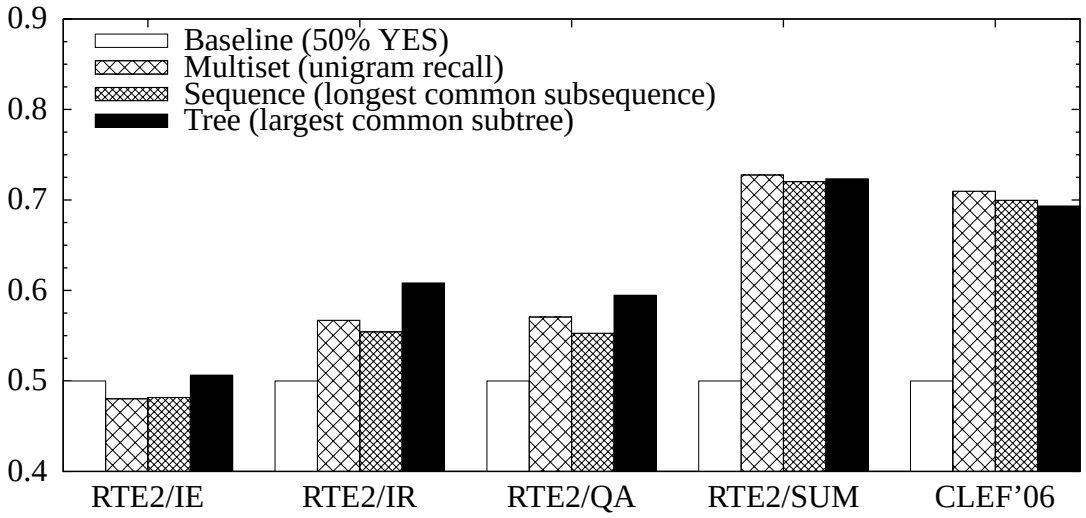

Figure 3.3: G-measure scores for RTE 2006 subsets (RTE2/IE, RTE2/IR, RTE2/QA, RTE2/SUM), and for CLEF-AVE 2006.

3. It skips the current text node (eq. 3.22). The current hypothesis node is aligned with the best possible child of the text node.

The entailment score is calculated as the ratio of nodes of H which are in the largest common subtree. This is to reflect the asymmetric nature of entailment: T entails H if information in H can be derived from T, but the reverse is not necessarily true. In effect, this means that nodes of T can be aligned with multiple nodes of H, but not vice versa, and only the number of aligned hypothesis nodes contribute to the entailment score, not the number of aligned text nodes.

The computational complexity of the algorithm is reduced by caching values $f_{subtree}(H,T)$ as soon as they are calculated. This reduces the complexity as a function of $||H||$ to $O(n^2)$.

3.3.1.3 Representation and performance

By using increasingly sophisticated data structures, more information is contained in the structure of the text itself in addition to its content. Figure 3.3 shows that using structural information increased performance on the RTE2/IE, RTE2/IR and RTE2/QA corpora: although the multiset representation performed better than or similarly to the sequence representation on all corpora, the tree representation outperformed the multiset by 5.5%, 7.3% and 4.2% respectively. On the other hand, the system which used

the tree representation performed worse on corpora which tend to have longer text and hypothesis snippets, i.e. RTE2/SUM (-0.6%) and CLEF-AVE (-2.3%). A possible explanation for this is that RTE2/SUM and CLEF-AVE contain longer sentences and thus more sentences with complex syntactic structures, which causes dependency parsing to be less reliable. In particular, the CLEF-AVE corpus contains more grammatical errors and incomplete sentences, again resulting in less reliable dependency parses.

With respect to differences in absolute performance on different corpora, the scores in Figure 3.3 are in line with earlier results. The IE subset of RTE2 appeared very hard during the RTE2 challenge, while the best results were achieved on SUM (Bar-Haim et al., 2006).

3.3.2 Alignment: IDF and paraphrasing

3.3.2.1 Inverse Document Frequency

Various methods have been proposed to assign weights to words. For instance, Rodrigo et al. (2006) performed Named Entity Recognition on the CLEF-AVE corpus to detect entailment by finding named entities which appeared in both the text and the hypothesis. In order to do so, they effectively assigned a weight of 0 to all words which are not part of a named entity. Using *inverse document frequencies* (IDF) is a more general way to assign greater weights to uncommon words – often content words – than to common words, which are usually function words (Spärck Jones, 1972). In contrast to named entity recognition, calculating IDF requires no linguistic knowledge and can be applied to any text.

The IDF value of a term in a set of documents is calculated as the logarithm of the number of documents in the document set, divided by the number of documents containing the term, or formally:

$$idf(t,D) = \log \frac{\|D\|}{\|\{d_i \mid d_i \in D \wedge t \in d_i\}\|} \tag{3.23}$$

where $idf(t,D)$ is the IDF value of term t in document set D.

Each of the presented alignment methods measure the percentage of the hypothesis that can be aligned in the representation used. The weight of a text string is its length, measured as the number of words. IDF weighting is applied by measuring the length of a text string as the sum of the IDF values of its words, instead of weighting each word as 1.

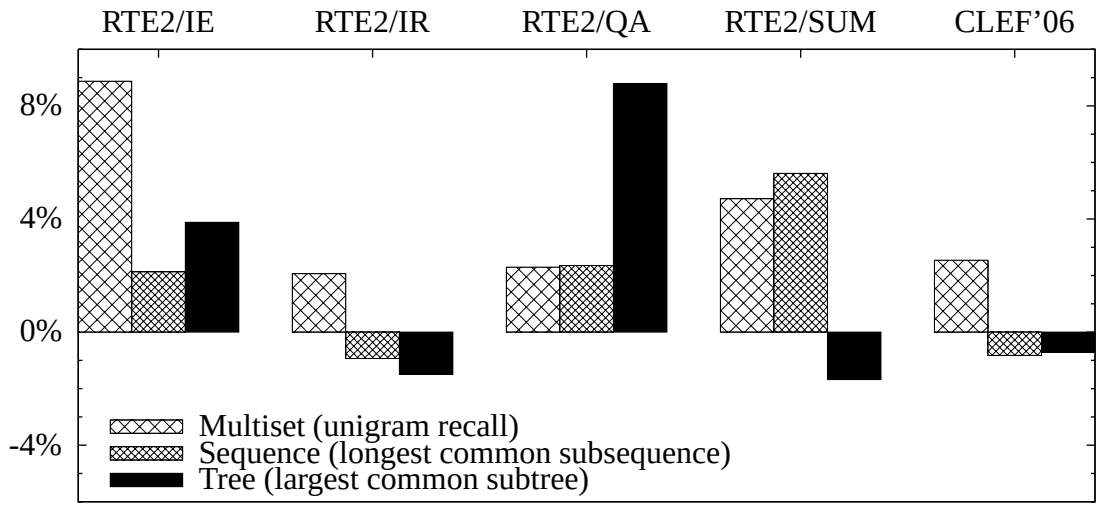

Figure 3.4: Relative improvement of performance when using IDF weighting of words.

Table 3.3: Examples paraphrases and probabilities for the phrase 'dead bodies'

Phrase	P
bodies	0.21
dead bodies	0.17
body	0.09
deaths	0.07
dead	0.07
corpses	0.06
bodies of those killed	0.03
the dead	0.02
carcasses	0.02
corpse	0.01

Figure 3.4 shows mixed performance effects of IDF weighting. Again, corpora containing short hypotheses appear to behave differently from corpora which use more elaborate text snippets, i.e. RTE2/SUM and CLEF-AVE. However, while IDF appeared to increase performance on RTE2/IE and RTE2/QA, it negatively affected RTE2/IR performance. This may be explained by the relatively frequent use of synonyms and paraphrases in RTE2/IR, RTE2/SUM and CLEF-AVE, compared to RTE2/IE and RTE2/QA, which is not covered by IDF weighting.

3.3.2.2 Paraphrase substitution

Integrating paraphrasing in an entailment system is a way to detect natural variation in language. Entailment can be detected if a paraphrase of the hypothesis is can be shown to be entailed by the text. For paraphrasing, a large corpus of paraphrases is used. Paraphrases were extracted from parallel corpora using the method of Bannard and Callison-Burch (2005). Bannard and Callison-Burch search for phrases which translate to the same phrase in another language. Candidate paraphrases were found by first identifying all occurrences of the English phrase to be paraphrased, then finding the corresponding foreign language translations of the phrase, and finally looking at what other English phrases those foreign languages translated back to. Note that "phrase" here simply means (ordered) sequence of words. As an example, Table 3.3 shows the paraphrases that were automatically extracted for an English phrase. Because paraphrases do not always correspond to syntactic boundaries, paraphrase substitution was only applied to text represented as sequences, not to syntactic trees.

In order to assign a ranking to a set of possible paraphrases, Bannard and Callison-Burch used a paraphrase probability $p(e_2 \mid e_1)$, which is defined in terms of two translation model probabilities: $p(f \mid e_1)$, the probability that the original English phrase e_1 translates as a particular phrase f in the other language, and $p(e_2 \mid f)$, the probability that the candidate paraphrase e_2 translates as the foreign language phrase. Since e_1 can translate as multiple foreign language phrases, f is marginalized out:

$$p(e_2 \mid e_1) = \sum_f p(f \mid e_1) p(e_2 \mid f) \tag{3.24}$$

The translation model probabilities can be computed using any standard formulation from phrase-based machine translation. For example, $p(e_2 \mid f)$ can be calculated straightforwardly using maximum likelihood estimation by counting how often the phrases e and f were aligned in the parallel corpus:

$$p(e_2 \mid f) \approx \frac{count(e_2, f)}{\sum_e count(e, f)} \tag{3.25}$$

The definition of the paraphrase probability is extended to include multiple corpora, as follows:

$$p(e_2 \mid e_1) \approx \frac{\sum_{c \in C} \sum_{f \text{ in } c} p(f \mid e_1) p(e_2 \mid f)}{\|C\|} \tag{3.26}$$

where c is a parallel corpus from a set of parallel corpora C. Thus multiple corpora may be used by summing over all paraphrase probabilities calculated from a single corpus

Text: *Clonaid* said, Sunday, *that* the cloned baby, allegedly born to an American woman, *and* her family were going to return *to the* United States *Monday*, but where they live and further details were not released.

Hypothesis: *Clonaid* announced *that* mother *and* daughter would be returning *to the* US on *Monday*.

Substitutions:

the US	→ the United States (score: $\frac{6}{14} \rightarrow \frac{7}{14}$)
returning	→ return (score: $\frac{7}{14} \rightarrow \frac{8}{14}$)
said	→ announced (score: $\frac{8}{14} \rightarrow \frac{9}{14}$)
on Monday	→ Monday (score: $\frac{9}{14} \rightarrow \frac{10}{14}$)

Paraphrased hypothesis: *Clonaid said that* mother *and* daughter would be *return to the United States Monday*.

Figure 3.5: Example of hypothesis paraphrasing by substitution.

(as in Equation 3.24) and normalizing by the number of parallel corpora. The paraphrase probabilities are calculated using the Europarl parallel corpus (Koehn, 2005), which contains parallel corpora for Danish, Dutch, English, French, Finnish, German, Greek, Italian, Portuguese, Spanish and Swedish.

Paraphrase extraction is attempted for every phrase in the hypothesis of up to 8 words. After generating these candidate mappings the hypothesis is iteratively transformed toward the text by substituting in paraphrases. Figure 3.5 exemplifies the paraphrase substitution process. At each iteration, the substitution is made which constitutes the greatest increase of the entailment score. The process stops when no more substitutions can be made which positively affect the entailment score. By example, in Figure 3.5, the paraphrase of the hypothesis is obtained by a number of substitutions. To prevent overgeneration, a word which was introduced in the hypothesis by a paraphrase substitution cannot be substituted itself. In addition, the relative contribution of a substitute to the entailment score equals the contribution of the substituted phrase. For instance, after the substitution *the US* → *the United States* is made, the entailment score has increased from $\frac{6}{14}$ to $\frac{5}{14} + \frac{3}{3} \cdot \frac{2}{14} = \frac{7}{14}$; not $\frac{8}{15}$, because the relative weight of *the US* (i.e. $\cdot\frac{2}{14}$) will be retained after substitution.

In the example of Figure 3.5, paraphrasing caused the length of the LCS to increase from 43% ($\frac{6}{14}$) to 71% ($\frac{10}{14}$). The words in italics are the words which are aligned with the text sentence, i.e. which are part of the longest common subsequence. Figure 3.6

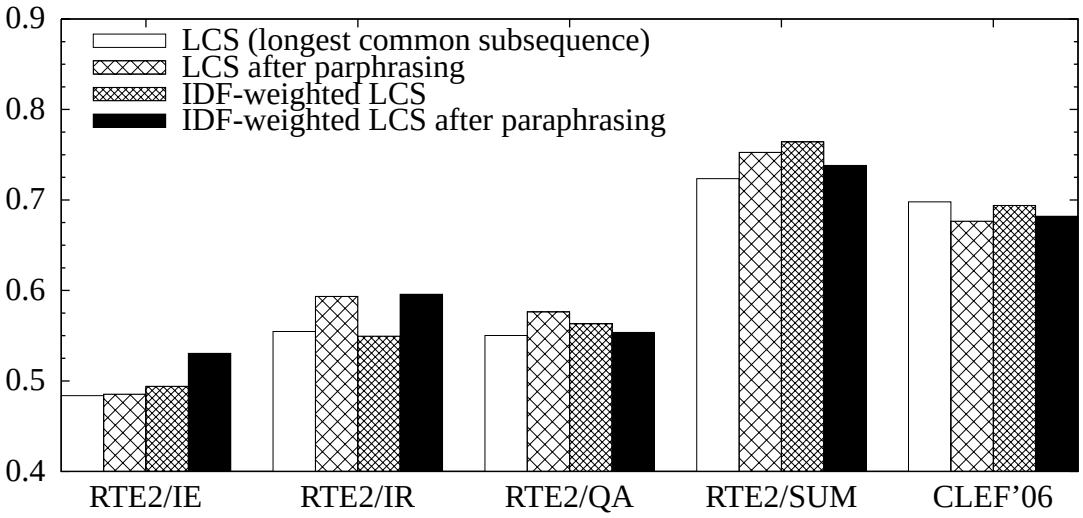

Figure 3.6: The effect of IDF weighting and paraphrasing on G-measure scores.

shows that when combined with LCS, paraphrasing adds considerably to performance if the hypothesis is a very short sentence, as was the case for RTE2/IE and RTE2/IR. On the QA and SUM subsets of RTE2, the corpora with longer sentences, paraphrasing only increases performance if no IDF weighting is used, and on the CLEF-AVE corpus paraphrasing reduced performance regardless of the use of IDF weighting.

Earlier experiments using the same paraphrasing method suggested that paraphrasing performed considerably better than a system which uses plain LCS (Bosma and Callison-Burch, 2007). The explanation for this is that in Bosma and Callison-Burch (2007), no representative training data was available and the threshold value was determined by training on PASCAL-RTE data. As a result, both systems (with and without paraphrasing) underperformed. The results as presented here are based on the average of a large number of cross-validation experiments.

A possible explanation of the poor performance of the paraphrasing system on the RTE2/SUM and CLEF-AVE corpora is that the paraphrase substitution algorithm is more prone to overgeneration on long sentences than on shorter sentences. For instance, Table 3.3 shows that the phrase *bodies of those killed* is a paraphrase of *dead bodies*. If, in a text/hypothesis pair, the hypothesis contained the phrase *dead bodies*, and the text was unrelated but contained the phrase *of those*, the substitution algorithm would substitute *dead bodies* by *bodies of those killed*. This would result in a common subsequence of at least the two words *of those*. This type of overgeneration occurs less frequently on shorter sentences.

3.4 Conclusion

Both the evaluation metrics of *accuracy* and *F-measure* depend on the generality (the ratio) of positive samples in the test set, although this dependence is unmotivated as part of the task of recognizing textual entailment. This poses problems to 'unconventional' corpora such as the machine-constructed corpus of CLEF-AVE, in which the number of positive and the number of negative samples depend on external systems whose behavior is unknown. In such cases, the proposed generality-independent alternatives of *G-accuracy* and *G-measure* (which average recall of positive samples and recall of negative samples) produce more consistent system rankings. Furthermore, G-accuracy and G-measure are shown to be fairly discriminative regardless the corpus composition, while accuracy and F-measure may perform considerably worse if the generality of the corpus is not tailored to the metric. The G-measure metric is used for all experiments with RTE systems in this chapter.

Rather than to focus on a specific approach, performance was compared of an entailment system using various increasingly sophisticated text representations and matching algorithms. We presented the tree-based algorithm earlier in Marsi et al. (2006), and the method for paraphrase matching in Bosma and Callison-Burch (2007). IDF weighting was combined with each text representation. Due to the phrase-based nature of the paraphrasing method, paraphrasing was only applied to the sequence representation, both with and without IDF weighting. Resources required for these methods are limited to a dependency parser (for the tree representation of text), and an aligned bilingual corpus (for paraphrase extraction). In these experiments, Minipar (Lin, 1998) was used for dependency parsing and Europarl (Koehn, 2005) for paraphrase extraction.

Although results show that more sophisticated text representations, IDF weighting and paraphrasing result in a performance gain in some circumstances, no consistent improvements were achieved on all five corpora. The corpora used are the four subsets of PASCAL-RTE 2006 (IE, IR, QA, SUM) and the CLEF 2006 Answer Validation corpus. The strong variation in performance gain on these corpora indicates that the decision whether or not to use these methods for recognizing entailment should depend on the type of corpus. Some but not all differences in performance between corpora can be accounted to physical properties of the text, such as sentence length. This suggests that different information is needed for solving RTE in different corpora (representing different applications). The variety of applications may be a considerable problem

for defining RTE. To my knowledge, no serious evaluative study of the entailment annotation task has been performed to show that the task is in fact well defined.

In order to enable better comparison of RTE systems, it may be worthwhile to evaluate RTE systems against corpora which are annotated with types of inference required (e.g. logical inference, paraphrasing, syntax), rather than or in addition to application-oriented annotation (information extraction, question answering, etc.).

4

Methods for automatic text summarization

This chapter reviews advances and issues in automatic summarization. The chapter contains an overview of applications of summarization, similarities with human abstracting, evaluation methodology, content selection methods, and text revision summarization methods.

In the 1950s, Luhn (1958) invented the "auto-abstract". It was an abstract generating by a computer system capable of condensing text automatically. It would facilitate decisions as to whether or not a document was worth reading. Doing so, it would save time to potential readers and thereby improving access to information. Since 1958, automatic abstracting has not only evolved but also diffused, and many applications appeared, from biography generation to query-focused abstracting. However, a set of core concepts and common methodology applies to a wide range of natural language processing applications we call *summarization*. Since 2001, directions in automatic summarization have been greatly influenced by DUC[1], a evaluation program by the National Institute of Standards and Technology (NIST) which includes a yearly automatic summarization competition.

Mani (2001) defines summarization quite accurately as *taking an information source, extracting content from it, and presenting the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs.* An assumption in this definition is that some content is more important than other content. Also, the definition is quite general and open to interpretation. First, it imposes no restrictions on the type or form of the source, i.e. the information being summarized.

[1] Document Understanding Conference, `http://duc.nist.gov/`

The source may be a single text or multiple documents by different authors, it may be a single paragraph or a multimedia document with a rich mark-up. Similarly, the presentation may vary from a fully grammatical text to a list of keywords characterizing the source. And finally, the user's or application's needs may affect the form of the summary (e.g. use of language and multimedia, summary length) or its content (e.g. a user query).

In other words, the techniques and strategies for summarization depend on the user's needs. Examples of summaries of news articles are shown in Figure 4.1. News headlines or abstracts can be used by potential readers to decide whether the article is worth reading or not, or they can be used to get a quick overview of today's news. Popular search engines also use some form of automatic summarization, e.g. by showing an excerpt of a web page where the query terms are found (see Figure 4.1).

Note that the definition of summarization of Mani deviates from popular use of the term *summarization*[2] by including applications such as bibliography generation and query-based summarization. Nevertheless, I follow this definition as it is accurate for the field of automatic summarization. Summarization can be subdivided into indicative or informative summarization, generation versus cut-and-paste summarization, single-document versus multi-document summarization, generic versus query-based summarization, and extracting versus abstracting.

Single-document versus multi-document. Early text summarization systems aimed at creating a condensed version of a document, containing only the most important information. In some cases it may be useful to extract information from a number of documents, and create a single summary containing the most salient information of those documents. Multi-document input to a summarization system may be preferred if no single document has sufficient coverage, but a set of documents has.

For instance, news articles often provide only the latest news. A search for *Kosovo* on Google news[3] today (May 2007) returns only the latest news of the UN plan for independence of Kosovo. It is not a good source of information about the run-up to the plan. If a user demands a complete overview of the course of events, s/he has to read a number of previous articles as well. Multi-document summarization provides a solution by summarizing a broader range of articles. A special case of multi-document

[2] summary. (n.d.)., *a comprehensive and usually brief abstract, recapitulation, or compendium of previously stated facts or statements*, http://dictionary.reference.com/browse/summary

[3] `http://news.google.com`

Headline: Former Russian leader Yeltsin dead (source: CNN)

Newspaper abstract: RUSSIA'S FIRST PRESIDENT HAS DIED: His legacy remains controversial: he won the first democratic election in Russia's history, and presided over war, political tumult and economic collapse. (source: International Herald Tribune)

Altavista search results:

World reacts to **Yeltsin's death** - CNN.com Former Russian President Boris Yeltsin, who presided over the demise of the ... World reacts to **Yeltsin's death**. Adjust font size: ...
edition.cnn.com/2007/WORLD/europe/04/23/yeltsin.reaction.reut
More pages from edition.cnn.com

Figure 4.1: Various types of summaries.

summarization is timeline summarization, where a summarizer creates a timeline of the course of events on a specific topic (e.g. Swan and Allan, 2000). Another branch of summarization are so-called 'update summaries', where the summarization system has knowledge of the user's prior knowledge, and mentions only what the user does not already know (e.g. Witte et al., 2007).

News articles may be written from different perspectives. An Albanian and a Serbian newspaper may emphasize different aspects of the same event, or may plainly disagree. Opinion mining is used to extract opinionated or comparative statements from multiple documents (Jindal and Liu, 2006; Mullen and Malouf, 2006). Similarly, a journalist writing about car safety measures may focus on the legal side of seat belt regulations, while another emphasizes how car manufacturers deal with this problem. The most relevant information from both perspectives may be presented in a single summary (Dang, 2005).

Finally, news articles may have information overlap, even if they are complementary. This can be exploited by a multi-document summarizer because occurrence of a particular concept in many source documents may be regarded a positive indicator of its importance (c.f. Erkan and Radev, 2004). On the other hand, it also complicates the summarization process, as it should be avoided that the same information is mentioned twice in a summary.

Multi-document summarization introduces several other challenges. Multi-document summarization requires additional preprocessing to cluster documents on the

same topic. A multi-document summarization system faces differences in publication dates of sources, conflicting opinions or perspectives, information overlap and other relations between documents (these have been addressed in more detail in section 2.3, and in chapter 2). Instances of these phenomena may provide useful information for improving a summary, but failing to detect them may negatively affect the quality of the summaries. In sum, a single-document summarizer may exploit expectations about the document and its organization, knowing it to belong to a certain genre and to adhere to corresponding writing conventions.

Indication versus information. The summaries of Luhn (1958) were intended to provide the reader with sufficient information to determine whether reading the full document is worthwhile or not. Condensing a document for this purpose is called *indicative summarization*. Or, as the ANSI 1997 abstracting guidelines (NISO, 1997) put it, an *abstract written in indicative mode describes rather than paraphrases the original document and its contents*. The indicative summary should just describe the purpose or scope of the source document. Another early example of indicative summarization is Edmundson (1969), who extracted key sentences from scientific publications to help researchers or information analysts find or categorize documents. Other types of indicative summarization are keyword extraction (Rau and Jacobs, 1991) and headline generation (e.g. Banko et al., 2000). Keyword extraction involves finding the terms which best describe a particular document. Keyword extraction and headline generation fall within the definition of summarization as it is essentially creating a condensed version of a longer document. The main purpose of keywords and headlines is to facilitate the user's decision to read on or not (Zajic et al., 2002).

In contrast to indicative summaries, informative summaries are intended to *inform* the reader, rather than to serve as a relevance indicator. A summary may serve both an indicative and an informative purpose, and the boundaries between indicative and informative summarization is not always obvious.

The difference between indication and information reflects in the way summaries are evaluated (see also section 4.2.3 on utility-based evaluation). For instance, the indicative summarization experiments of Minel et al. (1997) included a user evaluation of the presence of ideas in the summary which are essential for a relevance assessment. In addition, Minel et al. had an informative task which required the summary to retain also *relations* between ideas, such as cause-result, solutionhood, generalization, list and contrast. Thus, writing an informative summary is more demanding since more criteria must be satisfied.

Later work in summarization has focused mainly on creating summaries form multiple documents, which are presented to the user without pointers to the source documents (c.f. Dang, 2005). The lack of source pointers in the presentation of the summary makes this type of summarization informative by definition.

Generic versus query-based. The fact that a summary is requested implies a potential information need. *Generic summarization* is the type of summarization where this information need is not formulated in any way. In this case, the information need can only be derived from the fact that a summary of the document or document set is requested. The summary should resemble the information which the authors of the original documents deemed most important. Phrased differently, an author writes a document in order to answer a (possibly unspoken) question. Generic summarization is answering this question in a more concise form than the author originally did. In query-based summarization, an expressed information need is available to the system in the form of a query. The ideal answer would match the intersection between the author's implicit question and the user's query.

But what is a query? The aim of question answering systems is also to answer a query. Apart from the fact that query-based summarization systems usually provide more verbose answers than question answering systems, the term *query* in query-based summarization is less strict than a *question* in question answering. In the TREC question answering evaluation campaign, queries are from a closed set of question types. In TREC 2003, the questions were factoid questions (questions for years, names, etc.), list questions (e.g. *which cities have Crip gangs?*) or definition questions (Voorhees, 2003).

In query-based summarization, *the query* may pertain to the type of summary required or to the information requested in the summary. For instance, in DUC 2005, the query included a user profile which was used to express the demand for a *specific* or *general* summary (Dang, 2005). The query may also take the form of a question (e.g. *What devices and procedures have been implemented to improve automobile safety?*) or an assignment in imperative form (e.g. *Describe developments in the movement for the independence of Quebec from Canada.*). Mani and Bloedorn (1999) used arbitrary keywords as queries to a query-based summarization system. One of the tasks in DUC 2004 was to produce query-based summaries in response to queries of the form, *who is X*, where X is the name of a person or a group of people. Query-based summarization became the main task of DUC in 2005, where the query was phrased an assignment comprising one or more sentences phrased as questions or imperatives.

Nevertheless, there is no clear boundary between question answering and query-based summarization, as question answering is moving toward less focused questions and longer answers. For instance, TREC 2004 introduced *other* questions, to which the system has to respond with 'relevant remarks' about a general topic, such as *Crip gangs*.

Extracting versus abstracting. A human summarizer typically does not create a summary by copying (extracting) sentences verbatim from a source into the summary. Rather, s/he would abstract from the source, and write a summary containing the most relevant information in different wording. Kupiec et al. (1995) analyzed the manual summary creation process by comparing sentences of manually created summaries with sentences of the source text. They identified the following summary sentence types, based on their relation with the source document.

Type I: *direct sentence matches (79%)* – sentences which are copied verbatim or with minor modifications from the source.

Type II: *incomplete single sentences (4%)* – sentences which can be matched with a single source sentence, but they do not fulfill the constraints of direct sentence matches, because the content of the summary sentence is a subset or a superset of the content of the source sentence.

Type III: *direct joins (3%)* – sentences which are created by combining two or more sentences from the source with minor modifications.

Type IV: *incomplete joins (5%)* – sentences which can be match with two or more source sentences, but they do not fulfill the constraints of direct joins for the same reasons as that type II is incomplete.

Type V: *unmatchable sentences (9%)* – sentences which are created from a general understanding of the text, rather than one or more source sentences.

The term *extracting* is sometimes defined as creating summaries by copying parts of the source text (Radev et al., 2002a). The extracted text units may be paragraphs (e.g. Mitra et al., 1997), sentences (e.g. Goldstein et al., 1999) or even smaller (Witbrock and Mittal, 1999). In contrast, *abstracting* refers to creating summaries by paraphrasing the source in a concise manner. Many state-of-the-art summarization systems perform some form of post-processing on summaries, if only for esthetic reasons. This may

include for instance normalizing punctuation or removing conjuctive adjuncts (such as *but*). By definition of Radev et al., these summarization systems would be characterized as abstractive.

As an alternative to the extract/abstract distinction of Radev et al., summarization may be seen as a two-stage process, consisting of extracting and abstracting. Extracting is the process of *selecting* passages for inclusion in the summary. Abstraction is the further processing of those passages into a summary, possibly including sentence compression, sentence fusion, etc. I use the terms *extracting* and *content selection* synonymously.

The table of Kupiec et al. shows that most summary sentences (i.e. 79 percent) were used with only minor modifications, suggesting that an extraction system with no or a simplistic abstraction strategy may be reasonably successful. However, care should be taken when interpreting these numbers, as results of empirical studies such as these may be heavily dependent on the application and the corpus used. For instance, a similar study by Jing and McKeown (1999) resulted in substantially lower numbers: 42 percent for direct matches and another 36 percent of the (generic single-document) summary sentences matched 2–3 source sentences.

The majority of summarization systems use only shallow understanding of text, and is therefore restricted to using type I sentences to construct a summary. *Compression* techniques can be used to create type II sentences. Text can be compressed on a sentence level (Jing and McKeown, 1999; Knight and Marcu, 2000, e.g.) by reducing the sentence size while retaining grammaticality and the most important pieces of information. Witbrock and Mittal (1999) applied compression on a document level by using bigram-based language generation. Marsi and Krahmer (2005) use sentence fusion to aggregate information from multiple sentences into a single sentence. These techniques enable construction of sentences of type III and IV.

True abstracting however, involves rephrasing and possibly distilling information which is left implicit in the source text. For instance, newspaper articles concerning the murder of the politician Pim Fortuyn contained many quotations of politicians and statements of people's opinions. A human summarizer might generalize this in a type V sentence by stating: *Dutch as well as international politicians have expressed their grief and disbelief* (van Halteren and Teufel, 2003).

This type of summarization requires a deep understanding of the source and the ability to derive new information by means of inferencing or generalization. Since type V sentences constitute only 9 percent of the summary sentences, it is unclear how not being able to create these sentences would affect the quality of the summaries.

Kupiec et al. do not answer this question. Nevertheless, forming type V sentences are considered an important next step in automatic summarization.

Generation versus cut-and-paste. The quality of a summary may benefit from using domain knowledge. Most summarization systems use domain knowledge in one way or another. The features they use to determine how relevant information is to a summary (user) may be tailored to specific genres. For instance, named entities (i.e. names of persons, monetary units, expressions of time, etc) may be a key feature in newspaper summarization, while they are less relevant when summarizing medical encyclopedic text. Some summarization systems exploit the fact that some genres or sources tend to start with an abstract. Also, a system may be tuned to a particular domain by training on a particular data set. Newspaper articles are a popular domain of text summarization systems, most likely due to the abundance of resources and the proliferation of information, and they have been used each year to date in the main task of DUC.

Cut-and-paste systems extract information, possibly apply inferencing or generalization, and then revise where needed. *Generation-based* summarization systems are a class of summarization systems which make extensive use of domain knowledge. Generation-based systems generate summaries from structured data rather than directly from text. Possibly, the data source is derived from text by means of information extraction. Information extraction is the process of populating a structured information source (e.g. a database or a template) from an unstructured information source (such as free text). From the resulting structured information, natural language generation can be used to generate fluent text (Theune et al., 2001).

By example, suppose we are interested in company mergers in which a specific company is involved. We might search a database of newspaper text for articles describing mergers. Beforehand, we know that two parties and a date are related to any merger, and this information is likely to be found in an article describing such an event. We may add information (such as a price) as needed. The information may be extracted from the article using information extraction techniques by instantiating a template with three slots: $(company_1, company_2, date)$, and then scan the article for a date and the names of the companies in relation to the takeover. If a sufficiently large number of newspapers is searched, we (hopefully) have a complete overview of all mergers which interest the user, and we can present this as a timeline or as fluent text by means of language generation.

High quality summaries are possible when using generation, but the type of summaries such a summarizer creates is very specific to a particular domain. The system must be programmed with prior knowledge about its domain, including the type of information the user may be interested in. Summarization by generation uses techniques very different from cut-and-paste systems, where no intermediate structural representation is used. The focus of this thesis is cut-and-paste summarization systems.

Other applications of summarization. The list of possible summarization applications is endless. Many applications are based on or use similar techniques as applications mentioned earlier, but they may pose additional challenges. For instance, in multilingual summarization, the source language may be different from the target language (Evans, 2006). A similarity of spoken dialog summarization (Zechner, 2002) and multi-document summarization is the involvement of multiple authors, but spoken dialog summarization faces other challenges such as (a) the typical length of an utterance, (b) the relation between utterances, (c) the type of language used, and (d) errors in speech recognition. Other types of summarization include non-text media, such as diagram summarization (Futrelle, 1999).

4.1 Human summarization

Since human abstracting is often taken as an example and an ideal standard for automatic summarization, it is worthwhile to review similarities and differences between the two tasks. Services of professional abstractors include the following (Mani, 2001):

- providing abstracts for documents that lack abstracts (e.g. news articles);
- editing author-supplied (or machine-generated) abstracts to conform to guidelines or quality criteria;
- tailoring abstracts to different audiences (user-focused abstracts);
- translating abstracts to different languages;

Considering the automatic summarization applications mentioned earlier, there is large overlap between automatic summarization and professional abstracting applications. The contribution of human abstracting to automatic summarization and its applications is three-fold. First, professional abstracting has led to abstracting guidelines which may be relevant for automatic summarization as well. Most notably, NISO

(1997) describes a set of guidelines for content and style, e.g. desired abstract length, the use of language constructs, etc. Second, human abstracts have proven to be a valuable source for evaluation of automatic abstracts. Third, research in human abstracting processes and strategies results in knowledge which may be applicable automatic summarization.

4.1.1 The process

Studies in human abstracting distinguish a number of stages in the abstracting process (e.g. Cremmins, 1982; Pinto-Molina, 1995; Endres-Niggemeyer, 1998). Although very similar, the exact definition of these stages differs slightly as described by different authors. The model of Pinto-Molina (1995) describes abstracting in three stages:

- reading and understanding;
- interpretation and selection of relevant information;
- synthesis: producing a summary.

The abstractor's expectations play a central role in the first phase (Endres-Niggemeyer, 1998). This pertains also to the structure of the documents with respect to content and layout, and may be genre-specific. For instance, a scientific article most likely starts with an introduction and ends with a conclusion, while newspaper articles have a less strict organization.

In the second stage, the abstractor uses a mental representation of the source text to decide upon the main points of the author and which information is most relevant. The work of Endres-Niggemeyer is based on earlier research on discourse modeling (Kintsch and van Dijk, 1978; Mann and Thompson, 1988).

The final abstracting stage comprises summary production from relevant material from the source. A frequently used technique is to extract sentences or parts of sentences from the source, and applying minor revisions to create a coherent whole (c.f. Kupiec et al., 1995; Jing and McKeown, 1999).

4.1.2 The strategies

Endres-Niggemeyer (1998, cited in Mani (2001)) discovered a number of strategies that were used by all expert abstractors they examined. For instance, they never read

the whole document, but they relied on surface characteristics such as cue phrases, layout and headings to find relevant passages and how they relate.

During the *interpretation and selection* stage, human abstractors typically do not create content from scratch. Rather, they use a *cut-and-paste* strategy to 'cut' passages from the source and 'paste' them into the summary. A reason for this is that they are usually not an expert on the subject matter.

Revision techniques were investigated further by Jing and McKeown (1999), who analyzed the relation between summary and source sentences. Jing and McKeown found the following revision techniques to be used frequently.

Sentence reduction. Less important fractions of a sentence are omitted in the summary, such as attributions (e.g. *X said ...*), person names, adjectives or propositional phrases.

Sentence fusion. Two sentences may be merged into one by adding a connective.

Syntactic transformation. Constituents may be moved, or a passive form may be replaced by an active form or vice versa.

Lexical paraphrasing. A phrase such as *point out* is replaced by *note*, or the phrase *fits squarely* is replaced by *hits the head on the nail*.

Generalization and specification. Irrelevant details are replaced by generalizations, or if an element of the summary sentence becomes unclear due to lack of context (e.g. dangling anaphora), the abstractor may choose to add more detailed information.

Reordering. A summary does not necessarily present its content in the same order as the source document.

All of this has been addressed in Natural Language Processing and most of the techniques may be applied in automatic summarization. Automation of these operations is discussed in more detail in section 4.4.

4.2 What is a good summary?

Query-based summarization is a relatively new application of natural language processing, without established evaluation procedures. For evaluation, we may benefit

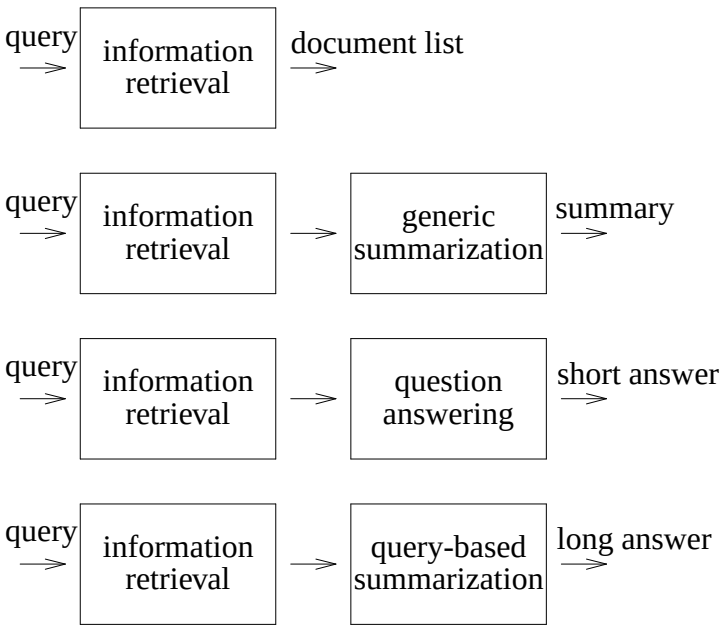

Figure 4.2: Generic summarization, query-based summarization and question answering as extensions to information retrieval.

from relating query-based summarization to related fields. Query-based summarization is similar to (generic) summarization (it aims to produce a concise version of the original) and to question answering (it aims to satisfy an expressed information need in the form of a query). Both question answering and (query-based and generic) summarization systems typically take input from an information retrieval system whose task is to filter out irrelevant documents, or take for granted that a small number of (potentially) relevant documents is available (Figure 4.2). It should be stressed that variations of summarization or question answering may be used in different contexts. Nevertheless, the fact that they can be used as substitutes makes it worthwhile to examine their differences and (dis)advantages.

Evaluation procedures in question answering has been shaped largely by TREC (Text Retrieval Conference), which has a question answering evaluation track since 1999 (Voorhees, 2001). The goal of the track was to 'foster research on systems that retrieve answers rather than documents' in response to a question. However similar the goals, their evaluation methods are hardly useful for query-based summarization, as the main criterion of performance in the question answering track is how well a system succeeds in retrieving 'correct' answers, as opposed to 'incorrect'. In contrast, the goal of query-based summarization is to produce a 'useful' answer rather than a correct one.

One could argue that not every correct answer is useful, and not every answer which is incorrect in a strict sense is useless. Especially for longer documents it is difficult to decide between the predicates 'correct' or 'incorrect'.

In contrast to question answering, summarization has a history of evaluating on usefulness. In this regard, query-based summarization is viewed as summarization guided by a query. The next problem is that *usefulness* in itself is a vague concept. What is useful in one situation may be less useful (or useless) in another. Nevertheless, a number of general quality considerations apply to a greater or lesser extent Alterman (1991):

1. a summary should reduce the workload for the interpreter/understander over the text;

2. a summary should maintain coherence;

3. a summary should maintain coverage.

Although the quality aspects of Alterman are not entirely disjunct, most evaluation metrics for summarization focus on one of these three quality aspects. The first requirement is strongly connected to the task at hand, and comes down to the effect on the user while fulfilling this task. The formulation of Alterman (1991) is too strict for our purposes, as it disregards qualities such as user appreciation. Therefore, I will use the more general *effectiveness* to a user or an application rather than *workload reduction*. Usefulness can be evaluated by *utility-based* or *extrinsic* evaluation, i.e. evaluation in context, as opposed to *intrinsic* evaluation, which refers to measuring qualities of a summary as such (Spärck Jones and Galliers, 1996). Requirements 2 and 3 are intrinsic. Coherence (quality 2) can be generalized to *linguistic quality*, while coverage (quality 3) is evaluated by means of *content-based* evaluation metrics, usually by measuring information overlap between automatically generated summaries and hand-crafted "gold-standard" summaries. The remainder of this section discusses metrics for content-based evaluation (section 4.2.1), linguistic quality (section 4.2.2) and utility-based evaluation (section 4.2.3).

4.2.1 Content-based evaluation

Perhaps the most difficult aspect of a summary to evaluate is how informative it is. A straight-forward way to do this is to ask human assessors to simply rate informativeness

of a summary as a whole. For instance, Mani et al. (1999) performed a QA-like evaluation where users judged correctness of a query-based summary. Similarly, DUC evaluations included responsiveness assessments of query-based summaries (Dang, 2005, see below). In generic summarization as such, there is no explicit information need. In order allow people to judge the informativeness of generic summaries, Brandow et al. (1995) used an Information Retrieval setting to measure responsiveness to a query of summaries of retrieved documents.

Another approach to summary evaluation is to regard summaries as a set of propositions or information units. Assuming that information units can be extracted from text, and that the relevance of each information unit can be determined, informativeness of a generated summary can be measured in terms of precision and recall of relevant information. *Precision* would be the amount of relevant information in the summary proportional to the size of the summary, and *recall* is the amount of relevant information in the summary proportional to the amount of relevant information that should be in the summary.

But which information is relevant? If each sentence in the source text is (manually) rated for importance, an extractive summary can be evaluated by calculating the sum of the ratings of its sentences (Radev et al., 2000; Wolf and Gibson, 2004). An indirect method of rating sentences was used by (Otterbacher et al., 2005), who asked annotators to come up with a list of questions to information key to understanding the story. Then, each sentence in the source which answered those questions was marked as relevant.

Alternatively, relevant information can be found by having professional summarizers make reference summaries under the assumption that human made summaries are 'ideal' summaries. However, inconsistent results are caused by the fact that summaries created by professional summarizers vary greatly in content, while there is no reason to believe that one man-made reference summary is better than another. Rath et al. (1961, cited in Mani (2001)) showed that human extractors often disagreed on which sentences should be in the extract. Moreover, intra-extractor agreement was only 55 percent if the same extractors were given the same assignment eight weeks later. On the other hand, it appears that the longer the extract is, the less extractors agree on its content (Jing et al., 1998). In other words, they agree more on what is the most essential information in a text than on peripheral information which may also be relevant. The presence of variation in summaries suggests that the task is underspecified. Differences among human summaries may be caused by different interpretations of the query (in the case of query-based summarization) or the source text (Donaway et al.,

2000). This may affect a summary's focus and granularity. A reason for DUC to move to query-based summarization was that it enables the use of tools to better control the task.

However, it appears very difficult to eliminate natural variability between human abstractors. To make this acceptable, summaries may be evaluated against multiple reference summaries of the same document set. The idea behind this is that each piece of information has some degree of significance. If the information is in each reference summary, it is very significant. If it is in only one of the reference summaries, it is marginally significant.

Another unsolved problem is how to compare information in summaries. An extractive summarization system can be evaluated by imposing this restriction to extract also to human summarizers: their summary must consist of sentences which also appear in the source text, or the source text is annotated with sentence relevance (Goldstein et al., 1999). Goldstein and Carbonell (1996) report 68 percent agreement for relevance assessments among three annotators. Apart from the obvious limitation that it does not apply to evaluation of abstracts, the main objection to this method is that checks for sentence identity rather than content: there would not be a match if a sentence of a reference summary is not identical but semantically equivalent to a peer summary sentence (Donaway et al., 2000).

In sum, the content of a summary can be evaluated for relevance in at least two ways: (1) by directly judging the query-relevance of a summary as a whole (in case of a query-based summary), or (2) by first determining which information *should* be in the summary, and then measuring how well the candidate summary resembles this. The second breaks down into four questions (Hovy et al., 2005b):

1. What is the unit of content?

2. Which content is important?

3. When do two content units match?

4. How can a score be derived from these matches?

The remainder of this section describes a number of evaluation metrics which are based on different answers to these questions.

4.2.1.1 Fidelity to source

Information retrieval and search methods aim at finding relevant documents and determining if a document is relevant. Indicative summaries can be part of a search process, but since information retrieval (IR) has received much more attention than indicative summarization, it makes sense to investigate how IR can be used for summary evaluation. If the content a document is relevant to the user, then its summary should also be relevant, and vice versa, as the summary should catch the main points of the document. More specifically, Donaway et al. (2000) argued that a search engine can be used to index a document and its (generic) summary, and the similarity of the indices is a measure of the summary's quality. Term frequency vectors have been successful in information retrieval as a model of document content. Sakai and Spärck Jones (2001) confirmed that a vector representation of a summary is as good an indicator of relevance as the full text index. Thus, a (generic) summary can be evaluated without the use of reference summaries, just by comparing it to the source text.

In order to evaluate the quality of summaries, Donaway et al. (2000) used cosine similarity (c.f. eq. 2.1, section 4.3.4.1) to compare vector representations of summary S and source text T.

$$cosim(S,T) = \frac{\sum_{i=1}^{N} s_i \cdot t_i}{\sqrt{\sum_{i=1}^{N} s_i^2} \cdot \sqrt{\sum_{i=1}^{N} t_i^2}} \tag{4.1}$$

where $N = \|S\| = \|T\|$. The values of s_i and t_i are the number of occurrences of term i in the summary and the source respectively.

Because the summary vector contained relatively few terms which may be synonyms of terms in the document, Donaway et al. applied LSA (Latent Semantic Analysis) transformations to the vector representations before measuring cosine similarity (Deerwester et al., 1990). LSA can be used for inferencing on a lexical level (c.f. Landauer et al., 1998). For instance, if the terms 'bank' and 'financial institution' are used (nearly) interchangibly, LSA would transform the vector representation of the summary in a way that it is similar to the vector representation of the source text, even if the source text uses mostly the term 'bank' while the summary uses 'financial institution'.

Donaway et al. claimed that similarity between summary and source mimics human judgment better than earlier (sentence identity-based) methods. A limitation of their experiments is that the data which is used for summarization (the source text)

is also used for testing. In fact, many summarization systems may use (and do use, see e.g. Edmundson, 1969) the same or similar information (e.g. term frequency) to determine importance of information. No external validation is performed to avoid circularity during evaluation: in the worst case, the summarization system evaluates itself. Techniques similar to that of Donaway et al. were employed by Radev et al. (2003).

Alternatively, Saggion and Lapalme (2000) asked users to describe a summary in terms of keywords. The idea behind this is that a summary can be described best using the keywords used to describe the full text. In order to prevent polluting the evaluation, the summarization system must not use the list of keywords in the summarization process.

Finally, experiments of Minel et al. (1997) included a user evaluation of the (lack of) presence of essential ideas and parasitic concepts in the summary which are important for a relevance assessment of the source. In addition, the informative (but not the indicative) task required the summary to retain also *relations* between concepts, such as cause-result, solutionhood, generalization, list and contrast.

4.2.1.2 SEE and responsiveness

For query-based summaries in DUC (Dang, 2005), *responsiveness* was used (alongside other metrics) as a measure of quality. Human subjects were asked to judge on a 5-point scale how well a summary as a whole provided the information requested in the query. Since a query is used for the relevance test, this method only applies to query-based summarization. Query-based summarization was introduced in DUC in 2003 and it is the main task since 2005.

Most state-of-the-art evaluation metrics require the use of hand-crafted *reference summaries* to which *candidate summaries* are compared. In DUC 2001 to 2004, summaries were evaluated by human judges who had to annotate information overlap between a system summary and a reference summary by marking pairs of sentences which shared information. They also assigned a rating to the degree of recall in the system summary, i.e. *all*, *most*, *some* or *hardly any*. In these experiments, information-carrying units of text are sentences. A tool called *Summarization Evaluation Environment* (SEE) was used to carry out the annotations. As Lin and Hovy (2002b) pointed out, this rating is unreliable because human judgment is involved, and in many instances they assigned a different rating to the same phrase compared to the same reference summary, when produced by a different system. Lin and Hovy (2002b) did not

investigate whether this variation in rating might be caused by a change in the actual content of the phrase due to the different context in which it appeared.

4.2.1.3 Lexical similarity

Saggion et al. (2002) proposed three similarity metrics for evaluation of summaries using reference summaries: cosine similarity, n-gram overlap, and longest common subsequence (LCS). While Saggion et al. measured *similarity* between a summary and a reference summary, Lin (2004) used n-grams and LCS to measure *recall* of a peer summary with respect to a reference summary: a summary achieves higher scores if it contains more information which is also in the reference summary. No penalty is given for irrelevant information. Instead, a brevity bonus is given to shorter summaries, or a restriction is imposed to the length of the summary.

Cosine similarity. Saggion et al. used cosine similarity to compare summaries. Saggion et al. employed a simpler version of cosine similarity than Donaway et al. without the use of LSA, but they experimented with different weighting schemes. The weighting schemes they used were term presence/absence and $tf \cdot idf$. In the first weighting scheme, a vector element for a term contained a 1 if the term occurred in the document, and 0 otherwise. In the $tf \cdot idf$ weighting scheme, the term frequency is multiplied by the inverse document frequency:

$$tf \cdot idf = tf \cdot log \frac{\|D\|}{\|\{d \in D \mid t \in d\}\|} \tag{4.2}$$

where *tf* is the frequency of a term, the numerator of the fraction is the number of documents in the corpus, and the denominator is the number of documents in which the term appears.

N-gram recall. The second similarity metric proposed by Saggion et al. to compare summaries is n-gram overlap, with $n = 1$ or $n = 2$. An n-gram is a substring of a sequence (in this case a sequence of words) with length n. The following example exhaustively shows 1-grams (unigrams), 2-grams (bigrams) and 3-grams of a sentence.

S: Spain prosecutes Rwandan leaders

1-grams: Spain; *prosecutes*; *Rwandan*; *leaders*.

2-grams: Spain prosecutes; *prosecutes Rwandan*; *Rwandan leaders*.

3-grams: Spain prosecutes Rwandan; *prosecutes Rwandan leaders*.

N-gram overlap between summaries was calculated as the size of the intersection between the sets of n-grams in both documents, divided by the total number of n-grams in both documents:

$$overlap_{saggion}(S,R) = \frac{\|S \cap R\|}{\|S \cup R\|} \tag{4.3}$$

where S and R are the sets of n-grams in the summaries under comparison.

Alternatively, Lin and Hovy (2003) defined n-gram overlap in terms of precision and recall, by example of the Bleu metric for evaluating machine translation systems (Papineni et al., 2001). Because BLEU is precision-oriented and DUC summary evaluations were recall-oriented, BLEU could not be used directly. Lin and Hovy proposed the following formula for measuring n-gram recall of a peer summary S with respect to reference summary R.

$$overlap_{single}(S,R) = \frac{\|S \cap R\|}{\|R\|} \tag{4.4}$$

In essence, formula 4.4 calculates the number of n-grams shared by a peer summary and a reference summary, divided by the total number of n-grams in the reference summary. When using a collection C of multiple reference summaries, the formula can be written as follows:

$$overlap_{multi}(S,C) = \frac{\sum_{R \in C} \|S \cap R\|}{\sum_{R \in C} \|R\|} \tag{4.5}$$

Because this formula calculates recall but not precision, it would give a bonus to longer summaries. To compensate for this, a brevity bonus *BB* is awarded to favor brief summaries. Furthermore, Lin and Hovy use n-grams with different values for n and average the result. The n-gram score with $n = i..j$ is calculated as follows:

$$N_{i,j}(S,C) = BB \cdot exp\left(\sum_{n=i}^{j} \frac{1}{j-i+1} logoverlap_n(S,C)\right) \tag{4.6}$$

where $N_{i,j}(S,C)$ is the average n-gram overlap of S and C with n-gram parameters i and j; $overlap_n(S,C)$ corresponds to $overlap_{multi}(S,C)$ with n-grams as content units.

R	police killed the gunman
12A	police kill the gunman
12B	the gunman kill police
12C	the gunman police killed

Figure 4.3: An example of paraphrases of phrase R.

Table 4.1: Comparison of different recall metrics.

	12A	12B	12C
unigram	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{4}{4}$
bigram	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
LCS	$\frac{3}{4}$	$\frac{2}{4}$	$\frac{2}{4}$
skip bigram	$\frac{3}{6}$	$\frac{1}{6}$	$\frac{2}{6}$

Lin and Hovy (2003) report good correlation between human assessments and the automatic score $N(1,4)$. N-gram recall is a widely accepted metric for content-based summary evaluation and it has been implemented under the name *Rouge-N* in the Rouge toolkit (Lin, 2004).

N-gram overlap with $n = 1$ behaves similarly to cosine similarity. But for $n > 1$, n-gram overlap is a more strict matching algorithm than cosine similarity, because it is sensitive to the ordering of words in a sentence. For instance, Given the paraphrases in Figure 4.3, Table 4.1 reads a unigram recall of $\frac{4}{4}$ for 12C with respect to R, while 12C contains only one of the four bigrams in R (i.e. “the gunman”).

Lin (2004) implemented two extensions to Rouge-N: skip-bigram co-occurrence (Rouge-S) and skip-bigram co-occurrence averaged with unigram co-occurrence (Rouge-SU). The way Rouge-S is calculated is identical to Rouge-2, except that skip bigrams are defined as *subsequences* rather than the regular definition of bigrams as *substrings*. In contrast to substrings, subsequences may be discontinuous. For instance, the sentence “police killed the gunman” contains six skip bigrams (among which are three bigrams), e.g. “killed gunman” is a skip bigram but not a bigram. Because Rouge-S in this form may be too lenient, Lin proposed an intermediate solution by imposing the restriction to skip bigrams that the ‘gap’ between both terms in the bigram may not exceed a specific size. For instance, if the maximum skip size is 1, “police gunman” is not counted as a skip bigram because both terms are separated by more than one term.

Rouge-SU is a weighted average between Rouge-S and Rouge-1. It can be calculated by adding an end-of-sentence marker to the end of each sentence, and measuring Rouge-S.

Longest common subsequence. The longest common subsequence (LCS) of two passages is the longest sequence of words which is a subsequence of both passages. A sequence A is said to be a subsequence of B is A can be obtained from B by deleting elements, in this case words. For instance, the sequence *the attacked the* would be the longest common subsequence of the following sentences:

13A *the* terrorist *attacked the* president
13B *the* president *attacked the* terrorist

In the above example, the vector representations of 13A and 13B would be identical, while the length of the longest common subsequence would be only $\frac{3}{5}$ of the sentence length. Saggion et al. (2002) measured normalized LCS on a sentence level by comparing each pair of sentences of two documents. This measure was implemented in the Mead summarization framework (Radev et al., 2002b) The length of the LCS between summaries S and R can be computed using the following formula (Crochemore and Rytter, 1994):

$$lcs(S,R) = \frac{1}{2}(\|S\| + \|R\| - edit_{di}(S,R)) \tag{4.7}$$

where $\|S\|$ and $\|R\|$ is the length of the summaries S and R respectively, and $edit_{di}(S,R)$ is the minimum number of deletion and insertion operations required to transform S into R.

The normalized pairwise LCS similarity of two summaries S and R is calculated as follows (Radev et al., 2002b):

$$LCS_{mead} = \frac{\sum_{s_i \in S} max_{t_j \in T} LCS(s_i,t_j) + \sum_{t_i \in T} max_{s_j \in S} LCS(t_i,s_j)}{\sum_{s_i \in S} \|s_i\| + \sum_{t_i \in T} \|t_i\|} \tag{4.8}$$

Similarly to the n-gram-based metric of Saggion et al., their LCS measure is symmetric in the sense that $LCS_{mead}(A,B) = LCS_{mead}(B,A)$. And similarly to the way Lin (2004) adapted the n-gram measure of Saggion et al., he reformulated their LCS measure information recall and precision separately. It has been claimed that the LCS measure in MEAD is in fact equivalent to the F-measure in Rouge-L (Lin, 2004), but

since MEAD is symmetric while F-measure is not, this is obviously not the case. Lin defines recall and precision of a candidate summary sentence with respect to a reference summary sentence as follows:

$$Recall(S,R) = \frac{lcs(R,S)}{\|R\|} \tag{4.9}$$

$$Precision(S,R) = \frac{lcs(R,S)}{\|S\|} \tag{4.10}$$

Formulas 4.9 and 4.10 calculate LCS scores on a per-sentence basis. In order to calculate LCS scores on a summary level, Lin introduce the term *union LCS* ($LCS_{\cup}(r,S)$ between a reference summary sentence r and a candidate summary S. The union LCS of r and S is size of the set of terms which appear in the longest common subsequence of r and any of the sentences in S. Now, summary-level LCS is defined as follows:

$$Recall(S,R) = \frac{\sum_{r \in R} lcs_{\cup}(r,S)}{\sum_{r \in R} \|r\|} \tag{4.11}$$

$$Precision(S,R) = \frac{\sum_{r \in R} lcs_{\cup}(r,S)}{\sum_{s \in S} \|s\|} \tag{4.12}$$

It should be noted that these formulas are slightly awkward, which can best be illustrated by an example. Suppose a reference summary R consists of five sentences each of which consist of only the word v. A candidate summary S consists of five sentences of which one sentence consists of the word v, and the remaining four sentences contain only the word w. We would expect the recall to be high (i.e. close to 1) because all information in the reference is also in the candidate. On the other hand, precision should be low (i.e. close to 0) because four out of five sentences consist of information which is not in the reference. When calculating recall, we find (as expected) that the recall value for summary-level LCS is 1, as $LCS_{\cup}(r,S) = 1$ for each $r \in R$. However, precision also appears to be 1 because we calculate the same $LCS_{\cup}$ values, except we divide by the size of the candidate instead of the reference. This is caused by the fact that the denominator is asymmetric.

A feature of LCS is that it takes discontinuous sequences into consideration. Lenient matching to some extent is desired for text similarity algorithms, but a consequence of using LCS is that two sequences of the same length receive the same scores, even if one of them contains gaps while the other is continuous. Intuitively, the continuous sequence should be preferred. Lin (2004) implemented an extension to Rouge-L

which discriminates between sequences with discontinuous subsequences of different lengths.

Rouge performance. Rouge was inspired by Bleu, a metric for n-gram precision widely used in machine translation evaluations. Bleu has been evaluated for performance in a summarization setting (Pastra and Saggion, 2003). Pastra and Saggion compared Bleu scores with human assessments of information overlap on a per-sentence basis, similar to DUC 2001 evaluations. Pastra and Saggion are optimistic but they also mention the limitation of n-grams not to capture a number of linguistic phenomena such as paraphrasing. As a result, Bleu may be useful only when multiple reference summaries are used.

The evaluations of Pastra and Saggion have been repeated for Rouge, a metric for n-gram recall and other statistical similarity measures (Lin, 2004; Lin and Hovy, 2003). Lin evaluated their evaluation algorithms on a number of corpora for different summarization tasks: generic single-document summaries, very short summaries (or headlines), and multi-document summaries. Not surprisingly, no measure outperformed all other measures for all tasks. In general, multi-document summaries appeared to be more difficult to evaluate than single-document summaries. Although various Rouge measures have been used for query-based summarization in DUC, I am not aware of any evaluation of Rouge performance on query-based summarization.

An important limitation of these performance evaluations of Rouge is that the ground truth was human assessments of information overlap between sentences. As mentioned earlier, these human assessment scores themselves appeared to be unreliable (Lin and Hovy, 2002b).

4.2.1.4 Factoids and Pyramids

Statistical similarity metrics such as used in Rouge address unreliability of human judgments of content overlap between sentences, but they are very rigid similarity metrics. Synonymy and paraphrasing are left largely unattended. Reformulation to some degree is allowed, but this inevitably introduces false matches as well. This dilemma between strict and lenient matching is reflected by the large number of algorithms implemented in Rouge (c.f. Lin, 2004). Nenkova and Passonneau (2004) argue that similarity metrics are less suitable for summarization evaluation than for evaluation of machine translations (c.f. Papineni et al., 2001), because what is a “good” and what is

a “bad” translation is better defined than what should be in a summary. As a result, variation in summarized text may be greater than variation in translated text.

Factoids. Rather than string matching, the *factoid* evaluation method uses information overlap for evaluation of summaries (van Halteren and Teufel, 2003; Teufel and van Halteren, 2004). The unit of information is called a factoid. A sentence contains at least one but potentially any number of factoids. For instance, consider the sentence: “The police have arrested a Dutch white man.” Teufel and van Halteren identify five factoids in this sentence: (a) a suspect was arrested; (b) the police did the arresting; (c) the suspect is white (d) the suspect is Dutch; (e) the suspect is male. The factoid method evaluates a summary in four phases:

1. factoids are identified in a set of reference summaries by human annotators;
2. each of the factoids is assigned a weight algorithmically;
3. occurrence of these factoids in candidate summaries are marked by human annotators;
4. a score is assigned to candidate summaries, based on the number of factoid occurrences and the weights of the factoids.

During phase 1, reference summaries are represented by a number of factoids. The set of factoids is *complete* and *disjunct*: the factoids cover all of the information in the set of reference summaries, and information overlap between factoids is not allowed. The factoid method is based on the assumption that reference summaries share common (but possibly differently phrased) factoids with each other and with candidate summaries. Factoids can be used to formalize information overlap between summaries, but partial overlap is not accepted: a factoid is either in a summary or it is not. If a factoid partially occurs in a summary, the factoid is divided into multiple factoids. For instance, if one summary contains the words “was killed” while another reads “was shot dead”, three factoids are identified: (a) there was an attack; (b) the victim died; (c) a gun was used. Splitting up factoids ad hoc implies that factoids may vary in granularity.

When the set of factoids is determined, each factoid is weighted (phase 2) depending on its number of occurrences in reference summaries and the position of the factoids in the reference summaries. This reflects the idea that information in many

reference summaries is important information, and that the author may indicate importance of information by mentioning it near the beginning of the summary.

A candidate summary is evaluated during phases 3 and 4 by marking occurrences factoids in the summary and adding the weights of the factoids. The final score is the sum of weights of factoids in the candidate summary, divided by the sum of weights of factoids in the set of reference summaries.

Pyramids. The Pyramid evaluation scheme (Nenkova and Passonneau, 2004) is very similar to the factoid evaluations of Teufel and van Halteren (2004). Pyramid scores of summaries are acquired in the same four phases as factoid scores, as discussed in the previous paragraph. The main difference between the pyramid and the factoid method is how SCUs are used when compared to factoids. An SCU is a Summary Content Unit, as the equivalent of factoids are called in the pyramid method. The form of factoids is defined more strict and matching factoids from different summaries is more strict than SCU matching. In practice, strict matching causes problems, e.g. if one summary phrases a factoid more general than another. Teufel and van Halteren solve this by adding annotation layers: they annotate generalization relations between factoids and they annotate the fact/opinion status of a factoid. Nenkova and Passonneau approach this problem differently by accepting an SCU match when a SCU contains "largely" the same information as another SCU.

Pyramids and factoids differ slightly in the way the scores are calculated. While the factoid score depends on a number of factors, the contribution of each SCU to the score of a candidate summary is simply linear to the number of reference summaries it occurs in. The factoid method prefers factoids which occur near the beginning of a reference summary.

The Pyramid method has been applied in DUC first in 2003, and the DUCView tool[4] for Pyramid annotation has been made available publicly. An obvious limitation of the Pyramid (and Factoid) method is that manual annotation is required. Attempts have been made to automate phase 3 Harnly et al. (2005). This still requires manual annotation in phase 1, but this has to be done only once for each document set, no matter how many candidate summaries have to be evaluated. Harnly et al. achieve better correlation with the Pyramid score when using their automated Pyramid annotation than when using Rouge n-gram recall.

[4] `http://www1.cs.columbia.edu/~ani/DUC2005/Tool.html`

Performance of factoids and pyramids. Teufel and van Halteren (2004) evaluated their factoids method on a small number of summaries, and claim that factoids can be extracted from and annotated in naturally occurring text with high agreement between annotators. They conclude that factoids are well defined and reproducible. With large sets of reference summaries, they also found high correlation between factoid scores of the same summary based on factoids from different reference summary sets. They suggest to use 20–30 reference summaries as a minimum required to obtain reproducible scores. However, correlation of factoid scores with human rankings as used in DUC 2001 was low, and so was correlation with Rouge scores. On the other hand, the DUC 2001 scores used for comparison are based on human relevance judgments of summary sentences, which Lin and Hovy (2002b) claimed is an unreliable evaluation method.

Experimental results of Teufel and van Halteren were confirmed by (Passonneau, 2005) who did similar experiments with Pyramid annotations on a larger document set: they measured high inter-annotator agreement and they were able to reproduce pyramid scores with independent annotations of the same document set. Like factoid scores, pyramid scores showed low correlation with manual DUC 2001 annotations and with n-gram recall. A point of concern Passonneau raised was that differences in Pyramid scores across document sets was greater in magnitude than differences across systems. They suggest this may complicate comparing performance of systems. On the other hand, Donaway et al. (2000) argued that in general, scores are not comparable across document sets – *rankings* of different systems should be compared rather than (the average of) the actual scores.

4.2.1.5 Nuggets and Basic Elements

Refined evaluation methods such as Factoids and Pyramids still require human annotation. The Basic Elements (BE) method (Hovy et al., 2006) uses content units smaller than factoids/SCUs, which (unlike factoids and SCUs) can be extracted automatically from text, but they are more refined than n-grams as used by Rouge. Hovy et al. (2006) envision the use of compound Basic Elements (as they call their content units) to mimic the Pyramid method. Basic Elements are defined as follows:

- the head of a major syntactic constituent (noun, verb, adjective or adverbial phrases), expressed as a single item; or
- a relation between a head-BE and a single dependent, expressed as a triple (*head*, *modifier*, *relation*).

Table 4.2: Correlation between responsiveness and various automatic metrics in DUC 2005 (Hovy et al., 2006).

Metric	Spearman	Pearson
Rouge-2	0.900	0.926
Basic Elements	0.905	0.902
Pyramids	0.785	0.818

Basic Elements can be considered an evaluation framework rather than a metric. How exactly BEs are extracted from text is left open to the implementation, and so is BE comparison, and how summary scores are derived. However, Hovy et al. (2006) offer a full implementation of the framework, including modules for BE matching and BE extraction using various dependency parsers. For instance, the extraction module using MiniPar (Lin, 1998) extracts the following Basic Elements form the sentence, "Two Libyans were indicted for the Lockerbie bombing in 1991":

- (libyans, two, nn)
- (indicted, libyans, obj)
- (bombing, lockerbie, nn)
- (indicted, bombing, for)
- (bombing, 1991, in)

Zhou et al. (2007) acknowledge that automatic matching of content units is error prone, but they insist that reliable automatic extraction of content units is feasible. As an alternative to the Basic Elements method, they proposed a semi-automatic evaluation method. First, using syntactic patterns, the content units (which they call *nuggets*) are extracted from reference summaries as well as candidate summaries. Then, human annotators decide which nuggets are equivalent, so that an information recall score of the candidate summaries can be calculated.

4.2.1.6 Discussion

Several methods and metrics have emerged for evaluating relevance of summaries' content, most of which rely on a ground truth in the form of hand-crafted reference summaries. Others use a manual relevance judgment of the summary or of content

Table 4.3: Correlation between responsiveness and other metrics in DUC 2006 (Dang, 2006).

Metric	Spearman	Pearson
Rouge-2	0.767	0.836
Rouge-SU4	0.790	0.850
Basic Elements	0.797	0.782
Pyramids	–	0.787[a]

[a] My own measurement using DUC 2006 data – not published in Dang (2006).

Table 4.4: Correlation between Pyramid scores and other metrics in DUC 2005 (Hovy et al., 2006).

Metric	Spearman	Pearson
responsiveness	0.785	0.818
Rouge-2	0.665	0.880
Basic Elements	0.807	0.815

elements in the source text. Evaluations based on relevance judgments have been criticized for being unreliable due to variability in human judgment (e.g. Lin and Hovy, 2002b; Hovy et al., 2006). However, the only way that fully automatic metrics are validated is by measuring correlation with metrics based on human judgment, usually SEE for generic summaries or responsiveness for query-based summaries. Because a number of metrics were used in DUC 2005 and 2006 (query-based multi-document) summarization evaluations, it is a great resource for comparison of evaluation metrics. The responsiveness and the Pyramid metrics both use fully manual procedures to determine relevance of information, and may be used to validate automatic metrics. Tables 4.2 and 4.3 show Spearman and Pearson correlations between responsiveness and other metrics in DUC 2005 and 2006 respectively; Table 4.4 shows correlations with Pyramid scores in DUC 2005. It is remarkable that mutual correlation of responsiveness and Pyramid scores is lower than correlations of either metric with automatic metrics.

Another point of concern is the supposed *context-independence* of information-carrying text units. Apart from responsiveness, metrics divide a summary into content units and assign a relevance score to each content unit more or less independently.

This contradicts models of text understanding (e.g. Kintsch and van Dijk, 1978; Mann and Thompson, 1988) which argue that the meaning and understandability of a content unit may depend on its context, or that a content unit may serve a particular purpose in discourse other than plainly to inform the user of its content.

Content-based evaluation methods described here focus entirely on *generic* summarization. How well are they applicable to query-based summarization? Evaluation methods which rely on hand-crafted model summaries are usable to a certain extent as long as the conditions pertaining to the creation of the reference summaries are maximally similar to the peer summary conditions. In fact, results may be more consistent if the summarization task is defined better. In query-based summarization, the summarizer has more clues to what is relevant and what is not. As a result, evaluation metrics may perform more consistently. To my knowledge, this has not been verified.

4.2.2 Linguistic quality

The *linguistic quality* of a summary is the overall readability of the text, regardless of content or potential use. This includes coherence, style and grammaticality. Typical problems in *automatically* created summaries include:

- a sentence cannot be properly understood because it lacks context;
- a changed meaning of a sentence due to appearance in a different context;
- anaphora which cannot be resolved because its antecedent is not in the summary;
- ungrammatical sentences caused by sentence revision or end-of-sentence detection.

4.2.2.1 Subjective criteria

The readability of a summary can be measured by having human subjects read the summary and rate it for specific qualities, but there is no consensus on the actual readability criteria to use. A fairly general approach is to ask judges to assign a single rating of ‘acceptability’ based on a number of guidelines (e.g. Brandow et al., 1995; Saggion and Lapalme, 2000). The judges of Saggion and Lapalme (2000) were free to use their own readability criteria, but they suggested a number of criteria to them which are derived from Rowley (1988).

- good spelling and grammar;

- clear indication of the topic of the source document;
- the use of an impersonal style;
- consisting of one paragraph;
- conciseness;
- readability and understandability;
- the presentation of acronyms along with their expansions;
- other criteria that the judge considered important for readability.

These criteria are derived from Rowley (1988). Rowley aim at assessing the quality of human abstracts. DUC 2005 introduced a questionnaire with readability criteria which specifically targets automatic abstracts (Dang, 2005). Furthermore, the assessment was broken down into five questions, instead of using one general question. Summaries were rated on a 5-point scale for each of the following criteria:

- grammaticality;
- non-redundancy;
- referential clarity;
- focus;
- structure and coherence.

4.2.2.2 Objective measurements

If readability qualities can be formalized as objective criteria, they would not be subject to human variability and thus easier to reproduce. Another advantage is that they may be easier to automate, as to reduce human effort.

Cherry and Vesterman (1981) developed the Unix Style tool for analyzing surface characteristics of text that are important for a good writing style. The Unix Style tool implements various readability indices proposed in literature, based on textual features such as sentence and word length. Although the tool is designed to measure the readability of human text, it has also been used for measuring readability of automatic summaries (e.g. Mani et al., 1999). Schwarm and Ostendorf (2005) argue that the use

of long sentences and long words does not necessarily imply that the text is difficult to read. As an alternative, they propose an approach to measuring readability using statistical language models. However, also their method is designed for assessing human text. It is unclear whether their method is suitable for measuring the quality of summaries.

Minel et al. (1997) specifically aim at evaluating the readability of automatic summaries. Along with subjective assessments, they counted grammar and style errors of a small number of priorly defined error types in order to measure the quality of a summary. More specifically, they counted the number of anaphora deprived of antecedents and the number of instances of improper use of linear integration markers. For instance, the use of *firstly* in a text suggests that a list of items will follow. It would count as a style error if *firstly* marks a list of only one item.

A more extensive list of countable grammar and style errors in automatic summaries was used in the DUC 2003 evaluation (Over and Yen, 2003) as a measure of quality. This includes grammatical errors, capitalization errors, dangling conjunctions, improper use of referential expressions (e.g. pronouns or definite noun phrases), and discontinuities in coherence (e.g. time sequences or cause-effect relationships). Some of these errors may be automatically detectable using heuristics or a general purpose spelling checker.

4.2.3 Utility-oriented evaluation

Previous sections described how specific qualities of a summary are evaluated, but it is not clear how each of these qualities contribute to a real-life environment in which the summarization system may be employed. In contrast to intrinsic (content-based and linguistic quality) metrics, extrinsic or utility- metrics measure how a summarization system affects the *user*. If a summary is supposed to help the user complete a specific task, we may be interested in a number measurements:

- the time it takes the user to perform the task (efficiency);
- how it affects the result (effectiveness);
- the user's perception of either of the above (appreciation).

4.2.3.1 Categorization and relevance assessment

Clearly, how the summary affects a task depends on the application. A summarization system may be used to help an information analyst categorize documents if using a summary instead of (or in addition to) the full text saves time or increases categorization results. Minel et al. (1997) measured accuracy of categorization decisions based on summaries in an experimental environment with eight categories such as political, sociological, scientific or technical, etc. It was allowed to assign more categories to a document. Klein et al. (1999) performed similar experiments on a larger scale in the context of the TIPSTER SUMMAC evaluation program with a number of summarization systems and five disjunct categories.

In an information search process, a summarization system may help assessing the relevance of documents more quickly and even more accurately. In DUC 2003 (Over and Yen, 2003), users were asked to rate the *usefulness* of indicative summaries, where usefulness was defined in terms of time efficiency, i.e. to decide whether or not to read the full text.

Tombros and Sanderson (1998) set up an experiment in an IR setting in which the users had to identify relevant documents to a given query. A list of potentially relevant documents was produced by an IR system, and subjects were given five minutes to select as many relevant documents as possible. A query-based summary was presented along with the full text each document to determine its relevance. Measurements taken were precision and recall of selected documents, the number of documents examined, the number of references to the full text, and the general appreciation of the users. Similar experiments were run using generic summaries instead of query-based summaries (Brandow et al., 1995; Jing et al., 1998; Miike et al., 1994).

The TIPSTER SUMMAC (Klein et al., 1999) experiments are also based on the same ideas but were more constrained, as access to the full text was not allowed. Klein et al. also did not have a five minute time limit. While Tombros and Sanderson used a lead summary (i.e. the first few sentences) as a baseline, Klein et al. used the full text as one of the experimental conditions. A serious limitation of the latter approach is that the full text must be of reasonably small size so that it can be read in a reasonable amount of time.

4.2.3.2 Information gathering

In summaries aiming at informing the user, reading understanding is an essential quality. In experiments of Morris et al. (1992), subjects read a number summaries, and

then had to answer a number of questions to which the answers were in the summaries. Morris et al. (1992) were specifically interested in the loss of information for different summary lengths, and concluded that the documents they used could be condensed to less than 25 percent of their size without significant information loss.

The reduction of workload can be measured by scenario-based studies, in which users are assigned a task they have to carry out. For instance, Bakshi et al. (2003b) measure the effectiveness of a question-answering system by the number of queries that users pose to the system before they find the required information.

In experiments of McKeown et al. (2005), subjects were presented a topic, consisting of a number of questions. Using a web interface to a number of documents grouped into four clusters, the subjects were given 30 minutes to write a report containing the answers to the questions posed in the topic. Under various conditions, the web interface showed along with the documents:

1. no summary (document headlines only);
2. a one-sentence lead summary for each document and a one-sentence summary for each cluster;
3. an automatically generated multi-document summary for each cluster;
4. a human multi-document summary for each cluster.

After the experiment, the users filled out a questionnaire in which they were asked how they experienced the task, including how difficult it was and how they felt the summaries affected the task. In total, the corpus included four topics and related document sets, and all 45 subjects covered each of the four topics once.

The quality of the users' reports was measured using the Pyramid method (Nenkova and Passonneau, 2004, described earlier), so user performance under different conditions could be measured. In order to measure how much of the information in the reports originated from summaries, users were asked to cite their sources while writing the report. Percentages of summary information in the report ranged from 8 percent (one-sentence summary) to 27 percent (human summary).

A problem of the experiments was that reports vary greatly in length (between 102 and 1525 words), and long reports tended to contain much redundant information. McKeown et al. attribute this to the setup of the experiment: subjects were allowed to cut-and-paste content directly in the reports without producing coherent text. They suggest that the report lengths would be more consisted if the task was more directed

toward synthesis rather than cut-and-paste. McKeown et al. also remarked that it is difficult to control all variables in user experiments involving a complex task.

4.3 Content selection

As mentioned in section 4.1, the human summarization process consists of three stages: understanding, interpretation and synthesis. Analogously, typical stages of automatic summarization are *preprocessing*, *content selection* and *revision*. During preprocessing, an internal representation of the source and the query (if any) is constructed. This may include NLP tasks such as indexing, part-of-speech tagging, dependency parsing, discourse analysis. During content selection, relevance decisions are made and relevant passages (usually sentences or paragraphs) are extracted. In a coherent source, the selection of passages to extract not only affects the subset of the information of the source to appear in the summary, but also coherence of the summary. Finally, revision may involve ordering of extracted passages, sentence fusion, sentence reduction and generalization/specification. This section discusses the content selection process. Since the kind of preprocessing required depends on the features used for content selection, preprocessing is not discussed separately. Revision is discussed in section 4.4.

In the interpretation stage, summarization systems (explicitly or implicitly) use knowledge about how the source is organized. This can be rooted in theories of discourse organization.

4.3.1 Discourse models for content selection

According to the discourse model of Kintsch and van Dijk (1978), a discourse consists of a set of propositions. As these propositions are processed by the reader, they may be generalized, constructed by combining propositions, or deleted if the proposition is insignificant. Coherence is modeled as interdependence between propositions, and is the cause of compression of the document in memory by reduction of the number of propositions. This we will call the *memory-based* discourse model. A simplified version of this model is used by most state-of-the-art summarization systems. They attempt to find the set of most significant propositions. Each passage is judged by its significance independently, based on correspondence with the user query (if any), and redundancy with respect to other candidate passages. The passages with the highest significance value are extracted. An implicit assumption behind this is that coherence is implied when including the most salient propositions in the summary. After all, coherence is

present if the reader can understand the summary using previous propositions and prior (background) knowledge.

In contrast, with their Rhetorical Structure Theory (RST), Mann and Thompson (1988) take a more discourse oriented perspective with respect to coherence. Other intention-based discourse models are those of Grosz and Sidner (1986), Wolf and Gibson (2005). Their position is that a text is intended to convey a piece of information or an opinion, and that the author uses most of the text to support that information. This process is recursive in the sense that peripheral information may in turn also be supported. Similarly to the memory-based model, a discourse consists of a set of propositions, and the atomic unit of discourse is the smallest phrase conveying a proposition. According to Mann and Thompson (1988), peripheral information can not be interpreted properly without the information it supports. The result is a hierarchical view on a discourse: the top is the information most close to the author's intention; each following layer consists of information supporting a particular piece of information in the previous layer. This *intention-based* discourse model is used by the summarization system of Marcu (1999). The content selection process consists of two phases. First, the discourse structure is identified in terms of RST. Then, the lower levels of the RST structure are discarded, until the remaining text is the desired size for the summary. The relevance of a sentence depends on the relevance of other (related) sentences. In doing so, this approach makes discourse-level coherence relations between passages explicit in source and summary. Intuitively, it is therefore more suitable for informative summarization, as informative summarization is more demanding with respect to coherence (c.f. NISO, 1997).

Radev (2000) attempted to scale RST to accommodate single analyses of multiple documents, but by doing so, he relaxed the definition of *rhetorical* relations by removing the intentional aspect. In RST, each relation comes with a proposition which is intentional but implicitly conveyed by the juxtaposition of the related sentences. According to Radev (2000), a CST relation may also be conflicting information in newspaper articles, e.g. conflicting numbers of victims in a plane accident. The cause of this conflict may be false sources of information, or out-dated information, but the conflict is not necessarily or even unlikely to be intended, as is a requirement for RST relations. Although good results have been reported with using CST information in multi-document summarization (Zhang et al., 2002), CST suffers from problems similar to RST's, namely that CST relations are difficult to detect automatically.

A third model for summarization is the *schema-based* discourse model (van Wijk and Sanders, 1999; Rumelhart, 1975). The schema-based model is hierarchical, simi-

larly to the intention-based model. Van Wijk en Sanders (1999) argue that knowledge about the genre can be exploited to predict the structure of the text because authors use writing conventions of the genre. This model has been used in combination with the memory-based model by Radev et al. (2004), who argue that newspaper articles describing an event start with a brief description of the event, followed by less and less important details. This knowledge can be used to create a summary by giving sentences near the beginning of the article preference over other sentences.

4.3.2 Features for content selection

Apart from a few exceptions, most summarization systems use the memory-based model. The main reason for this is that it is fairly generally applicable, requiring a minimum of domain knowledge. Schema-based summarization relies on specific expectations of the format of the source. The main disadvantage of the intention-based summarization system of Marcu (1999) is that it relies on an RST analysis, for which there is presently no reliable way to generate it automatically.

Content selection systems select passages – typically sentences but possibly paragraphs – from one or more source text. Thus, for each passage, they decide whether or not to include it in the summary. This decision may rely on features of the **passages as such** or its position in the document, e.g. cue phrases, weighted term frequencies, the presence of anaphoric references. It may rely on the **query** (if any), e.g. the presence of query terms, or on its **relation with summary candidate passages** by any definition of 'relation'. For instance, a passage may be relevant because it is required to understand another passage. This aspect of summarization is widely neglected in literature, possibly because the conditional relevance of a passage is harder to establish.

A number of authors have implied that some passages are inherently more salient than others (e.g. Otterbacher et al., 2005; Jagarlamudi et al., 2006). These authors argue that there is a prior probability that a passage in a document is salient, independent of the query, and that the salience value of a passage is a weighted average of the relevance based on the document alone and the query-relevance.

$$R_{passage} = (1-\lambda) \cdot R_{doc} + \lambda \cdot R_{query} \tag{4.13}$$

4.3.3 Lexical knowledge and cue phrases

Edmundson (1969) hypothesized that authors indicate importance by using pragmatic *cue words* such as “significant”, “impossible”, and “hardly”. Edmundson stored three types of words in a dictionary: bonus words, stigma words, and null words. If a sentence contains a bonus word, its relevance probability is increased. Stigma words indicate irrelevance, and null words do not affect the sentence’s relevance. Edmundson used a supervised learning algorithm (section 4.3.8) to derive the cue dictionary from a set of documents and reference extracts.

Paice (1981) generalized this idea by using cue *phrases* instead of cue words as indicators of relevance. In the target genre of the summarization system of Paice — scientific papers — authors use linguistic structures such as “we found that ...” and “the results of this study confirm that ...”. These phrases may occur in numerous variations, but Paice argue that the underlying structures are limited in number. For each group of phrases, seven in total, templates were constructed, similar to regular expressions. If a sentence matched the template, it has an increased probability of relevance.

Due to the relative infrequent occurrence of cue words and phrases, both Edmundson and Paice use the cue method in conjunction with other indicators of relevance.

4.3.3.1 Sentence structure

Not only cue phrases may indicate importance, but authors may also alternate syntactic constructs to this end. Preliminary investigations by Earl (1970) to find syntactic patterns for use in summarization was unsuccessful. Earl suggests that syntactic patterns may positively affect a summary’s quality when used in conjunction with other methods such as word frequencies. Janoš (1979) suggests that a functional analysis of a sentence may help for summarization.

4.3.4 Term frequency

A writer who discusses a specific topic is likely to frequently use words related to that topic. On the other hand, the most frequent words are usually *function words* (e.g. articles, propositions) which are not topic-specific and may appear in any text. Thus, for generic summaries, Luhn (1958) proposed to use term frequency in a document to derive describing power: both high and low frequency terms are less likely to be topic-related than mid range frequency terms. The high frequency words can be filtered out

by using a cut-off frequency or a *stop list*, i.e. a list of words which can be ignored safely.

Based on a term's occurrence frequency, the extraction system of Luhn (1958) distinguished *significant* from *insignificant* terms. It extracted sentences based on a *sentence significance* value, which was derived from a sentence's number of significant words and the linear distance between them.

4.3.4.1 The vector space model

Besides the occurrence frequency, Luhn relied for summarization on the position of terms in a sentence relative to each other. Later summarization systems tend to drop the order of words as a distinguishing feature, and represent passages as term-frequency vectors. The vector space model was introduced in information retrieval by Salton (1988) and gained popularity in summarization as well (e.g. Radev et al., 2004; Erkan and Radev, 2004). In this model, each passage is represented as a vector of terms $[f_1 ... f_n]$, where f_i is the number of occurrences of term i in the passage, and n is the number of distinct terms in the corpus (c.f. section 2.1.4). Two passages can be compared by measuring the similarity of their vector representations in n-dimensional space. This is typically done by measuring the cosine of the angle between the two vectors.

In a corpus with n distinct terms, a document can be written as a vector of length n. Given documents A and B and their respective vector representations $[a_1..a_n]$ and $[b_1..b_n]$, the cosine similarity of those documents is their angle in n-dimensional space, calculated as follows (Salton, 1988):

$$cosim(A,B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \tag{4.14}$$

$$= \frac{\sum_{i=1}^{n} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}} \tag{4.15}$$

In the term-frequency vector representation in its pure form, all words are of equal importance. This does not reflect reality, as a *content word* such as *establishment* tells more about the nature of a text than a *function word* such as *the* or *of*, which may appear in any text. So, some words should weigh heavier than others. Term weighting is presently considered the most important challenge in information retrieval, and many ways for weighting terms have been proposed in literature. In eq. 4.15, terms are weighted by their occurrence frequency in the compared vectors. Weights can be applied as a transformation of the term frequency vectors, so that the same formula 4.15

can be applied. The vector representation of a sentence $A = [a_1..a_m]$ is now derived from the term frequency vector by the weighting function. In case term frequencies are used directly, $a_i = tf_i$, where tf_i is the occurrence frequency of term i in the sentence.

Luhn (1958) maintained a *stop list* of non-content words, so these words can be disregarded by assigning a weight of 0. A disadvantage of stop lists is that they usually rely on hand-crafted rules and there is no general consensus on a list of stop words. Alternatively, weights are assigned to terms depending on their class. For instance, nouns and verbs may be considered more meaningful than articles or propositions. In particular, named entities are important to understand what a text is about (Conroy and Schlesinger, 2006). A notable limitation of this approach is that term classes must be recognized as such, requiring a deeper text analysis.

The most wide spread method for weighting terms is to use corpus statistics to distinguish common words from uncommon words. In its most basic form, a cut-off frequency is used as a threshold; all terms occurring in more than a specific number of documents receive a weight of 0 (Luhn, 1958). A more refined weighting method is *inverse document frequency* (IDF), proposed by Salton (1988). Rather than dividing terms in two classes — content words and stop words — a word's IDF value is higher (and thus is more descriptive) if the word appears in less documents, hence the term *inverse* document frequency. $tf \cdot idf$ is the *term frequency* (*tf*) multiplied by the *inverse document frequency* (*idf*). Inverse document frequency is a weighting function which returns a value greater than 0. Prevalent terms in the corpus (occurring in many documents) are less characteristic for a topic than uncommon terms if they occur equally frequently in a document. These terms have a lower *idf* value. The term frequency *tf* is the number of occurrences of a term in a document; *idf* of a term is defined as the logarithm of the ratio of number of documents in the corpus ($\|D\|$) and the number of documents in which the term occurs ($\|\{d \in D \mid t_i \in d\}\|$):

$$idf_i = log \frac{\|D\|}{\|\{d \in D \mid t_i \in d\}\|} \tag{4.16}$$

Salton and Buckley (1988) compared a number of weighting schemes in information retrieval. They concluded that $tf \cdot idf$ performs best for weighting document terms. For weighting query terms, the best weighting scheme appeared the following.

$$(0.5 + \frac{0.5 \cdot tf}{maxtf}) log \frac{\|D\|}{\|\{d \in D \mid t_i \in d\}\|} \tag{4.17}$$

4.3.4.2 Similarity-based summarization

In query-based summarization, the query is represented as a vector as well. The core idea is that, in order to maximize relevance to the query, sentences most similar to the query constitute a query-based summary. In multi-document summarization, redundancy is a complicating factor. Carbonell and Goldstein (1998) attempt to reduce the risk of selecting sentences with redundant information by maximizing the *marginal* relevance. The marginal relevance of a sentence indicates how much a sentence contributes to the summary *in addition* to the already-selected sentences. Given a query vector q and a summary S, the marginal relevance of a sentence d is measured as follows.

$$\text{marginal relevance} = \lambda \cdot cosim(d,q) - \lambda \cdot (1-\lambda) \cdot \max_{s_i \in S} cosim(d,s_i) \tag{4.18}$$

In words, the marginal relevance of a sentence is positively affected by similarity with the query, and negatively by similarity with previously selected summary sentences. The set summary sentences S is initially empty, after which sentences are added one by one, based on their marginal relevance, until the desired length of the summary is reached.

Similar methods of preventing redundancy are adopted in generic summarization. For instance, Hovy et al. (2005a) veto inclusion of a sentence in a summary if it is too similar to already selected sentences.

In generic single-document summarization, a good summary should be representative of its source (section 4.2.1.1). If a term-frequency vector of the source is available, fidelity to source can be measured as similarity between the term-frequency vector of the summary and that of the source. An extractor can then simply select the sentence which brings the summary representation closest the the source representation, by taking the sentence with the highest cosine similarity with respect to the full document vector. This is repeated until specified criteria (e.g. desired summary length) are fulfilled. This method was refined and extended for multi-document summarization by Radev et al. (2004). Radev et al. measured relevance as a combination of factors. One of them was similarity to the *centroid* of the document cluster. The centroid is a pseudo-document, a term-frequency vector derived from the document cluster. The centroid is used as a query in a way similar to how Carbonell and Goldstein (1998) maximize marginal relevance. Radev et al. used the $tf \cdot idf$ weighting scheme for doc-

ument sentences. The centroid was also $tf \cdot idf$-weighted, but all terms with a $tf \cdot idf$ value below a certain threshold were weighted 0.

In some corpora, descriptive titles and headings may be present. Viewing them as a very short indicative summary, title words are likely to appear in a longer, more elaborate summary as well. Hence, the presence of title words increases the salience of a sentence (Edmundson, 1969). The same is true for the presence of first-sentence words (Radev et al., 2004).

A problem for cosine similarity is that terms are either different or equal, terms cannot be 'similar'. Even the same word appears in a different form (e.g. *establish*, *established*, *establishment*), it is treated as another term. Normalization of words can be applied to address this problem, e.g. *word stemming*. By using stems of words, a term like *establishment* is recognized as an alternative form of *establish*. Methods to acquire word stems include dictionary-based stemming and algorithmic stemming. A widely used algorithmic stemmer is the Porter stemmer Porter (1980), which uses a set of heuristics to normalize terms to a common form. Although the original algorithm was limited to English text, it lead to a generic 'programming' language for stemmers (Porter, 2001) in which stemmers for a variety of languages are designed. Algorithmic stemming may improve performance, but, as Harman (1999) point out, it may also cause false positives which balanced out the performance gain in her IR experiments.

4.3.5 Cohesion

While coherence is used to describes links between ideas, cohesion describes links between textual elements on a linguistic level. How is cohesion – anaphoric references, ellipsis, lexical iteration, etc – relevant for summarization? First, the cohesive structure of a text provides information about topic shifts and the general flow of a text. Second, a cohesive tie links to items of which interpretation of one relies on the other. If one is included in a summary but not the other, the cohesive structure of the summary is broken. As a result of this lack of context, the summary is more difficult to read or become illegible. It is also possible that unintended cohesive ties are introduced if a textual item in the summary becomes a likely antecedent candidate for a dangling reference. Taking cohesion into account may prevent such problems.

4.3.5.1 Lexical chains

Barzilay and Elhadad use the WordNet thesaurus (Miller, 1995) and an algorithm based on Morris and Hirst (1991) to extract lexical chains from text. They rank the lexical

chains by the number of distinct words that are part of the chain. Then, Barzilay and Elhadad devise a summary from one sentences for each of the strongest chains of the source text. They invented a set of heuristics to decide which sentence is actually picked to represent a chain. For instance, one of their methods selected the sentence containing the first word of a chain. By covering the strongest chains of the source text, Barzilay and Elhadad aim to maximize coverage and minimize redundancy.

Witte and Bergler (2003) combined a number of heuristics to determine the similarity of noun phrases in order to create what they call *co-reference chains*. They used thesaurus relations (synonymy and hyponymy), substring matching, acronym matching, pronoun resolution, and head identity to determine whether two noun phrases co-refer to the same item.

4.3.5.2 Matrix decomposition

Constructing a thesaurus is laborious and expensive, and coverage of presently available thesauruses varies by language and domain. Instead of relying on a thesaurus, Manabu and Hajime (2000) used text statistics measure of word similarity for discovering lexical patterns (see section 2.1.4).

Manabu and Hajime (2000) still need words to co-occur to detect a semantic relation between them. Matrix decomposition can be used to discover similarity of words if they are commonly used in a similar context, even if they never co-occur. Gong and Liu (2001) use singular value decomposition (SVD) to discover 'latent concepts' from corpus statistics. More recently, Park et al. (2006) proposed non-negative matrix factorization (NMF) for the same purpose. Latent concepts can be roughly described as sets of terms which share a lexical environment, similar to lexical chains. For instance, if the terms *PM* and *Balkenende* were used interchangibly along with words such as *elected, parliament, jurisdiction,* these terms would all contribute to the same concept. In the term-frequency vector representation, text documents are represented as a set of terms. Matrix decomposition is used to represent documents (or passages) as a set of concepts; concepts are represented as a set of terms.

Similarly to lexical chains, terms are grouped by topic, the main difference being that corpus (co-occurrence) statistics are used, while these lexical chains extraction algorithms use semantic relations from thesauruses. Also, what is a 'concept' is loosely defined, while thesauruses contain well-defined relations such as hypernymy, antonymy and synonymy.

If each of m document is represented as a term-frequency vector of n terms, an $m \times n$ matrix X can be constructed to map terms to documents. This matrix can be decomposed into two matrices, W and H, and a residual matrix E:

$$X = WH + E \tag{4.19}$$

where W is an $m \times r$ matrix; H is an $r \times n$ matrix; and E is the residual matrix with dimensions $m \times n$. The value of r is the maximum number of concepts, and is chosen as $r \leq n$. The matrix W maps terms to concepts: it has a row for each term and a column for each concept. Matrix H maps concepts to passages. If $r < n$, there may not exist matrices W and H such that $X = WH$. In that case, X is approximated by WH with error E.

At least two forms of matrix decomposition have been used for generic summarization. Inspired by Latent Semantic Analysis (Deerwester et al., 1990), Gong and Liu (2001) use singular value decomposition in summarization to recognize 'relatedness' between terms by means of corpus statistics. SVD decomposes an $m \times n$ matrix X as follows:

$$X = W\Sigma H \tag{4.20}$$

where W and H correspond to the identically-named matrices in eq. 4.19. W is an $m \times n$, H is an $n \times n$ matrix. Σ is an $n \times n$ diagonal matrix containing non-negative singular values, which can be viewed as prominence values of corresponding concepts.

Park et al. (2006) proposed to use non-negative matrix factorization for matrix decomposition in summarization. NMF factorizes a matrix in correspondence with eq. 4.19 with matrices W and H containing non-negative values only. In SVD, the matrices W and H may contain negative numbers, meaning that a term may negatively 'contribute' to a concept and a concept may negatively contribute to a document. Park et al. argue that this is counter intuitive. As the matrices W and H contain only non-negative values in NMF, they favor the use of NMF for summarization.

Gong and Liu represent sentences of the source text as term-frequency vectors, and apply SVD to the resulting matrix. SVD not only derives concepts, but also orders them by prominence in the text. Gong and Liu (2001) use concepts for content selection in a way similar to how Barzilay and Elhadad (1997) exploit lexical chains. Gong and Liu start with the most prominent concept and select the sentence most representative for that topic. Then, they extract a sentence for the second most prominent concept, and so on, until a pre-defined number of sentences is extracted. Gong and Liu compose a

summary of the most representative sentences of the most salient concepts. Doing so, Gong and Liu aim to maximize coverage and minimize redundancy.

Independent from the decomposition method of choice, Park et al. also used a different content selection algorithm from Gong and Liu because their goal was to create query-based rather than generic summaries. Park et al. first select the concept most similar to the query (in matrix W) and extract the sentence to which the concept contributes most (in matrix H). Then, they extract a sentence for the concept second most similar to the query, and so on, until a pre-defined number of sentences is extracted.

4.3.5.3 Graph based methods

Lexical chains and matrix decomposition are methods for clustering words (into chains or concepts respectively). A content selection algorithm uses the clusters to select the best sentences for a summary. Alternatively, relations between sentences or passages are represented as a graph, and properties of the graph are used to determine which passages are most salient.

Degree. Salton et al. (1997) construct a graph in which each vertex represents a sentence, and each edge a semantic relation between sentences. The nature of the semantic relations is independent from the content selection strategy, but Salton et al. use IDF-weighted cosine similarity. They regard two sentences related if their similarity exceeds a certain threshold.

Sentences which are 'central' in the graph are regarded central to the topic of the document. Centrality of a sentence is measured as their *degree*, i.e. the number of related sentences. Salton et al. expect sentences of a peripheral topic to be relatively isolated in the similarity graph.

Normalized centrality. This rationale was taken further by Erkan and Radev (2004). In analogy with social networks, they argue, it matters not only how many people you know, but also *who* you know: sentences linked to well connected sentences are likely more salient themselves. Erkan and Radev made two modifications to the source text graph representation. First, the edges (links) of the graph are weighted and normalized, so that the sum of weights of all outgoing edges is normalized to 1. Since each sentence is similar to itself, all nodes have outgoing edges. Second, an 'entropy' value is assigned to each passage. Because of the weight normalization, the graph behaves as a Markov chain.

Their summarization algorithm iteratively alters the entropy of the sentences, until the algorithm converges to a stationary state. In subsequent iterations, each passage will activate related passages by dividing its entropy among connected nodes, thus allowing entropy to propagate through the network. Erkan and Radev argue that a node which is highly connected (i.e. a passage with many related passages) is more likely to be important and thus receives a higher entropy than less connected nodes, given a sufficient number of iterations. As the graph is a Markov chain, the salience of a node – the centrality – is the probability that someone would end up at that node after an arbitrary number of steps traversing the graph from node to node, regardless of where one started, each time randomly following an edge to a related node. Erkan and Radev (2004) formalize this as follows.

$$\mu_j(t+1) = \sum_{i \in adj_j} \frac{\mu_i(t)}{deg_i} \tag{4.21}$$

where $\mu_i(t)$ is the activation level of a passage i at iteration t; adj_i is the set of passages adjacent to i; deg_i is the *out degree* of i, i.e. the number of outgoing edges. Dividing by the degree of the node ensures that the graph behaves as a *Markov chain*, as the activation of each node is shared among adjacent nodes – no activation is created or lost. If the relations between passages may be of unequal strength (i.e. the probability of transitions between nodes differ), the weighted variant is applied by analogy:

$$\mu_j(t+1) = \sum_{i \in adj_i} \frac{w_{ij}\mu_i(t)}{\Sigma_{k \in adj_i} w_{ik}} \tag{4.22}$$

The 'final' activation levels are reached when the algorithm converges to a stationary state. This can be checked by estimating the error as a stopping criterion:

$$\sum_{i=0}^{n-1} \|\mu_i(t+1) - \mu_i(t)\| \leq \varepsilon \tag{4.23}$$

Convergence is essential, as the algorithm depends on this feature to determine salience. Erkan and Radev (2004) address a number of issues with respect to the adequacy of the algorithm, including irreducibility and periodicity. A graph is *irreducible*, if there is a path between any two nodes in the graph. If this feature does not pertain to a graph, all activation may be trapped in a small number of nodes. A graph is *periodic*, if there is an integer $k > 1$ that divides the length of each cycle in the graph. If

the graph is periodic, the algorithm may never reach a stationary state. On the other hand, a Markov chain is guaranteed to converge if the graph is irreducible and aperiodic. Following the PageRank algorithm of Brin and Page (1998), Erkan and Radev added a smoothing function to ensure these properties, by assigning a small non-zero weight to the transition between each two nodes, thus ensuring convergence. Neither Erkan and Radev nor Mani and Bloedorn make use of directed graphs, as both used lexical similarity for measuring relatedness. Nevertheless, the algorithm converges for directed graphs as well.

$$\mu_j(t+1) = d\frac{1}{N} + (1-d)\sum_{i \in ad j_i} \frac{w_{ij}\mu_i(t)}{\sum_{k \in ad j_i} w_{ik}} \tag{4.24}$$

where N is the number of nodes in the graph, and d is the non-zero weight (the "damping factor").

The sentences ending up with the highest centrality rank are the sentences which would receive more activation than they would yield to their neighbors in a state of equal activation. Erkan and Radev (2004) use a commutative similarity measure similar to that of Salton et al. (1997), meaning that $w_{ij} = w_{ji}$ for all (i, j). For any two sentences, if their activation is equal and the weight of their link is equal in both directions, the link does not affect the activation of either, since an equal amount of activation is carried in both directions. Due to the outgoing degree normalization, the weight of the link from i to j is not the same as from j to i if the out-degree of the nodes differ, despite the commutative similarity measure. In other words, unlike in Salton et al. (1997), the degree itself does not contribute to the centrality of a sentence; the degree of a node relative to the degree of its neighbors does.

The central idea behind centrality is that, after a sufficient number of iterations, highly connected nodes receive a higher amount of activation than less connected nodes. It appears that this is not necessarily the case in undirected graphs. In the extreme case, a passage which is not related to any other passage then itself, will directly return its activation level to itself. The result is that its activation will not be changed from its initial (average) value, which is likely higher than the final activation level of better connected nodes. In a practical situation, it is well possible that a to-be-summarized document cluster contains at least one document which is slightly off-topic which is hardly related to any of the other documents. Even if this isolated document did not have any relation with any other document, the average activation level of passages of this document would still be equal to the overall average activation level. In essence, the algorithm prefers nodes which are connected to nodes less

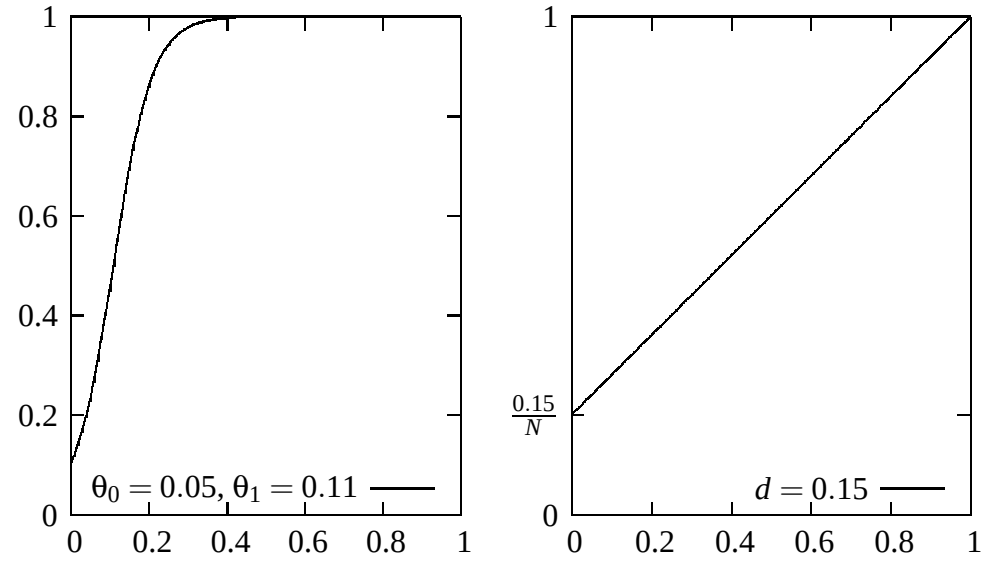

Figure 4.4: *Left:* hopfield transformation function of Chen and Ng (1995); *right:* centrality transformation function of Erkan and Radev (2004).

connected than themselves, as less connected nodes distribute their activation over less nodes.

Query-biased centrality. The algorithm for measuring centrality (as in Erkan and Radev, 2004) converges to the same state, regardless of initial activation levels. Without modification, this renders the algorithm useless for query-based summarization. Mani and Bloedorn (1997) set the initial activation levels to the sentence's query-relevance; the algorithm simply stops after a pre-set small number of iterations, rather than waiting until the algorithm converges. This way, the initially set activation has only propagated slightly, resulting in a mix of centrality and query-relevance. Otterbacher et al. (2005) modified the transformation function of Erkan and Radev (2004) by adding a query-dependant component to the transformation function:

$$\mu_j(t+1) = d\frac{w_{qj}}{\sum_k w_{qk}} + (1-d)\sum_{i \in adj_j} \frac{w_{ij}\mu_i(t)}{\sum_{k \in adj_i} w_{ik}} \tag{4.25}$$

where w_{qj} is the query-relevance of sentence j.

Other than that their algorithm stops after a few iterations, the approach of Mani and Bloedorn (1997) is almost identical to Erkan and Radev (2004). The most notable difference is the transformation function applied to the activation level after each it-

eration, which is a linear redistribution of weights in Erkan and Radev (2004) and a Sigmoid transformation in Mani and Bloedorn (1997). Both functions are plotted in Figure 4.4. The algorithm of Mani and Bloedorn is derived from Chen and Ng (1995), using a threshold value to discriminate 'activated' from 'unactivated' nodes. This allows the algorithm to converge after less iterations, at the cost of requiring the threshold to be tuned.

$$\mu_j(t+1) = \left(1 + exp\left(\frac{1}{\theta_0} \cdot (\theta_1 - \sum_{i=0}^{n-1} t_{ij}\mu_i(t))\right)\right)^{-1} \tag{4.26}$$

where θ_0 and θ_1 are the threshold values. Chen and Ng use the initial values of $\theta_0 = 0.05$ and $\theta_1 = 0.11$. If this does not give satisfactory results, the algorithm is rerun with other values. It should be noted that Chen and Ng used the algorithm for knowledge exploration in thesauruses, rather than summarization. Mani and Bloedorn (1997) did not publish their threshold values or number of iterations before stopping.

4.3.5.4 Anaphoric references

Although cohesion can be used to find salient passages for summary generation, a sentence which has a strong embedding in a text may be better avoided for use in extracts (Pollock and Zamora, 1975). In particular, an extracted sentence containing referring expressions cannot be interpreted as intended by the author if the antecedent is not present in the summary. The extractive summarization system of Pollock and Zamora tries to avoid dangling references in the summary, by including sentences containing referring expressions such as 'this' only if the preceding sentence can also be included in the summary. Even though this may not enhance a summary's content relevance, it partially relieves the revision task and reliance on anaphora resolution. While anaphora can be used as a positive (Mani and Bloedorn, 1997; Manabu and Hajime, 2000) as well as a negative (Pollock and Zamora, 1975; Brandow et al., 1995) cue for sentence scoring, lexical cohesion is typically used as a positive indicator for inclusion in a summary.

4.3.6 Coherence

According to theories of discourse organization (Mann and Thompson, 1988; Grosz and Sidner, 1986), documents are organized hierarchically: the author's main points are higher in the hierarchy than peripheral ideas. The summarization system of Marcu

(2000) exploits this information by ranking passages by their position in the hierarchy. Selecting the sentences highest in the hierarchy tree still leaves many possibilities. Marcu used rules for selecting the most salient sentences from the tree. O'Leary et al. (2001) experimented with machine learning algorithms to determine the optimal set of rules automatically. They used a hidden Markov model, a Bayesian model, and a dynamic decision-based model. In their evaluation measurements, the dynamic decision-based model achieved near-human performance.

The representation of discourse as a hierarchical tree is disputed by Wolf and Gibson (2005). They built a summarization system which used their non-hierarchical discourse model for content selection (Wolf and Gibson, 2004). They claim to have achieved a higher quality summaries with their graph-based approach than with the (RST) tree-based approach of Marcu (2000). Both systems used hand annotated coherence structures.

Blair-Goldensohn and McKeown (2006) used RST for conditional inclusion of sentences in a summary. They use the RST relation recognition of Marcu (2000), but they abandoned the idea that single coherence structure is required for summarization. Instead, they determine the salience of sentences using features such as centroid and lexical cohesion. If a sentence is selected for inclusion in a summary, they check whether the sentence participates in a rhetorical relation with another sentence. If that is the case, they include also that sentence in the summary.

4.3.7 Layout

Writing conventions of the genre may provide clues as to where particular information can be found. Baxendale (1958) concluded that in 85 percent of the 200 paragraphs of technical documents they examined, the first sentence contained the 'prime thought' of the paragraph. In 7 percent of the paragraphs, the prime thought was in the last sentences. Edmundson (1969) used this knowledge alongside other methods in an automatic summarization system. Lin and Hovy (1997) designed a method for using corpus statistics to derive the most salient sentence positions. They confirmed that also in newspaper text, the first sentence of a paragraph is often the most representative. Lin and Hovy (2002a) claim that taking the sentence position into account is also a good method to enhance coherence.

Using layout information such as sentence position is proven successful in summarization (Mani and Bloedorn, 1997; Lin and Hovy, 2002a) – it is also popular as a simple but competitive baseline. For instance, Dang (2006) used the leading sentences

of the most recent documents to create baseline summaries. Wolf and Gibson (2004) used the leading sentences of paragraphs in baseline summaries.

4.3.8 Machine learning for extraction

Lately, machine learning has been a major driving force behind NLP tasks. This includes sub tasks of automatic summarization, but extraction has lagged behind in this matter. Nevertheless, I will discuss two forms machine learning may take in extracting: *supervised learning* and *reinforcement learning*.

Both forms use the *noisy channel* model: there is an input x, an output value y, and an unknown (noisy channel) function f to transform x into $y = f(x)$. In an extraction setting, the input would be a representation of the source and the query, the noisy channel is an extractor, and the output is (a representation of) the extract created by the extractor. The goal of machine learning is to 'learn' the behavior of the noisy channel f and create a function g which estimates f. The performance of the extractor depends the quality of the model (i.e. the features) and the error $f(x) - g(x)$

For instance, suppose we suspect that certain cue phrases (e.g. "in sum", "most importantly") may indicate that a passage is relevant. In addition, lexical similarity to the query may be useful for extracting. These features are observable for each passage and may be encoded in x. The output value $f(x)$ could encode whether or not a passage is relevant. The function f encodes decisions based on features in x as to whether a passage is relevant or not. This function may be performed by a human (e.g. in manual abstracting) or by a machine (automatic extracting).

4.3.8.1 Supervised learning

In supervised learning, a set of input/output pairs $(x, f(x))$ is priorly known and is used as training data, and the task of machine learning is to infer $f(x)$ from these samples. If $f(x)$ is a transformation function from passage feature observations to passage relevance, this means the relevance must be known in advance. Edmundson (1969) used manually created reference extracts from single documents; extracted sentences were considered 'relevant'. As the original passages stay intact, this procedure is relatively straight forward. In general, any content evaluation method based on sentence (or passage) identity applies.

If the reference summaries are abstracts rather than extracts, abstract passages must be mapped on source passages in order to determine their relevance. Mani (2001) automatically mapped abstract sentences on the source sentence with the highest cosine

similarity, but this may introduce an additional learning error. However, renewed interest in tracing the origin of content in abstracts may open up new opportunities for applying supervised learning with abstracts as training material.

4.3.8.2 Reinforcement learning

Reinforcement learning is applied if the 'correct' output of function $f(x)$ is unknown, but the error $f(x) - g(x)$ *can* be estimated for individual samples. In the case of extracting, an extraction learning algorithm would start with a function f, and the resulting extract can be evaluated with an (automatic) evaluation metric. Next, f is slightly altered, and the learning algorithm learns from the way changes in f affect the summary score.

Genetic algorithms (Russell and Norvig, 1995) are an example of reinforcement learning which is applied in (extractive) summarization (Yeh et al., 2005). The summarization system of Yeh et al. uses a combination of features to judge sentence relevance: sentence position, keywords, centrality, and title words. Each of these features results in a score for each sentence; the final sentence score was a weighted average of the sentence feature scores. The genetic algorithm was used to iteratively tweak the feature weights. For evaluation, Yeh et al. used annotators who ranked sentences of the source text by relevance. A set of weights was successful if it caused relevant sentences to be picked.

It should be noted that reinforcement learning relies on an evaluation after each iteration, and it may take many iterations for the algorithm to converge. This makes this rationale especially suitable if a quick automatic evaluation method is available. In the case of abstractive summarization, there is no single generally accepted automatic evaluation measure. The most commonly used metric which runs without human intervention is n-gram recall (see section 4.2), but n-gram recall ignores a number of essential aspects of summary quality. In fact, n-gram recall is an *indication* of certain aspects of summary quality without measuring the quality itself. As a result, reinforcement learning may lead to improved n-gram recall scores without improving the quality of the summary in any way. Thus, after the iterative learning process has finished, a different evaluation metric would be required to validate the results. Yeh et al. avoid this by using sentence identity rather than n-grams, but sentence identity itself is disputed for use in summary evaluation (Donaway et al., 2000).

14A	RSI can be caused by repeating the same sequence of movements many times an hour or day.
14B	RSI is generally caused by a mixture of poor ergonomics, stress and poor posture.
14C	RSI can be caused by a mixture of poor ergonomics, stress and poor posture, and by repeating the same sequence of movements many times an hour or day.

Figure 4.5: Sentences 14A and 14B are fused into 14C.

4.4 Revision

In the *extraction* stage of summarization, passages are extracted from text documents. This inevitably leads to a lack (or change) of context when these passages are (re-)used in a summary. Revision can help recreate coherence in the summary. Another reason to revise is to increase the information density by removing unnecessary parts of sentences. Revision techniques Jing and McKeown identified in manual abstracting (see section 4.1) also apply to automatic summarization:

Sentence reduction: Text can be condensed by omitting the least salient sentences. It can be condensed even further by leaving out the less important words, i.e. by sentence reduction.. A way of doing this automatically is to create a dependency analysis for a sentence, and stripping parts of the tree for which is heuristically determined that they are unlikely to contain necessary information (Jing, 2000; Hovy et al., 2005a).

Sentence fusion: If two summary sentences are closely related, they may be merged into one sentence in order to indicate this relation (Jing and McKeown, 1999), but also to remove redundancy (Barzilay, 2003; Marsi and Krahmer, 2005). For instance, if a user asks for causes of RSI, the query-based summary may contain a number of answers, such 14A and 14B in Figure 4.5. This may be merged into the single, shorter sentence 14C without losing information.

Syntactic transformation and lexical paraphrasing: In NLP, paraphrasing is used to detect equivalence relations, but to my knowledge no attempts have been made to use paraphrasing as a revision technique in order to improve summary presentations.

Generalization and specification: If a sentence from a text is re-used in a summary, its meaning as well as the reader's capability to understand it may change due to the change of context. Examples are anaphora whose referent is excluded from

the summary, co-references which cannot be correctly interpreted because their context has changed, or relative expressions of time referring to the source text's publication date, e.g. 'yesterday', 'Bill'. This can be solved if the summarizer resolves these references and replaces them by absolute expressions, e.g. '10 April 2007' or 'a Microsoft spokesman'. The reverse is also being applied in text summarization. To increase variation in text, absolute expressions may be replaced by referring expressions.

References to complex concepts introduced in the text are more difficult to resolve, such as 'this' referring to the previous paragraph. Current-day summarization systems solve this problem mostly by avoiding it during the *selection* phase. This can be done by penalizing the appearance of such expressions, or by preferring longer sentences. By ratio, longer sentences tend to contain less unresolvable references.

Reordering: In single-document summarization, machine understanding of text is usually not good enough to outperform the original ordering in which the sentences appeared in the source. Following the original presentation is not possible in multi-document summarization, but since most summarization systems summarize (dated) newspaper articles, it is common practice to present content from different documents in the order of publication time. The use of alternative ordering methods has been studied by (Barzilay et al., 2002; Lapata, 2003; Madnani et al., 2007).

4.5 Conclusion

Summarization systems extract content from the source text. Most systems apply only minor revision or present extracted content verbatim. Research showed that professional human summarizers also use this strategy.

For content selection, a variety of methods have been proposed. Most systems select particular sentences by the presence of salient words. Individual words are salient for instance if they belong to a priorly defined set of cue phrases (such as *important*, *in sum*), to a set of words which are descriptive for the source, or to the set of query terms (in case of query-based summarization). Alternative methods exploit cohesion, coherence, or layout.

Evaluation methods are divided into three categories. Content-based methods intend to measure how successful a summarization system is in extracting relevant con-

tent. Usually, this is done by comparing generated candidate summaries with hand written reference summaries. Other methods measure the linguistic quality of a summary, including aspects such as grammaticality and coherence. Finally, utility-oriented methods measure the utility of a summarization system. This involves designing an evaluation environment which resembles a realistic use case as closely as possible. Then, the effect of the summarization system on this environment is measured. A disadvantage of utility-oriented methods is that experiments are expensive to carry out, as they typically require human participants.

5

The role of discourse in summarization

Even if a question can be answered with a concise and precise answer, it pays off to be more verbose. Given an answer pinpointed in text by a question answering machine, related content can be found using the layout or the discourse structure of the text. This chapter shows by means of a user study that both of these methods help the user to verify that the question was correctly 'understood'. In addition, the discourse-based method also improved the relevance of the presented information. Although hand-crafted Rhetorical Structure Theory analyses as used in this study limits applicability in automated systems, results show that discourse relations are relevant for query-based summarization.

Issues in query-based summarization have been addressed in question answering (QA) and generic summarization. Much can be gained by integrating existing techniques from these areas (Strzalkowski et al., 2000; Mori et al., 2004). A question answering system works by pinpointing an *answer* to a user-provided question in a set of documents. A *response* is then generated for this answer, and presented to the user (c.f. Hirschman and Gaizauskas, 2001). Making a distinction between *answer* and *response* makes it possible to view question answering and summarization as different tasks of query-based summarization. A question answering system locates the sentence which best matches the question, the *answer*, in a corpus of text documents. What remains is the task of generating an appropriate response and presenting it to the user, for which summarization techniques can be employed.

Question answering systems traditionally try to find and present an 'exact answer'. This is also the focus of large-scale question answering evaluation programs such as TREC (Voorhees and Tice, 2000). Although what is 'complete' may be subject to discussion, Voorhees (2002) defines an exact answer as *a text string consisting of a complete answer and nothing else*. Strings that contain a correct answer with additional text are considered 'inexact'.

Studies have shown, however, that users appreciate receiving more information than *only* the exact answer (Strzalkowski et al., 2000). The user may not only be interested in the answer to the question, but also in related information. The 'exact answer approach' fails to show leads to related information that might also be of interest to the user. Bakshi et al. (2003a) show that when searching for information, increasing the amount of text returned to users significantly reduces the number of queries that they pose to the system, suggesting that users utilize related information from supporting text.

Bakshi et al. (2003a) did not specify *how* context is put to use by users. Does context contain information that contributes to the answer? Is it used to answer (implicit) follow-up questions? Or is the context simply used to verify that the question has been answered adequately? In this chapter, a user experiment is set up to find out which summarization strategies help the user, and how.

Both commercial and academic QA systems tend to present more to the user than only the exact answer, but the sophistication of their responses varies from system to system. There are three degrees of sophistication in response generation: giving the exact answer, giving the answer plus context, and giving an extensive answer. The first is the most basic form of answer presentation. The second includes text surrounding the exact answer as well, which may allow the user to assess the accuracy of the answer extraction, and thus to verify whether the answer is correct (Bakshi et al., 2003a). The extensive answer approach aims at not just including the immediate context, but generating a response in a more intelligent way, aiming at optimizing the amount of useful information while maintaining verifiability. Thus, three types of answers can be distinguished:

Concise answer. The most basic form of answer presentation is to present only an exact answer. For instance, an exact answer to the question "what is the cause of RSI?" could be:

movements [which] involve repetitive contraction of the same muscles.

Answer plus context. If only an exact answer is provided, users have great difficulty assessing whether the answer matches the question, and thus whether the answer is correct. If the user is provided with more context (i.e. surrounding text), s/he will exploit this in order to find out whether the answer is indeed an answer to the question (Bakshi et al., 2003a). Most of the current QA systems follow this approach, and return not only the answer but also part of the surrounding text, in which the answer itself may be highlighted. This can be a few lines of text, or only the single sentence in which the answer occurs. For instance, the response to the question about RSI causes could consist of the answer, the preceding sentence and the sentence following the answer sentence:

Despite the fewer working hours, the same amount of work had to be done. A possible explanation of the development of RSI as a result of repeated low-exertion movements is that **the movements involve repetitive contraction of the same muscles.** *This happens for instance when working with a display device.*

Extensive answer. Bakshi et al. (2003a) have shown that users prefer to receive more information than only an exact answer, but simply returning to the user a particular quantity of surrounding text is likely to produce incoherent results. Furthermore, the surrounding text may include irrelevant information or unnecessary details. In the example answer plus context above, this is illustrated by the first sentence which is a poor introduction to the actual answer. Although – similarly to an answer plus context – an extensive answer includes more information than just the exact answer, the difference is that the extensive answer approach specifically aims at producing a coherent response that includes, apart from the answer, also related information which might interest the user. For instance, an extensive answer to the question about RSI causes could be:

A possible explanation of the development of RSI as a result of repeated low-exertion movements is that **the movements involve repetitive contraction of the same muscles.** *This happens for instance when working with a display device. Eventually they can cease to function and the muscle will lose power.*

In this chapter, extensive answers are generated by means of query-based summarization. The use of summarization for formulating answers is based on the following hypotheses:

- presenting more of the original document allows the user to verify whether the answer matches the question;
- summarization techniques can help increasing informativeness of the (long) answer, without weakening the user's opportunity to verify.

This chapter describes a semi-automatic query-based alternative to the RST-based summarization method for generic summarization of Marcu (1997a). Furthermore, a user experiment is designed and performed in order to test the abovementioned hypotheses.

5.1 RST-based summarization

RST has already been used to facilitate summarization (Marcu, 1997a). In his summarization effort, Marcu used the nuclearity of relations in the rhetorical structure to determine which sentence is more salient, but he also explored other features as additional indicators of importance, such as sentence length (Marcu, 1997a, 1998). This section shows how a graph-based algorithm for generating query-based summaries can be used with the same RST features as the algorithms of Marcu, which are intended for generic summarization.

I take a two-step approach to query-based summarization. First, the relations between sentences are defined in a discourse graph. Then, this graph is used to perform the summarization. During the first step, the rhetorical structure is transformed into a graph representation. The second step exploits a graph search algorithm in order to extract the most salient sentences from the graph. The starting node of the search is the node representing the answer segment.

The summary should consist of the most salient sentences, given the answer segment. This can be realized by determining the *distance* between the answer segment and each of the other sentences. The sentences which are most closely related to the answer segment are included in the summary. The distance between sentences is measured by their distance in the RST graph. In the graph construction, and therefore in the distance calculation, the presence of RST relations are the main factor. Details are

described later. RST defines relations between two spans of text, which can be used to derive the distance from one sentence to another.

The most nuclear sentence of an RST analysis is the sentence which is most central to the writer's purpose. The strategy for generic summarization of Marcu uses this knowledge by starting with the most nuclear sentence of the source text. A minimal summary created by following the summarization method of Marcu consists of only the most nuclear sentence. Similarly to Marcu's approach, my graph-based method prefers a nucleus over a satellite: in both summarization approaches, a satellite cannot be included in a generic summary without its nucleus. The consequence is that in the specific case that the entry point of the summarization – the answer segment – is the most nuclear sentence in the RST analysis, the result resembles the result of the summarization approach by Marcu (1997a) (depending on the summarization parameters). However, the graph-based approach is more general in the sense that summarization can start from any specific sentence rather than only the most nuclear sentence of the analysis.

Allowing an arbitrary (answer) segment to be used as an entry point for summarization raises new questions. For instance, if the answer segment is not the most nuclear sentence, should its nucleus also be included in the summary, or just its satellites? The answer segment's nuclei may provide general background information, while its satellites provide more specific information in relation to the answer. This section focuses on composing a summary of an answer and answer-supporting content, i.e. satellites of the answer. Later, I will discuss the possibility of including also more general content.

5.1.1 RST analyses as graphs

Although RST is not designed as a computational framework, it is relatively straightforward to derive a graph from a rhetorical structure. Graph theory is very suitable for this purpose. A rhetorical structure tree can be converted to a discourse graph by means of the following steps.

1. For each segment in the rhetorical structure, create a vertex associated with it.
2. For each subordinating relation, create an edge from the *nuclear segments* to the *satellite segments*.

A segment is a nuclear segment of a text span if it is not part of any sub span (of the text span) which participates as a satellite in a subordinating relation with any other

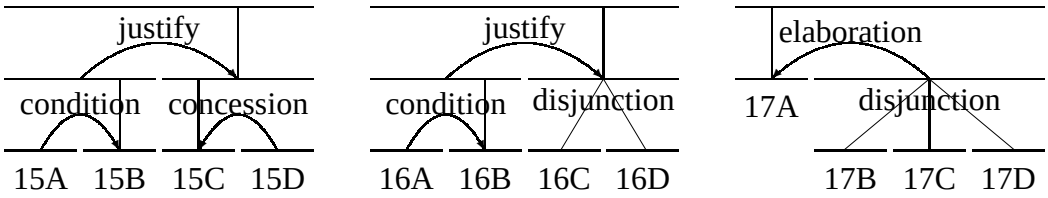

Figure 5.1: *Left:* a rhetorical structure analysis containing just subordinating relations; *center, right:* rhetorical structure analyses containing coordinating (multinuclear) relations. The symbols 15A, 15B, etc. refer to text segments (i.e. clauses or sentences).

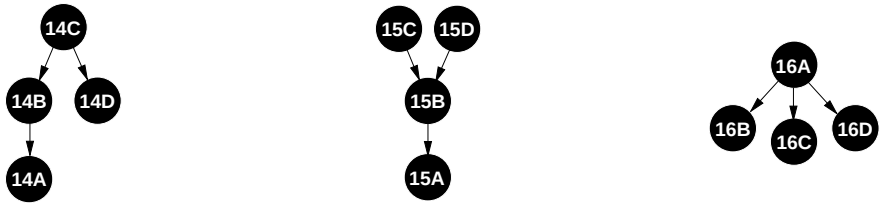

Figure 5.2: Discourse graphs corresponding to the rhetorical structures in Figure 5.1.

subspan. The satellite segments are the nuclear segments of the satellite. A text span can have multiple nuclear segments if multinuclear relations are involved. For instance, in the RST diagram on the left in Figure 5.1, the set of nuclear segments of the entire document (denoted as 15A:15D) contains only sentence 15C. The nuclear segment of 15A:15B is 15B. The middle diagram shows a rhetorical structure in which the set of nuclear sentences of 16A:16D consists of segments 16C and 16D.

The result of the transformation is an a-cyclic directed graph of which the vertices correspond to segments, and the edges define relations between them. Figure 5.2 shows the discourse graphs created for the rhetorical structures in Figure 5.1. During the transformation from RST to graph, part of the structural information is lost because segments of the graph are directly connected to other segments, while in RST, one end of a relation can also span more than one segment. If in RST one segment is related to a text span of two segments, the graph construction algorithm connects it to the nucleus of the two segments in the discourse graph. In practice, this means that if the inclusion of a segment in a summary is justified by a rhetorical relation, the nucleus of that relation must be included in the summary as well. This is in line with the role of subordinating relations in RST as defined by (Mann and Thompson, 1988), who state that a nucleus has meaning without the satellite, but not the other way round.

In the case of a multinuclear relation, each of the segments participating in the multinuclear relation (Figure 5.2, right: segments 17B, 17C and 17D) is connected

with the nucleus of the multinuclear span. That is, in the example, segment 17A is connected to each of the segments 17B, 17C and 17D, but segments 17B–17D are not directly mutually connected. The reason for this is that in terms of RST, there is a (multinuclear) relation between the segments 17B, 17C and 17D, but they are mutually independent: if we know that 17B contains relevant information in a particular context, there is no way to be sure that, to any extent, 17C or 17D is relevant as well, based on the relevance of 17B and the multinuclear relation between the three segments.

Now that we have a discourse graph T, we assume that given two segments $a, b \in T$ for which there is a path from a to b, we can say that they are related, and therefore if a is relevant to the answer, b is also relevant to the answer. If a path contains more than one edge, the segments are related only indirectly and an indirect relation is weaker than a direct relation between two segments.

The strength of a relation between two segments could be calculated by just counting the number of edges in the path between the vertices of the segments. However, it may be the case that there is more than one segment with an equally long path to the starting point of the summarization. In that case, the two segments would be equally likely to be included in the summary, although there might be other indications of one segment being better suited for inclusion in the summarization than the other.

In order to remedy this situation, we can assign a cost to edges in the discourse graph. A greater rhetorical distance is reflected by a greater cost. A cheap path from a to b indicates a high probability that b is relevant, given that a is relevant. The total cost of the path from a to b is denoted as $cost(a,b)$. The cost of a path between two segments is defined iff there is a path that connects them. The cost of a path is the sum of the costs of the edges in the path.

Given the entry point of the summarization (the answer segment), the shortest path from this segment to any other segment defines the relevance of the other segment to the final answer.

5.1.2 Determining costs

The distance between two segments is affected by the costs of the edges that connect the nodes corresponding to the segments in the discourse graph. These costs are determined by using features of the rhetorical structure from which the graph was created, such as features of the text spans on either side of the relation for which the edge was created. The cost of an edge also depends on features of the segment corresponding

to the vertex which is targeted by the edge. The only constraint is that all costs are non-negative.

The rhetorical structure has many features that may be relevant for determining costs to edges or vertices. Currently, only three features are considered when assigning costs. For these features, there is at least some evidence that they can contribute to the quality of a summarization. Further research may motivate the use of other features as well. The following features are considered, in order of relative importance.

1. Each edge has a basic cost, which is the same for all edges in the graph. This makes the distinction between directly and indirectly related sentences explicit. Two sentences are less closely related if the path that connects them consists of more edges.

2. For each edge, a cost is added depending on the number of sentences in the satellite of the corresponding rhetorical relation. If a particular satellite contains more sentences than another satellite of the same nucleus, the author apparently spent more words on it, which may indicate that the author finds this topic more important than a shorter one, although they both are a satellite of the same nucleus.

3. For each vertex, a cost is added depending on the number of words in the sentence. According to Marcu (1998), this is a good measure for the amount of new information contained in the sentence.

The cost of edge e from n to s is calculated as follows:

$$w_e = \alpha + \beta \cdot \sqrt{\|S\| + 1}^{-1} + \gamma \cdot log(\|s\|) \tag{5.1}$$

where $\|S\|$ is the number of sentences in the satellite of the relation; and $\|s\|$ is the number of words in the sentence s. Two subsequent sentences have a linear distance of one. The constants α, β and γ are used to balance the three factors of the distance between two sentences. Their values reflect that the number of edges (the α factor) is more important than the number of sentences in the satellite (the β factor), and the number of sentences in the satellite is more important than the number of words (the factor γ). In experiments described in this chapter, the value of β is chosen so small that it is just used to decide in favor of one sentence in case $\beta = 0$, $\gamma = 0$ would result in a tie. Similarly, the value of γ is chosen so small that the number of words just makes a difference when other factors do not.

18A Heavy workload, stress and repeatedly carrying out the same movements for a long period of time are the most important causes of RSI.
18B In the Netherlands the work pressure has increased by approximately 1.5 percent per year.

18C This is the result of shorter working hours in the eighties and nineties of the twentieth century.
18D Despite the fewer working hours, the same amount of work had to be done.
18E A possible explanation of the development of RSI as a result of repeated low-exertion movements is that the movements involve repetitive contraction of the same muscles.
18F This happens for instance when working with a display device.
18G The motorial units can be damaged due to lack of oxygen and difficulty in disposing waste products.
18H Eventually they can cease to function and the muscle will lose power.
18I There is however also evidence that the complaints do not arise from damaged muscles.
18J Instead, they supposedly arise from abnormalities in the response of the brain to signals from the muscles.
18K Another possibility is that psychological factors can lead to symptoms of RSI.

Figure 5.3: A text about RSI (translated from Dutch), used for query-based summarization.

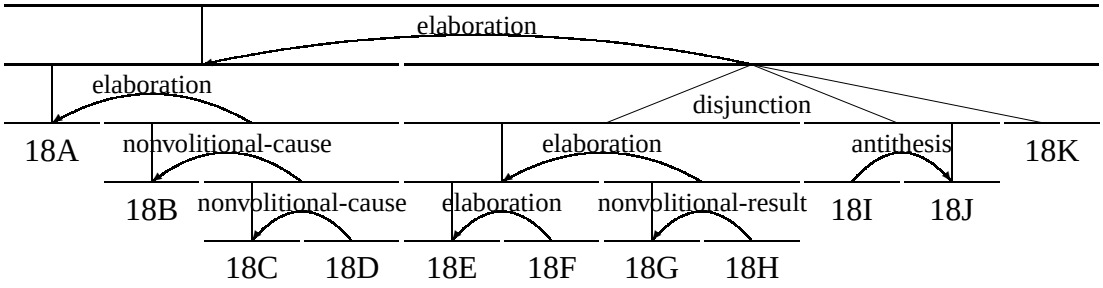

Figure 5.4: Rhetorical structure tree of the text fragment in Figure 5.3.

5.1.3 An Example

This example shows how three segments can be extracted from the source text in Figure 5.3, based on its RST analysis, and given the entry point of the summarization. In a QA context, the entry point would be the answer segment. Two of the extracted segments are direct or indirect satellites of the answer segment, the third is the answer segment itself. An RST analysis of the text in Figure 5.3 is shown in Figure 5.4. The entry point for extraction is segment 18E. This segment could for instance be the QA output for the question: *what can be the cause of RSI?*

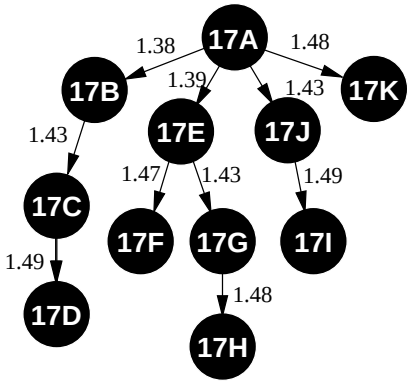

Figure 5.5: Weighted rhetorical structure graph of the text fragment in Figure 5.4. The vertex labels refer to their corresponding sentences. Edges are labeled by the distance they represent, and are calculated with parameters $\alpha = 1$; $\beta = 0.5$; $\gamma = 0.05$.

Figure 5.6: Extraction graph of the three sentences selected for inclusion in the summary, and the corresponding structure in RST notation, which is derived from the original RST analysis.

First, a discourse graph is created from an RST analysis (as shown in Figure 5.5). For this graph, the total cost of the paths from sentence 18E to each sentence in the graph is calculated using Dijkstra's shortest path algorithm (Dijkstra, 1959). A path in a graph is an alternating sequence of vertices and edges, beginning and ending with a vertex. For instance, in the graph of Figure 5.5, there is a path over three vertices and two edges from 18E to 18H. The cost of this path is the sum of the costs of all of its edges.

Only four sentences are reachable from 18E. Since the selection of sentences is based on the cost of their path from 18E, a sentence which is associated with an unreachable vertex cannot be included in the extract. In this case, the sentences with the cheapest path from the entry point 18E are selected. The selected sentences are extracted, resulting in the discourse graph on the left in Figure 5.6. For the sentences in this graph, the rhetorical structure can be derived using the original RST analysis in Figure 5.4. The result is the rhetorical structure on the right in Figure 5.6. This

rhetorical structure may be used for further processing. The output of the extraction process would be the following text. The answer segment is highlighted.

> **A possible explanation of the development of RSI as a result of repeated low-exertion movements is that the movements involve repetitive contraction of the same muscles.** This happens for instance when working with a display device. The motorial units can be damaged due to lack of oxygen and difficulty in disposing waste products.

5.2 Evaluation

5.2.1 The data

The utility of discourse structure in summarization is measured by means of an online experiment. Material for the experiment was extracted from an RST-annotated corpus of news releases in English (Carlson et al., 2002). The corpus[1] contains 385 RST-annotated text documents with a total of 21789 discourse units and 176,383 words. The text is segmented and annotated using the guidelines of Carlson et al. (2001). These guidelines build on RST but deviate from Mann and Thompson (1988) at some points. The segmentation into discourse units as well as the taxonomy of discourse relations is more fine-grained than described by Mann and Thompson (1988). Using the style analysis tool of Cherry and Vesterman (1981), I counted 8056 sentences in the corpus, which amounts to an average of 2.7 discourse units per sentence.

For 30 documents, the corpus contains between 3 and 26 questions to which the document has an answer, 250 in sum. Not all of the questions could be used because the interpretation of some of the questions rely on the previous questions or their answers, or even on the text document that contains the answer. Only self-contained questions are considered. For example, the question *what will the toilet cost* cannot be properly interpreted because the antecedent of the referential expression *the toilet* cannot be resolved without a proper context. Other questions, such as *what has been recovered?* are so general that they need more context to reduce ambiguity. The self-contained questions were matched with a discourse unit by manually selecting the discourse unit with the most direct answer to the question. If no discourse unit could be selected, the question was discarded, as the experiment concentrates on situations

[1]The corpus can be obtained from the LDC (Linguistic Data Consortium, `http://www.ldc.upenn.edu/`) under catalog number LDC2002T07.

where answer extraction is straight forward once an answer is found. The answer presentation approach taken presumes that a 'correct' answer is pinpointed in the text. Furthermore, not for all questions was it possible to create a summary of the desired length. These questions were also discarded.

From the 30 documents, 4 were randomly selected. After filtering as described above, 12 self-contained questions which had answers remained. For each of these questions, four answers were formulated:

Concise answer: the *core segment*, the discourse unit which most directly answers the question.

Answer plus context: a query-based summary of a maximum of 160 characters. The answer plus context consists of the core segment, the immediately preceding discourse unit, and one or more discourse units immediately following the core segment in the linear flow of text, up to 160 characters.

Extensive answer I: a query-based summary of a maximum of 160 characters, created using the summarization method described in section 5.1.

Extensive answer II: a query-based summary of a maximum of 160 characters. This summary consists of the core segment, its immediate superordinated discourse unit, and one or more discourse units which are selected by means of the summarization method described in section 5.1, up to 160 characters. This summary is added in order to measure the effect of presenting general information in a summary, rather than specific information.

5.2.2 Manual postprocessing

As mentioned previously, the four answers are compiled by selecting a number of discourse segments from the RST analysis of the document containing the answer. Because sentences may consist of multiple discourse units and the unit of summarization is the discourse unit, the answers may contain incomplete sentences. In addition, they may contain anaphora or referential expressions to other parts of the text. Because participants are asked to judge *content* rather than readability, minimal revisions are made to the answers in order to avoid that problems with readability cause artifacts in

the outcome of the experiment. The answers were revised by following a number of rules:

1. If a selected discourse unit participates in a SAME-UNIT relation, the related units are selected as well. The SAME-UNIT relation is not a rhetorical relation, but is a special relation used in the RST corpus to indicate that two physically separated text fragments are actually one discourse unit.

2. If one or more selected discourse units cannot be interpreted due to lack of context, starting from the unit closest (within the summarization model) to the core:

   - As long as the 160 character maximum length is not exceeded, and the discourse unit is not interpretable, contiguous units are added.
   - If no more discourse units can be added due to the 160 character boundary, the uninterpretable discourse unit is discarded.

3. Arrange the selected discourse units in the original order of the text.

4. Make minor revisions to enhance the fluency of the text and to make sure the text is well-formed.

5. Resolve anaphora and substitute their antecedents, unless the antecedent contains a verb.

The result is four fluent answers for each of twelve questions. An overview of all questions and answers used in the experiment is shown in appendix A.

5.2.3 Experimental setup

Sixteen people between the age of 24 and 58 participated in the experiment – ten male, six female – all in good command of the English language. Some of them were computer linguists, but none of them had prior knowledge of the experiment. All participants were recruited from e-mail lists. The experiment was presented to the participants as a question answering experiment, in which they had to rate answers for various qualities.

The experiment consisted of three stages. First, the participants entered some personal information such as age, gender, native language. Then, in stage 2, the participants were shown each of the 12 questions in a random order. Each question was presented along with an answer of (at random) one of the four answer types, so that

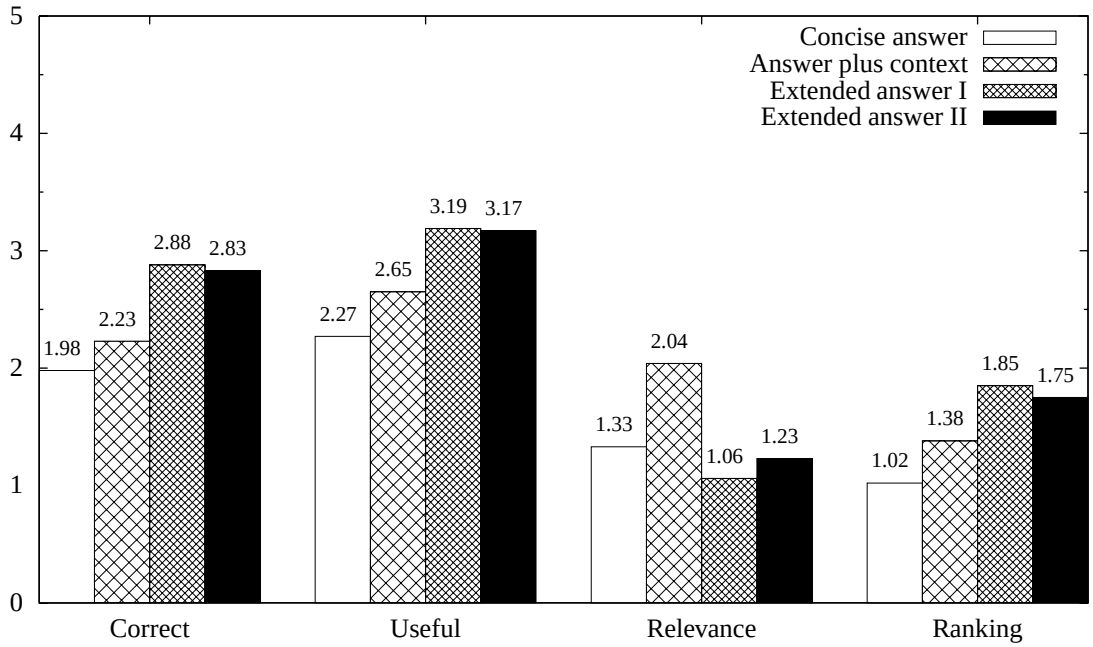

Figure 5.7: Average ratings of the four answer presentations.

each participant assessed three answers of each of the four types, but not every user assessed the same answer to the same question. This amounts to $3 \cdot 16 = 48$ assessments for each type of answer. The subjects rated the quality of the summaries on three dimensions. First, the user was asked to indicate on a 5-point Likert scale to what extent s/he was able to verify whether the answer was accurate. Secondly, the user was to judge how useful the provided information was with respect to the question. And finally, the user was asked how much irrelevant information was contained in the answer. On each occasion, the answer presentation received a score between 0 and 4 (from worst to best).

In stage 3, when the participants finished assessing answers, they were shown the same questions again in the same order. This time, they were shown all four answers to the question and asked to order them from best to worst. The participants were free to use any criteria for determining the quality of the answers – no criteria were specified. The best ranked presentation receives a score of 3; the worst 0.

5.2.4 Results

The average ratings assigned by the subjects are shown in Figure 5.7. A t-test performed on the results showed that the RST-based summaries were judged to be signifi-

cantly more verifiable than the answers plus context ($p < 0.01$). One of the participants remarked, *I would like to see contextual information of the answer to know whether the question was understood, especially because I have no clue what the correct answer is*, thereby supporting the claim that verification of the answer is an important function of a summary. Furthermore, RST-based summaries contained significantly less irrelevant information than the answer plus context ($p < 0.01$). Despite this, there was no significant difference in the amount of useful information between RST-based summaries and answers plus context. Also, the results of the two types of RST-based summaries were very similar.

The statistics of the ranks assigned to answer presentations by the participants show a clear improvement with more sophisticated presentation methods. On average, the answer plus context was given a significantly higher ($p < 0.01$) rank than the concise answer. The extensive answers are ranked significantly higher ($p < 0.01$) than the answer plus context. No significant differences were found between the rank averages of extensive answers I and II.

The results suggest that RST-based summarization compares favorably to generating an answer presentation simply by including the answer segment and surrounding sentences: using RST helps reducing the amount of irrelevant information, and increases the verifiability of the answer.

5.3 Conclusion

This chapter is a first attempt to exploit discourse structure in query-bsaed summarization, and to explore what discourse structure can contribute. I proposed graph-based framework for query-based summarization which is implemented to use an RST analysis of the source text, but it can be easily extended to use other features of the text. The summarization framework aims at producing coherent summaries containing an answer to the query, but also information supporting this answer. Previous work on query-based summarization has mainly focused on extracting the set of sentences which best match the query, not on producing coherent summaries or providing information not explicitly asked for.

A user study indicates that this method outperforms a competitive baseline with respect to the amount of relevant information included in the summary. Furthermore, when users were asked to rank answer presentations of different types, the RST-based summaries ranked significantly higher than the baseline summaries and the concise an-

swers. The user study also confirmed that users are more capable of verifying whether the summary is actually responsive to the query when they are given a summary than when they receive a concise answer. However, with respect to verifiability, there is no great difference between a simple baseline summary (answer plus context) and an extensive answer which is composed using RST.

6

Graph search algorithms for summarization

This chapter presents a graph based framework which is used for automatic discourse oriented text summarization. Increasingly sophisticated content selection algorithms are evaluated using automatic quality measures. A simple but competitive baseline (which selects the sentences most similar to the query) appeared hard to beat: only one method significantly outperformed this baseline ($p < 0.1$, DUC 2005 data). On the other hand, this method would have ranked first (Rouge-2) or second (Rouge-SU4) if it had participated in DUC 2005. It should be noted that automatic measures provide just a partial evaluation, as they are notoriously poor at measuring some aspects of summarization, such as readability and linguistic quality. A full evaluation in the context of DUC 2006 using manual and automatic quality measures provides a more complete picture on performance of one implementation of the summarization framework. While showing average results in content based measures, the discourse oriented summarization method presented here beat (all but one of the) competing summarization methods when it comes to readability.

Summarization systems typically extract content by applying a ranking and selecting the highest ranking passages. The passage ranking is a result of a relevance assessment of each sentence individually. This is based on the idea that some content is more relevant to the user than other content. However, this is contrary to theories of discourse organization which claim that meaning is tightly related to discourse organization. If a

passage is presented without its original context, then its meaning may change or even disappear.

19A RSI can be cured,
19B provided you're not too late.

In the text of 19A–19B, information in the second passage imposes a condition to what is said in the first. Presenting the first passage without this condition is misleading. In other words, if 19A is considered sufficiently relevant to include in a summary, the relevance of 19B is justified by their conditional relation.

Inspired by the RST based approach to generic summarization of Marcu (1999), the previous chapter showed how coherence structure can be used for query-based summarization. Specific types of RST relations are well detectable automatically, but unfortunately, it remains difficult to obtain a full automatic analysis in a reliable manner. As an alternative, Barzilay and Elhadad (1997) used lexical cohesion as evidence of semantic relations in text. Their approach relied on semantic similarity of words rather than intentional relations between passages, which is easier to detect automatically.

The generic multi-document summarization system of Erkan and Radev (2004) exhibits similarities with both the coherence based approach of chapter 5 and the cohesion based approach of Barzilay and Elhadad. Both Barzilay and Elhadad and Erkan and Radev use lexical repetition to detect relatedness between sentences, but similarly to chapter 5, Erkan and Radev use a graph based summarization system. In both chapter 5 and in Erkan and Radev (2004), graphs are used to express conditional relevance of text passages, and a graph search algorithm is used to find relevant sentences, but Erkan and Radev used those graphs for producing generic summaries. While chapter 5 uses query distance as a measure of relevance, Erkan and Radev use 'centrality' as a measure of relevance. According to Erkan and Radev (2004), 'central' (and thus salient) sentences are sentences which are related to many other sentences of high centrality. Because no prior amount of salience is attributed to sentences, salience is updated iteratively from the structure of the graph (see also section 4.3.5.3).

Otterbacher et al. (2005) brought the graph based approach of Erkan and Radev to query-based summarization. Both Erkan and Radev (2004) and Otterbacher et al. (2005) used centrality in multi-document summarization, but neither take redundancy into account. More generally, they do not make an attempt to combine evidence for determining salience.

The work in this chapter is based on three observations:

1. concepts that play a role in summarization include query-relevance (in query-based summarization) (Mani and Bloedorn, 1997), redundancy (Carbonell and Goldstein, 1998), and coherence (Blair-Goldensohn and McKeown, 2006);

2. intuitively, it seems plausible that these concepts demand employing different types of information and that these types of information should be combined to determine the salience of passages;

3. to my knowledge, no summarization system exists which integrates query-relevance, redundancy as well as coherence for determining salience of a passage.

This chapter investigates what is required for content selection in generation of high-quality summaries. The next section decomposes the task of summarization into several sub tasks. This paves the way for experiments with a query-based multi-document summarization system. Two aspects of content selection are investigated: graph search algorithms for determining sentence salience, and combinations of features for graph construction. The graph search algorithms include a measure of query distance based on Bosma (2005c), and measures of centrality based on Erkan and Radev (2004). Features include a graph to express relations between sentences of the same document based on cosine similarity, and a graph to express redundancy, also based on cosine similarity. For measuring the quality of generated summaries, I use data and automatic methods of DUC 2006. Section 6.3 describes an implementation of the framework using query distance to determine salience, utilizing a combination of feature graphs, and its evaluation in the DUC 2006 query-based summarization challenge.

6.1 A framework for summarization

This chapter aims to investigate the content selection sub task of summarization. Nonetheless, the evaluation methods used are designed to measure the quality of abstracts, and require a full summarization system. For the purpose of these experiments, I decomposed the summarization process in a number of sub tasks. This system presents a framework which allocates these sub tasks to different components, which allows for substituting part of the system while leaving everything else unchanged. Separate modules are responsible for the following sub tasks.

Segmentation. The source documents as well as the query are segmented into *content units*. A content unit is an undividable text passage which is candidate for inclusion in the summary. Experiments in this chapter involved only one implementation of this component which uses sentences as content units. Alternative implementations may perform paragraph segmentation or clause segmentation.

For each segment, a unique identifier is generated which is composed of the document name, the paragraph number and the sentence number. For instance, the second sentence of the eighth paragraph of document `D0601A_NYT20000228_0358` is labeled `D0601A_NYT20000228_0358.8.2`. These data can be used in any later stage of the summarization process to identify the sentence but also for other purposes such as to restore the sentence ordering during *presentation*. The query undergoes the same treatment as any document. Because the source document of a sentence is kept, query sentences can be distinguished from candidate (non-query) sentences.

The sentence segmenter uses a set of punctuation rules to determine sentence boundaries. A list of abbreviations is maintained to avoid falsely detected boundaries. Paragraph boundaries are derived from annotations provided with the source documents. The segmenter also attempts to remove meta data from the text, such as the date and location of publication. The rules were based on seventy five documents of the DUC 2006 development set. These documents are representative for the test data but not used for evaluation. No formal evaluation of the segmenter was performed.

Feature extraction. The source text and the query are processed and converted to a feature graph to prepare for content selection. Multiple modules may be used in parallel so that multiple graphs are generated. This may include coherence analysis, measuring redundancy, etc. The generated graphs are integrated, as described later.

Salience estimation. This model runs an algorithm which derives a salience value for each sentence from the (possibly combined) feature graph.

Presentation. A summary is created using the most salient content units. The summaries used in this chapter are extracts consisting of the set of most salient sentences. If adding the next-salient sentence would cause the 250-word limit to be exceeded, no more sentences are added. Where possible, the linear ordering of the sentences in the source text is retained. If the summary contains sentences

from multiple sources documents, sentences from the document containing the largest number of sentences are presented first. Although the ordering of the sentences may be important for readability, it has little effect on Rouge scores.

6.2 Toward discourse-oriented summarization

The quality of summaries is measured using the Rouge performance metrics also used in DUC 2006. These are Rouge-2 (i.e. bigram recall with respect to reference summaries) and Rouge-SU4 (skip bigram recall). Both metrics are described in greater detail in section section 4.2.1.3. The DUC 2006 evaluations by NIST also included a Basic Elements recall measurement using the Basic Elements toolkit of Hovy et al. (2006), but unfortunately, this toolkit is no longer available. It should be born in mind that Rouge metrics only partially quantify the quality of summaries. Equally important is a qualitative discussion of a manual inspection of the summaries in relation to the behavior of the algorithms used.

6.2.1 The data

Experiments in this chapter use DUC 2006 data for query-based multi-document summarization evaluation. This data set consists of fifty 'topics'. A topic consists of (a) a title, (b) a query consisting of one or more sentences in the form of a question or an assignment for the summarizer to respond to, and (c) a set of source documents. An example of the title and a query of a DUC 2006 topic is given below (topic D0650E). Section B.1 (in appendix B) shows one of the reference summaries used for evaluation of summaries generated for this topic.

Title: former President Carter's international activities

Query: Describe former President Carter's international efforts including activities of the Carter Center.

In addition, 25 newspaper articles are associated with each topic. The summarization task is to respond to the topic in the form of a summary, using content from the set of 25 documents. The response may comprise at most 250 words. This task is given to professional human summarizers as well as automatic summarization systems. The human summaries are used as reference summaries for evaluating system summaries

(henceforth *candidate summaries*). Each DUC topic has four corresponding reference summaries.

The source text contained errors and pollution which has a greater or lesser impact on the quality of summaries, depending on the summarization method used. For instance, the text contain notes to editors, such as 'BEGIN OPTIONAL TRIM'. I encountered also inconsistencies in punctuation, such as variation in the use of quotations and dashes. The articles were marked up in XML. It also occurred at least once that errors in the file format caused an article to be almost completely lost because most of an article was written in the header of the source file. Some of the reference summaries have spelling errors, such as 'The U.N. Genral Assembly'. Some articles contain sections from different authors, ending with a single line containing only the author name. If one author is mentioned several times, a summarization algorithm may find this name important and include it in a summary, possibly resulting in a summary containing only author names. Two measures were taken to alleviate these problems. First, parenthesized sentences were ignored during summarization. This removed most of the editorial notes. Second, if a document contained identical sentences, all but one of the sentences are removed.

6.2.2 Pair-wise significance

The Rouge-2 and Rouge-SU4 metrics produce a score for each summary, representing its quality. Performance of summarization systems is often measured by averaging the scores of individual summaries. The average score over all summaries of particular system gives a general indication of the quality of the summarization method. To see if one system is better than another, one could simply check whether the difference in average scores is significant. However, while a single average score is useful as a rough indication of the quality of a system, it may not be the best method for a pair-wise comparison.

When can a system be said to be better than another system? If a summary receives twice the score of another summary, is the summary twice as good? Evaluation metrics such as Rouge are notoriously bad at *quantifying* differences in quality in absolute terms. Donaway et al. (2000) suggest that a system which consistently produces higher scoring summaries than another system may be regarded a 'better' system. Note that the average scores of both systems may be very similar, even if one system performed consistently better.

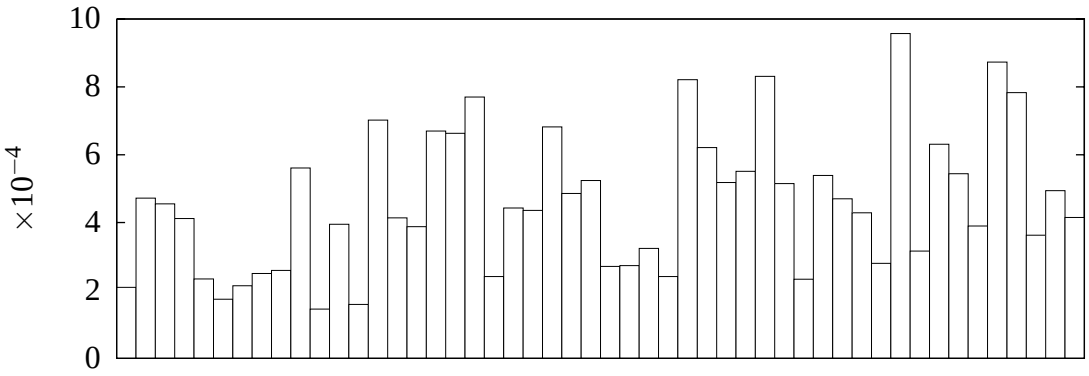

Figure 6.1: Standard deviations of Rouge-2 scores of participant systems of DUC 2006 for each of the 50 topics.

If the difference in scores between two systems does not tell anything about the difference in quality other than which one is better, a comparison can have only three possible outcomes: one is better, the other is better, equal quality. If this is accepted, it remains to be seen whether the system score averages are reliable for significance testing. The variability of scores for some topics is far greater than for other topics (see Figure 6.1). As a result, the topics with a high score variability affect the system average more than other topics. This makes it less likely to find significant differences in system performance. Therefore, I use not absolute values but the system rank in pair-wise comparisons.

When comparing two systems, I measure the percentage of topics for which one system beats the other. For each topic, the system which produced a higher quality summary than the other system (according to the metric of choice) receives a relative score of 1; the other receives a score of 0. In the event that both systems perform equally good, both systems receive 0. If the sum of scores is better for one system than for the other, its average rank is higher and the system performed better.

These data are used for significance testing by means of approximate randomization. The set of scores of both systems together (i.e. a set of values of 0 or 1) is randomly re-divided into two sets of scores. Then, we check whether the sums of scores in the random sets differ at least as much as the difference in the sum of scores of the systems under comparison. This is repeated one million times. The number of times the difference between the random sets of scores is indeed greater than the difference between systems corresponds to the probability that, if the null hypothesis holds, a performance difference occurs which is at least as great as the observed difference

between the two systems. The null hypothesis (that the performance of both systems is equal) can be rejected with probability p, defined as:

$$p < \frac{t+1}{N+1} \tag{6.1}$$

where t is the number of positive tests; and N is the total number of tests. Throughout this chapter, I reject a null hypothesis if $p < 0.05$, unless stated otherwise. Nonetheless, significance probabilities are reported where relevant.

If the system rank are preferable over absolute averages, why measure the system ranking for pair-wise comparison only? If a number of systems are compared, one could measure a system's rank for each topic, and take the average rank as a performance measure instead of the average score. An objection to this method is that the average rank of a system depends on which other systems it is compared to. When testing the significance of a difference between two systems, the outcome depends on which *other* systems are in the group of systems ranked.

In sum, I use the average scores as a rough indication of a system's performance; a binary system ranking is used for pair-wise significance testing.

6.2.3 Query-relevance

A simple form of query-based summarization is to determine sentence salience by measuring its cosine similarity with the query. The sentences most similar to the query are presented as a summary. A summarization system based on cosine similarity of a content unit and the query is a competitive baseline for query-based summarization. This method is compatible with the framework of section 6.1. The graph used for salience estimation is the graph where each candidate sentence is related to each query sentence, and the strength of this relation is the cosine similarity of the two sentences. The sentences closest to a query sentence are then included in the summary.

The cosine similarity graph is generated in three steps:

1. words of all sentences are stemmed using Porter's stemmer (Porter, 2001);
2. the inverse document frequency (IDF) is calculated for each word;
3. the cosine similarity of each candidate sentence and each query sentence is calculated using the $tf \cdot idf$ weighting scheme.

Stemming is a way to normalize syntactic variation. The inverse document frequency is used to weight words higher than other words if they occur in fewer sentences. Rare words are characteristic of the sentence they appear in.

This method for calculating IDF values appeared not appropriate for query terms because there is a mismatch between the language use in the query and in the source documents. For instance, queries frequently used phrases such as 'Discuss ...' or 'Describe ...'. These words have a low frequency in the source documents, and are thus assigned a high IDF value, but they are hardly descriptive if they appear in the query. Therefore, the IDF values for query terms are calculated from the set of sentences from all DUC 2006 queries instead of the source document sentences specific for the topic. The query-relevance graph is defined by a function determining the strength of the relation between two sentences:

$$\begin{aligned} \delta_q(i,j) &= cosim(i,j) && \text{, if } i \in Q;\ j \in S \\ \delta_q(i,j) &= 0 && \text{, otherwise} \end{aligned} \tag{6.2}$$

where $\delta_q(i,j)$ is the strength of the relation between sentences i and j; Q is the set of query sentences; S is the set of candidate sentences; $cosim(q,i)$ is the cosine similarity of sentences q and i. The strength of a relation is a value in the range of 0 (no relation) to 1 (a strong relation).

The query-relevance $R_{query\_relevance}(j)$ of a sentence j is then calculated as follows.

$$R_{query\_relevance}(j) = \min_{q \in Q} \delta_q(q,j) \tag{6.3}$$

where $R_{query\_relevance}(j)$ is the salience of sentence j; Q is the set of query sentences.

A summary is then generated from the most salient sentences. An example summary is shown in appendix B.2. The system achieves an average Rouge-2 score of 0.0818, and an average Rouge-SU4 score of 0.138. Compared to the scores of participant systems of DUC 2006, the query-relevance summarization system would rank 11^{th} of 36 in both Rouge-2 and Rouge-SU4 (in DUC, Rouge-2 and Rouge-SU4 was measured for 35 systems). This is well above the median, and remarkably high for a system this simple.

6.2.4 Query-distance

In an attempt to increase the coherence of the summaries, the role of query-relevance in summarization may be replaced by the more general concept of *query-distance*. A sentence is salient not only if it is relevant to the query, but also if there is indirect evidence of relevance, e.g. if it is similar to a query-relevant sentence of the same document. Similarity with sentences of other documents than the source document of

a sentence is not used as evidence of relevance, because similarity of sentences from different documents may as well indicate redundancy. Redundancy is to be avoided.

In order to realize query-distance as a salience measure, I measure cosine similarity not only between query and source document sentences, but also between two candidate sentences. The latter is to enable selection of sentences which are not directly query-relevant, but which are closely related to query-relevant sentences.

The query-relevance graph δ_q (eq. 6.2) is used again to express relevance to the query. A second graph is generated to express relatedness of sentences of same document:

$$\begin{aligned} \delta_c(i,j) &= cosim(i,j) && \text{, if } i,j \in S; doc(i) = doc(j) \\ \delta_c(i,j) &= 0 && \text{, otherwise} \end{aligned} \tag{6.4}$$

where $\delta_c(i,j)$ is the strength of the relation between sentences of the same document; $doc(i)$ is the source document of sentence i; S is the set of candidate sentences. Although $\delta_c(i,j) = \delta_c(j,i)$ for all i,j, the two relations in the graph are distinct because they have opposite directions. If there are more sentence pairs with a non-zero value of $\delta_c(i,j)$

The graphs δ_q and δ_c are integrated into a single multi-graph $\Delta_{q,c}$. A multi-graph is a graph that can have two edges between the same two vertices, expressing simultaneous relations. As a result, not a single relation but a set of relations hold between two sentences, and each relation may have a different strength between 0 and 1. The integrated graph is expressed as follows.

$$\Delta_{q,c}(i,j) = \left\{ w_q \delta_q(i,j), w_c \delta_c(i,j) \right\} \tag{6.5}$$

where $\Delta_{q,i}(i,j)$ is a set of values, each representing the strength of an edge from i to j in the multi-graph $\Delta_{q,c}$. The values of $w_q, w_c \in [0..1]$ are weighting factors. The smaller w_q and the greater w_c, the greater the relative importance of indirect evidence of relevance. The smaller w_q and the greater w_c, the more sentences are selected which are not directly query-relevant.

The query distance of a sentence i is calculated as the shortest path from a query sentence to i. A path between two sentences is a sequence of edges that connect them. The shortest path consists of edges from the graphs $\Delta_{q,c}$. Based on $\Delta_{q,c}$, the query distance is calculated as the follows:

$$D(j) = 0 \qquad \text{, if } j \in Q \tag{6.6}$$

$$D(j) = min\left\{ D(i) + r^{-1} - 1 \mid i \in Q \cup S; r \in \Delta_{q,c}(i,j) \right\} \qquad \text{, otherwise} \tag{6.7}$$

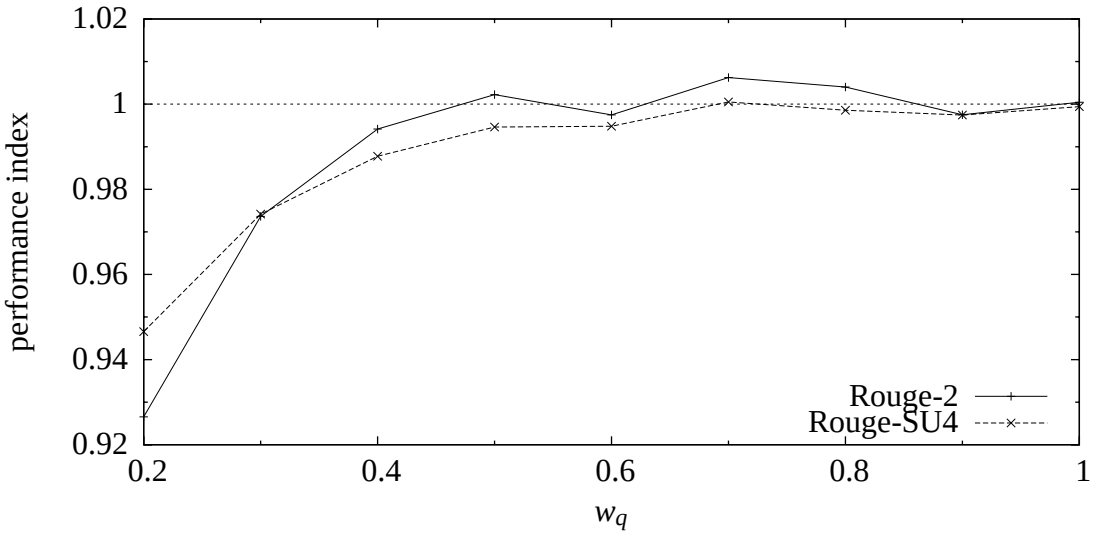

Figure 6.2: Indexed performance of query-distance summarization for different values of w_q. An indexed performance of 1 indicates the performance of the query-relevance system.

where $D(j)$ is the query-distance of sentence j. In each step of the path, an edge is followed; the strength of the relation represented by that edge is r. The distance of the step in the path is the reciprocal of the strength of the relation (expressed by $r^{-1}-1$). The distance of a path is the sum of the distances of its steps.

As all elements of $\Delta_{q,c}(i,j)$ are guaranteed to be in the range $[0..1]$, the term $r^{-1}-1$ is in the range $[0,\infty\rangle$, if defined. If the strength of a relation between two sentences is 1, their distance is 0; if strength is 0 (there is no relation), the distance is undefined unless there is an alternative path, in which case that path is taken.

Note that query-distance is here used as the reciprocal of salience. The salience as calculated using query-distance is given by

$$R_{query\_distance}(j) = 1 - D(j) \tag{6.8}$$

where $R_{query\_distance}(j)$ is the salience of sentence j. Like in the query-relevance summarization system, the most salient sentences are selected for inclusion in the summary.

Figure 6.2 plots the performance of the query-distance system for different values of w_q. In order to make results more comparable, Figure 6.2 shows the indexed performance rather than the actual performance. The performance index 1 is the baseline system, in this case the performance of the query-relevance system. Given $w_c = 1$, the highest performance of query-distance summarization is reached with

$w_q = 0.7$. At $w_q = 0.7$, the average quality of query-distance summaries is slightly higher than query-relevance summaries according to Rouge-2 (0.0823, DUC 2006 rank 11 of 36) and Rouge-SU4 (0.138, rank 11). This performance difference is not significant ($p < 0.12$). At $w_q \geq 0.6$, the query-distance summaries are very similar to query-relevance summaries.

Appendix B.3 contains an example of a query-distance summary with $w_q = 0.5$. Compared to the query-relevance summary in appendix B.2, the sentences 37G and 37J are added at the cost of removing 36C. The evidence of the relevance of the added sentences is based on their relation with 37H and 37I respectively.

6.2.5 Centrality

The summarization systems of Mani and Bloedorn (1997) and Erkan and Radev (2004) iteratively recalculate the salience of a sentence from a similarity graph and the salience of neighboring sentences. Erkan and Radev termed this method of calculating salience *centrality-based* summarization, because sentences which are more 'central' in the feature graph (i.e., sentences which have more neighbors) receive a higher salience value. The summarization systems of Mani and Bloedorn and Erkan and Radev are described in section 4.3.5.3.

I attempted to run both the algorithm of Mani and Bloedorn and the algorithm of Erkan and Radev using the same features (the graphs δ_q and δ_c) as used for query-distance summarization. Both algorithms are designed for generic summarization, so they had to be adapted for query-based summarization. Although Otterbacher et al. (2005) presented a query-based alternative to the generic summarization algorithm of Erkan and Radev (2004), I chose to adapt the latter algorithm for query-based summarization instead of using the query-based alternative of Otterbacher et al.. Otterbacher et al. included the concept of query-relevance in their salience computation rather than in a feature graph. This is not compatible with my graph based summarization framework. My implementation deviates from both Erkan and Radev (2004) and Otterbacher et al. (2005) in the sentences between which similarity is measured. Erkan and Radev measure cosine similarity between any two candidate sentences; I use the graph δ_c (eq. 6.4) which expresses similarity between sentences of the same document only.

The salience is calculated as follows.

1. Initiate the salience of all candidate sentences (source document sentences) at 0. The salience of query sentences is initiated at 1.

2. Recalculate the salience of each candidate sentence, using the feature graphs and the salience of neighboring (i.e. related) sentences. Salient sentences increase the salience of their neighbors.

3. Repeat step 2 unless the change in salience in the last iteration falls below a certain (pre-defined) threshold.

How salience is recalculated depends on the centrality algorithm used. The algorithm of Erkan and Radev (2004) uses normalization, so that a fixed 'amount of salience' is in the system: the sum of the salience values of all sentences to converge to a fixed value, e.g. to 1. This prevents the salience to keep increasing and never converge.

The algorithm of Mani and Bloedorn (1997) does not normalize salience, but uses a Sigmoid transformation to ensure that the salience value converges after sufficient iterations. After each iteration, the transformation maps the salience of each sentence to a value in the range $[0..1]$. The Sigmoid transformation works as a salience threshold: sentences above a certain salience value are assigned a value of 1. If the threshold is not set appropriately, all sentences may be receive a salience value of 1, or all sentences may receive a salience of 0. Chen and Ng (1995) suggest experimenting with different values until desired results are achieved. However, which threshold is appropriate depends on the topic and the source documents.

Instead of using the Sigmoid-based algorithm, I designed an algorithm which does not normalize and which does not work with thresholds. This algorithm is based on a probabilistic interpretation of semantic networks.

The following describes a normalized and a probabilistic approach to centrality-based summarization. Thereafter, centrality-based summarization is illustrated with an example.

6.2.5.1 Normalized centrality

The normalized centrality (based on Erkan and Radev, 2004) is calculated as follows:

$$\begin{aligned} \mu_j(t) &= 1 && \text{, if } j \in Q \\ \mu_j(0) &= 0 && \text{, if } j \in S \\ \mu_j(t+1) &= \frac{d}{\|D\|} + (1-d) \sum_{i \in D} \sum_{(r \in \Delta_{q,c}(i,j))} r \cdot \mu_i(t) \cdot degree(i)^{-1} && \text{, if } j \in S \end{aligned} \tag{6.9}$$

where $D = Q \cup S$; and $\mu_j(t)$ is the normalized centrality of sentence j at iteration $t \geq 0$; and $\Delta_{q,c}(i,j)$ is the set of edges between i and j (see eq. 6.5). The constant d is a small value Erkan and Radev used in generic summarization in order to guarantee a salience ranking under all circumstances by giving each sentence a small prior non-zero salience. Throughout this section, the value of 0.15 is used, as suggested by Erkan and Radev (2004). In contrast to query-based summarization, the graph weight factors w_q and w_c (eq. 6.5) do affect the salience of sentences in normalized centrality summarization, because the salience is normalized and the sets of sentences with outgoing edges in δ_q and δ_c are disjunct. The degree of a sentence i in the graph ($degree(i)$) is measured as the number of outgoing edges:

$$degree(i) = \sum_{k \in D} \sum_{(r \in \Delta_{q,c}(i,k))} r \tag{6.10}$$

At each iteration, an 'error value' is calculated. The algorithm terminates when it reaches a stationary state, i.e. once the error value is sufficiently small. The error value is calculated as follows.

$$\varepsilon(t) = \sum_{i \in S} abs(\mu_i(t) - \mu_i(t-1)) \tag{6.11}$$

where $\varepsilon(t)$ is the error value at iteration t, and $t > 0$. A proper error threshold is calculated dynamically.

The result is a salience value μ between 0 and 1 associated with each passage. The content units with the highest salience values are selected for inclusion in the summary.

The quality of normalized centrality summaries is slightly lower than that of query-distance summaries according to Rouge-2 and Rouge-SU4. Compared to query-relevance summaries, the Rouge-2 score is higher and the Rouge-SU4 score was lower. At a Rouge-2 score of 0.0820 and a Rouge-SU4 score of 0.136, the average quality of normalized centrality summaries does not significantly differ (at $p < 0.05$) from query-relevance or query-distance summarization. An example of a normalized centrality summary is printed in appendix B.4. The normalization cancels the effect of graph weighting – the summaries are the same regardless the value of the (non-zero) graph weights.

6.2.5.2 Probabilistic centrality

The algorithm of Erkan and Radev (2004) was inspired by social networks. Analogously to people, content units have a certain status (salience) and use this status to

increase their friends' status (i.e. to increase similar sentences' salience). In the work of Erkan and Radev, the weights are normalized by degree, so that total amount of salience (status) in the network remains approximately the same over iterations. If a person has many friends, those friends individually have less benefits from the friendship because they have to share. In other words, if a relation is added, existing relations to the same content unit are devaluated.

In the probabilistic approach, an edge in the semantic network from sentence a to b is viewed as the probability that b is relevant, given a is relevant. In this case, contrary to the normalized approach, the relevance of b given a is unaffected by any other sentence whose relevance may depend on a. Viewing edges as relevance probabilities also has implications on how evidence of relevance is combined. Rather than accumulating weighted salience of neighbors, salience of a sentence is calculated as the product of inverse conditional probabilities. This is based on the idea that, if we have several pieces of evidence that a sentence is salient, it suffices if one of them is true. I used the following probabilistic centrality algorithm for calculating salience.

$$\begin{aligned} \nu_j(t) &= 1 && \text{, if } j \in Q \\ \nu_j(0) &= 0 && \text{, if } j \in S \\ \nu_j(t+1) &= 1 - \prod_{(i \in Q \cup S)} \prod_{(r \in \Delta_{q,c}(i,j))} (1 - r \cdot \nu_i(t) \cdot y) && \text{, if } j \in S \end{aligned} \tag{6.12}$$

where $\nu_j(t)$ is the probabilistic centrality value of sentence j at iteration t. The value of y is the *decay* value, a global constant in the range $\langle 0..1 \rangle$. The constant y has a function similar to the constant d in normalized centrality: it is necessary to ensure that the salience value keeps increasing at each iteration.

Figure 6.3 shows the performance of the probabilistic centrality system is higher if the relative contribution of the query-relevance graph δ_q is greater. The best performance measured is at $w_q = 1$, $w_c = 0.1$ (Rouge-2 average 0.0888, DUC rank 3; Rouge-SU4 average 0.143, DUC rank 7). With this configuration, the system significantly outperforms the query-relevance system ($p < 0.01$ for Rouge-2 and Rouge-SU4), query-distance with $w_q = 0.7$ ($p < 0.01$ for Rouge-2 and Rouge-SU4), and normalized centrality ($p < 0.05$ for Rouge-2 and Rouge-SU4). An example of a probabilistic centrality summary is shown in appendix B.5.

6.2.5.3 An example

To clarify the iterative centrality-based summarization process, an instance of probabilistic centrality summarization is exemplified. The text in Figure 6.4 consists of the

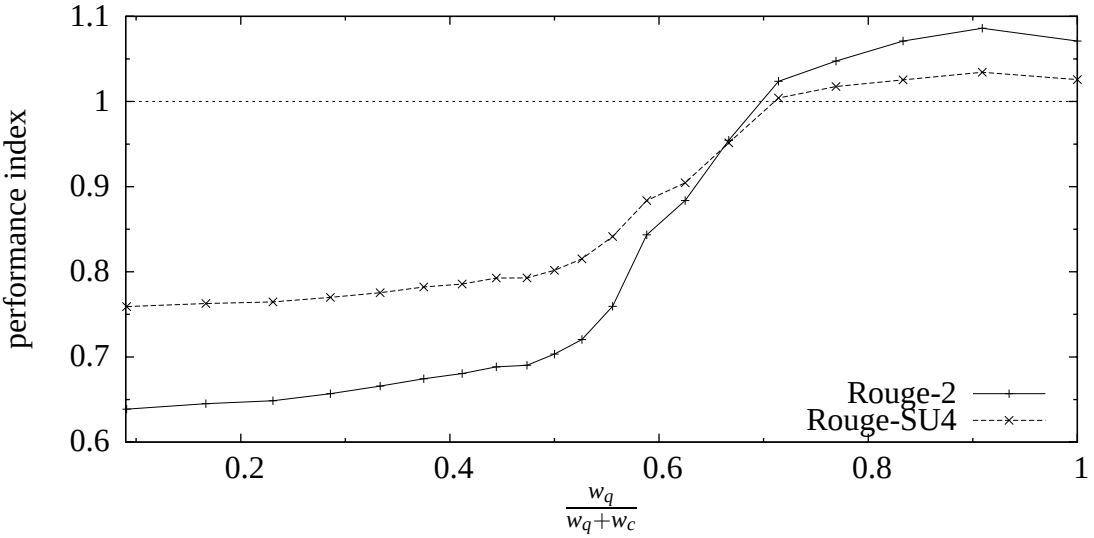

Figure 6.3: Indexed performance of probabilistic centrality summarization for different values of w_q and w_c. An indexed performance of 1 indicates the performance of the query-relevance system. Performance is measured for $w_c = 1$ and w_q incrementing in steps of 0.1 from 0 to 1. Then, performance was measured for $w_q = 1$ and w_c incrementing in steps of 0.1 from 0 to 1. The configuration of highest performance was $w_q = 1$; $w_c = 0.1$.

Q	Hugo Chavez
20A	CARACAS, Venezuela (AP) – Former Lt. Col. Hugo Chavez, who staged a bloody coup attempt six years ago, was elected president of Venezuela – a stunning blow to the political and economic establishment that has ruled for 40 years.
20B	People poured into the streets late Sunday, dancing, setting off fireworks and honking horns in celebration of what many viewed a victory of the poor over a political elite that has failed to ease poverty and control rampant corruption.
20C	"Venezuela is being born again," Chavez declared as results were revealed.
20D	"Once again, the people of Simon Bolivar have shown themselves to be a grand people," he told the Venevision television network.
20E	Chavez often invokes South American liberation hero Bolivar in his speeches.
20F	With 78 percent of the vote counted, Chavez had 56 percent compared to 40 percent for Yale-educated businessman Henrique Salas Romer, according to official results from the National Electoral Council.
20G	"I want to say to all Venezuelans that not only do I accept my adversary's victory, I also wish him much luck because his luck is Venezuela's luck," Salas said at his campaign headquarters in Caracas.

Figure 6.4: Excerpt from APW19981206.1106.

Table 6.1: Stemmed cosine similarity network.

δ	Q	20A	20B	20C	20D	20E	20F	20G
Q	1.00	0.34	–	0.36	–	0.26	0.17	–
20A	0.34	1.00	0.01	0.36	–	0.09	0.07	0.14
20B	–	0.01	1.00	–	–	–	–	0.03
20C	0.36	0.36	–	1.00	0.05	0.18	0.14	0.10
20D	–	–	–	0.05	1.00	0.25	–	–
20E	0.26	0.09	–	0.18	0.25	1.00	0.09	–
20F	0.17	0.07	–	0.14	–	0.09	1.00	0.11
20G	–	0.14	0.03	0.10	–	–	0.11	1.00

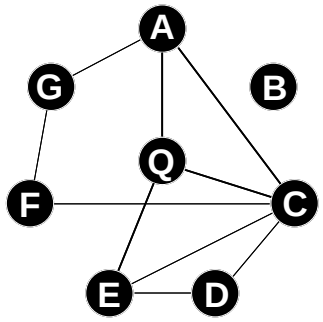

Figure 6.5: Graphical representation of Table 6.1. A thick line indicates strong similarity (> 0.25); a thin line indicates moderate similarity (> 0.1); no line indicates marginal or no similarity.

Table 6.2: Activation after t iterations. The algorithm continues until $\varepsilon < 0.001$. The final ranking in order of activation after 15 iterations is Q, 20C, 20A, 20E, 20F, 20G, 20D, 20B.

t	(ε)	Q	20A	20B	20C	20D	20E	20F	20G
0		1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	(0.563)	1.000	0.169	0.000	0.181	0.000	0.129	0.084	0.000
2	(0.395)	1.000	0.270	0.001	0.291	0.020	0.207	0.143	0.025
4	(0.168)	1.000	0.364	0.004	0.391	0.059	0.285	0.209	0.075
8	(0.024)	1.000	0.406	0.007	0.436	0.090	0.324	0.245	0.112
15	(0.001)	1.000	0.412	0.008	0.442	0.096	0.330	0.250	0.118

first seven sentences of document APW19981206.1106 from the DUC 2006 corpus. The query used for summarization is *Hugo Chavez*. The text is sentence-segmented, and thus the sentences are the content units. After stemming (Porter, 1980) and creating IDF-weighted term frequency vectors for the sentences, the semantic network in Table 6.1 (graphically represented in Figure 6.5) is derived by measuring the cosine similarity between each pair of sentences. At this point, we have a feature graph in the form of an adjacency matrix that can be used for summarization:

$$\Delta_{q,c}(i,j) = \{cosim(i,j)\} \tag{6.13}$$

The cosine similarity between 20A and 20E can be read from the cell of row 20A and column 20E (i.e. 0.09). Because cosine similarity is commutative, $cosim(i,j) = cosim(j,i)$ for all i, j.

The relevance level of query content units $\nu_{q\in Q}$ is fixed at 1; the relevance of the document content units $\nu_{q\in S}$ is initialized as 0. Relevance levels are recalculated in subsequent iterations and can be read from Table 6.2. At iteration 0, the salience of the query is initialized at 1; other salience values are initialized at 0. At iteration 1, the salience of query-related sentences is increased with the strength of their relation with the query. Then, their salience is multiplied by 0.5. For instance, the salience of 20A is $1 \cdot 0.35 \cdot 0.5 = 0.169$. At iteration 2, the salience of sentences which are related with the query or a query-related sentence is increased. For instance, the salience of sentence 20A is:

$$\begin{aligned}
\nu_{20A}(2) &= 1 - \prod_{(i\in Q\cup S)} \prod_{(r\in\Delta_{q,c}(i,20A))} (1 - r\cdot\nu_i(1)\cdot y) \\
&= 1 - \prod_{(i\in Q\cup S)} (1 - cosim(i,20A)\cdot\nu_i(1)\cdot y) \\
&= 1 - (1 - cosim(Q,20A)\cdot\nu_Q(1)\cdot y)\cdot(\ldots) \\
&= 1 - 0.831\cdot 0.916\cdot 1.000\cdot 0.967\cdot 1.000\cdot 0.994\cdot 0.997\cdot 1.000 \\
&= 0.270
\end{aligned} \tag{6.14}$$

The salience of all sentences are updated until the stopping condition (here chosen as $\varepsilon < 0.001$) is met. In this case, this is after 15 iterations. The final salience values of individual sentences are the relevance levels after the last iteration before terminating.

After the sentences are ranked by salience, we can select the most salient sentences and present them as a summary to the user. Supposing the presentation algorithm simply picks the three highest-ranking non-query sentences and puts them in the same

21A	People poured into the streets late Sunday, dancing, setting off fireworks and honking horns in celebration of what many viewed a victory of the poor over a political elite that has failed to ease poverty and control rampant corruption.
21B	"Once again, the people of Simon Bolivar have shown themselves to be a grand people," he told the Venevision television network.
21C	With 78 percent of the vote counted, Chavez had 56 percent compared to 40 percent for Yale-educated businessman Henrique Salas Romer, according to official results from the National Electoral Council.

Figure 6.6: Example probabilistic centrality summary for the query *Hugo Chavez*.

order as they appeared in the original document, the summary consists of the sequence of sentences 20A, 20C and 20E. This summary is shown in Figure 6.6.

6.2.6 Redundancy-aware summarization

One of the assumptions usually made implicitly in the design of single-document summarization systems, is that the source document does not contain redundancy. Consequently, there is no risk of including a sentence in the summary which does not contain any information not already present. This changes when a summary is generated from multiple source documents, where non-redundancy of sentences from different documents cannot taken for granted. The content selection procedures outlined previously concentrate entirely on relevancy, not redundancy. However, in multi-document summarization, presented content should be relevant to the query and novel with respect to what is already mentioned in the summary (c.f. Carbonell and Goldstein, 1998).

6.2.6.1 Adding redundancy

To accommodate representing novelty, the model is extended with a redundancy feature graph P which is used in addition to the previously mentioned relevancy feature graph Δ. Similarly to relevance, redundancy relations have a strength in the range $[0..1]$. The strength of a redundancy relation between two sentences expresses the likelihood that a sentence is redundant, given the fact that another sentence is redundant. The redundancy of sentence j, given sentence i, is defined by $\delta_r(i,j)$. The form of the redundancy graph is identical to that of the relevancy graph. The strengths of relations in the redundancy feature graph δ_r are defined as follows:

$$\begin{aligned} &\delta_r(i,j) = cosim(i,j) && \text{, if } i,j \in S; doc(i) \neq doc(j) \\ &\delta_r(i,j) = 0 && \text{, otherwise} \end{aligned} \tag{6.15}$$

The redundancy-aware summarization system uses a set of redundancy feature graphs P for determining salience of sentences, in addition to the relevancy feature graphs Δ:

$$\Delta_{q,c,r}(i,j) = \left\{ w_q \cdot \delta_q(i,j), w_c \cdot \delta_c(i,j), w_{r\Delta} \cdot \delta_r(i,j) \right\} \tag{6.16}$$

$$P_r(i,j) = \{ w_{rP} \cdot \delta_r(i,j) \} \tag{6.17}$$

where $\delta_q(i,j)$, $\delta_c(i,j)$ and $\delta_r(i,j)$ refer to eq. 6.2, 6.4, and 6.15 respectively. The set of relations between sentences i and j are represented by $\Delta_{q,c,r}(i,j)$ (relevancy) and $P_r(i,j)$ (redundancy). Since redundancy implies 'relatedness', I regard a redundancy graph a special case of a relevancy graph. Therefore, δ_r is not only included in P_r but also in $\Delta_{q,c,r}$.

The calculation of redundancy-adjusted salience was inspired by Carbonell and Goldstein (1998). First, the relevance of each sentence is calculated using $\Delta_{q,c,r}$ using whatever method appropriate. Then, the novelty is calculated – novelty is the reciprocal of redundancy. If two sentences are redundant, this affects only the novelty of the less-relevant of the two. The stronger the redundancy relation, the greater the reduction of novelty. Novelty is calculated as follows:

$$N(j) = \prod_{i \in F_j} \prod_{r \in P_r(i,j)} (1 - r \cdot R(i)) \tag{6.18}$$

$$F_j = \{ k : S \mid R(k) > R(j) \} \tag{6.19}$$

where $N(j)$ is a value in the range $[0..1]$, representing the novelty of sentence j; $P_r(i,j)$ is a set of redundancy relations, expressing the redundancy of j given i; F_j is the set of content units more relevant than j. The function $R(i)$ denotes the relevance of sentence i, as previously calculated.

Now, the redundancy-adjusted salience can be calculated as the product of relevancy and novelty:

$$\sigma_j = R(j) \cdot N(j) \tag{6.20}$$

where σ_j is the redundancy-adjusted salience of sentence j.

The calculation of σ_j ensures that:

- if one content unit is selected, all content units redundant to that unit are less likely to be selected: if two content units are redundant with respect to each other, the salience of the less-relevant content unit is reduced;
- redundancy of a content unit does not prevent relevancy to propagate: a redundant content unit may still be relevant.

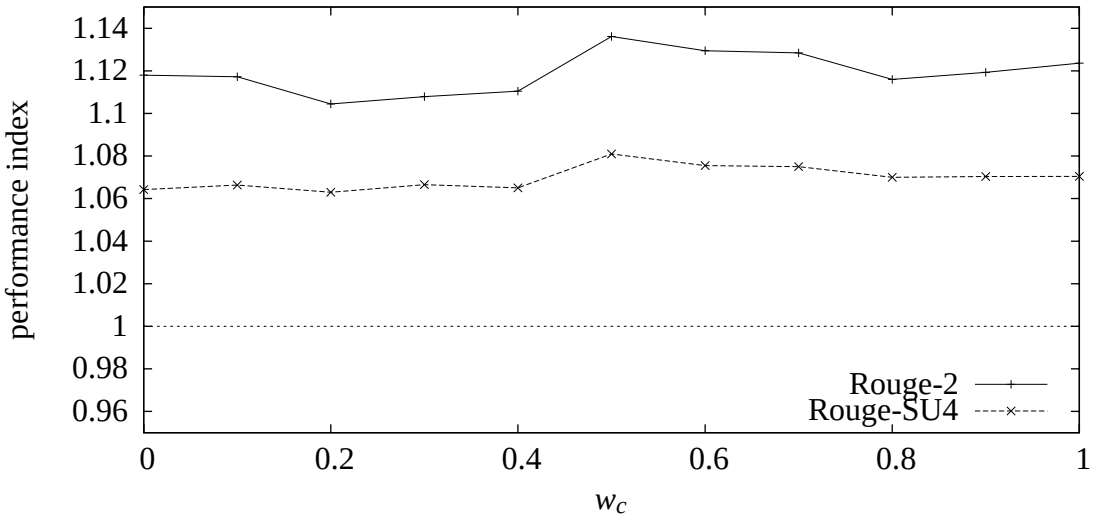

Figure 6.7: Indexed performance of normalized centrality summarization with $w_q = 1$; $w_{r\Delta} = 1$; $w_{rP} = 0$; for different values of w_c. An indexed performance of 1 indicates the performance of the query-relevance system.

6.2.6.2 Query-distance

For query-distance summarization, I started with the configuration which performed best in section 6.2.4. That is, $w_q = 0.7$ and $w_c = 1$. After adding the redundancy graph, I started with $w_{r\Delta} = 0$ and $w_{rP} = 0$. Then, I incremented $w_{r\Delta}$ to a value of 1 in steps of 0.1 while leaving the w_{rP} at 0, and vice versa. The best performance was achieved with $w_{r\Delta} = 0.5$; $w_{rP} = 0$ (Rouge-2 0.0824, Rouge-SU4 0.138, DUC rank 11). However, in this configuration, the content of summaries constructed by the query-distance algorithm was little affected by adding the redundancy graph. This may be caused by the fact that the query-distance algorithm is relatively conservative with respect to selecting sentences for which there is indirect evidence of relevance. I did not measure a significant improvement of the quality of summaries for any value of $w_{r\Delta}$ and w_{rP}.

6.2.6.3 Normalized centrality

For the normalized centrality algorithm, the best performing configuration with $w_q = 1$, $w_c = 1$ and $w_{rP} = 0$ is at $w_{r\Delta} = 1$. Starting with this configuration, I decreased the value of w_c to 0 in steps of 0.1. The performance is plotted in Figure 6.7. The best measured performance was at $w_c = 0.5$ with a Rouge-2 score of 0.0929 (DUC rank 2)

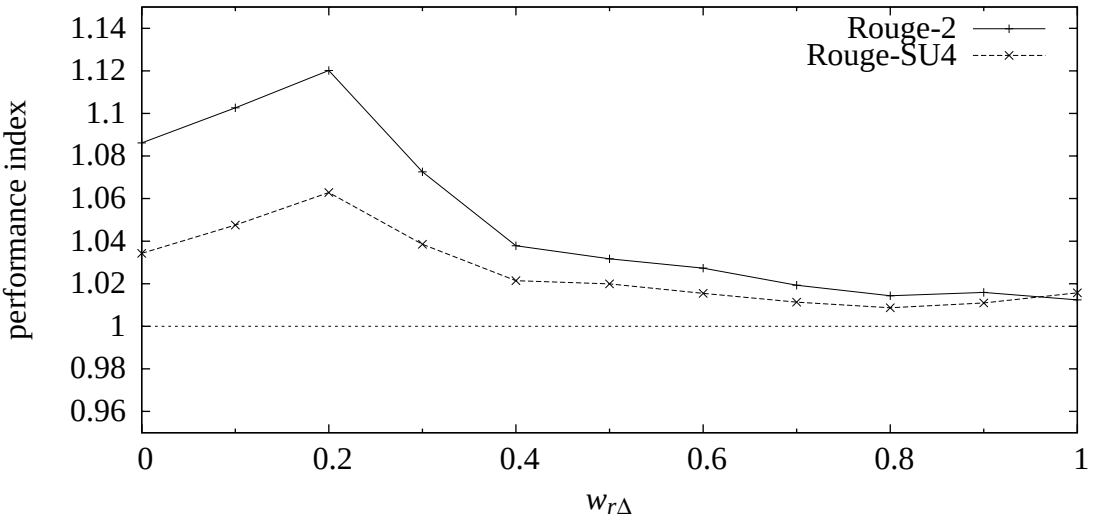

Figure 6.8: Indexed performance of probabilistic centrality summarization with $w_q = 1$; $w_c = 0.1$; $w_{rP} = 0$; for different values of $w_{r\Delta}$. An indexed performance of 1 indicates the performance of the query-relevance system.

and a Rouge-SU4 score of 0.150 (DUC rank 2). An example of a summary produced using this configuration is shown in appendix B.6. This configuration significantly outperforms previously measured configurations of the query-distance and normalized centrality systems (Rouge-2 and Rouge-SU4). It also performs slightly better than the probabilistic centrality system, but this difference is not significant. Increasing the value of w_{rP} had no effect on the quality of the summaries.

6.2.6.4 Probabilistic centrality

For the probabilistic centrality algorithm, the best performing configuration in section 6.2.5.2 was at $w_q = 1$, $w_c = 0.1$. Starting with this configuration and the additional redundancy graph, I increased the value of $w_{r\Delta}$ from 0 to 1 in steps of 0.1, while $w_{rP} = 0$. The performance is plotted in Figure 6.8. The best measured performance is at $w_{r\Delta} = 0.2$ with a Rouge-2 score of 0.0916 (DUC rank 2) and a Rouge-SU4 score of 0.147 (DUC rank 3).

Starting with the best probabilistic centrality configuration so far, I increased the value of w_{rP} from 0 to 1 in steps of 0.1. The performance is plotted in Figure 6.9. The best measured performance is at $w_{rP} = 1.0$ with a Rouge-2 score of 0.0930 (DUC rank 2) and a Rouge-SU4 score of 0.150 (DUC rank 2). An example of a summary produced using this configuration is shown in appendix B.7. This configuration sig-

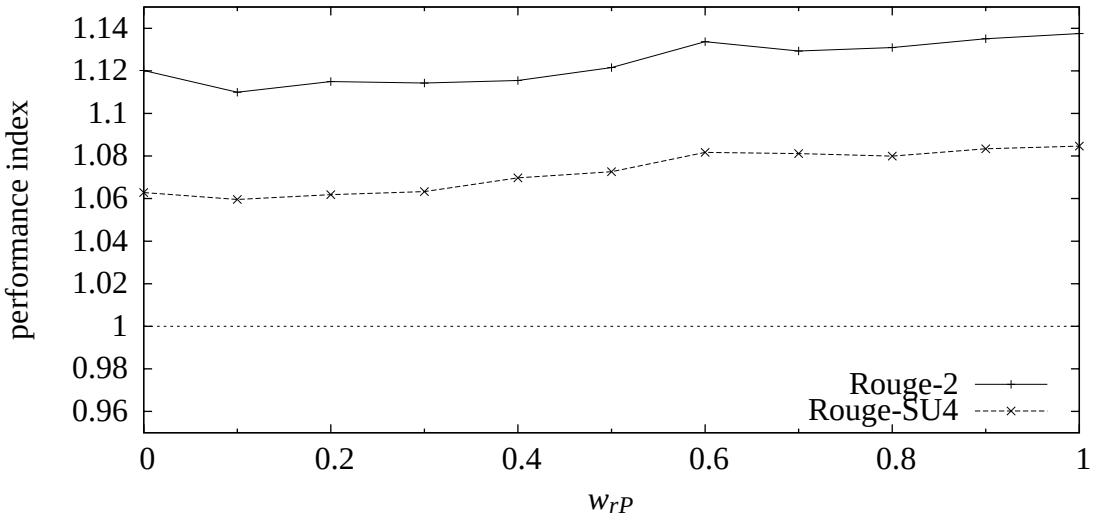

Figure 6.9: Indexed performance of probabilistic centrality summarization with $w_q = 1$; $w_c = 0.1$; $w_{r\Delta} = 0.2$; for different values of w_{rP}. An indexed performance of 1 indicates the performance of the query-relevance system.

nificantly outperforms the query-distance system and the normalized centrality system without the use of a redundancy graph (Rouge-2 and Rouge-SU4). It also performs slightly better than the probabilistic centrality system without the redundancy graph. This difference is significant for Rouge-SU4 but not for Rouge-2.

6.2.7 Validating the results

In this section, I compare the performance of several configurations of the summarization framework proposed in section 6.1. An overview of the results is shown in Figure 6.10. The best performance was achieved with the probabilistic centrality content selection algorithm, although it did not perform significantly better than the normalized centrality algorithm. For pair-wise comparison of summarization systems, I used as a measure of quality the percentage of DUC queries for which one system received a higher Rouge-2 or Rouge-SU4 score than the other. An overview of the results is given in Table 6.3.

The way the graph weight configurations are determined implies that the weights are tailored to the DUC 2006 data set. As a result, there is a risk that the weights are overfitted to this particular set. In order to validate the results, I ran the experiments on the DUC 2005 data set with the graph weight configurations of the systems in Figure

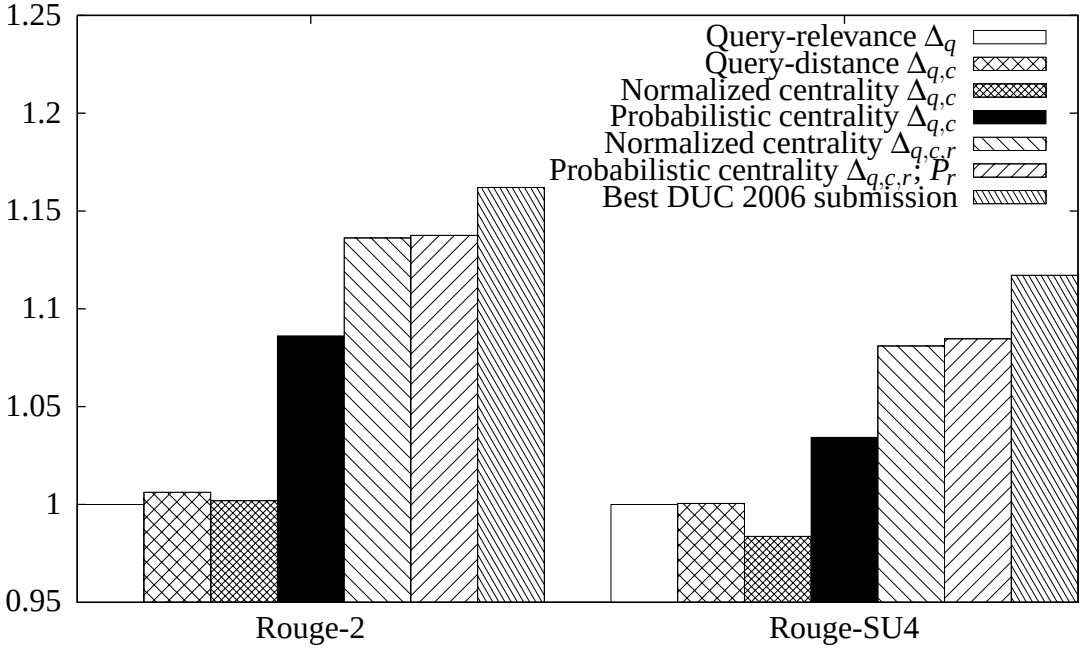

Figure 6.10: Indexed performance on DUC 2006 data: 1 indicates the performance of the query-relevance system.

Table 6.3: Percentage of DUC 2006 topics (Rouge-2/Rouge-SU4) for which one system (rows) beat another (columns). Note that percentages do not add up to 100 if both systems receive the same score for at least one topic. The compared systems are (a) query-relevance (Δ_q); (b) query-distance ($\Delta_{q,c}$); (c) normalized centrality ($\Delta_{q,c}$); (d) probabilistic centrality ($\Delta_{q,c}$); (e) normalized centrality ($\Delta_{q,c,r}$); (f) probabilistic centrality ($\Delta_{q,c,r}$; P_r).

%	(a)	(b)	(c)	(d)	(e)	(f)
(a)	–	12/18	50/52	34[a]/28[a]	30[a]/28[a]	26[a]/26[a]
(b)	22/18	–	52/54	38[b]/30[a]	32[a]/30[a]	30[a]/26[a]
(c)	46/48	44/46	–	34[a]/36[b]	38[b]/34[a]	30[a]/24[a]
(d)	64[a]/70[a]	60[b]/68[a]	66[a]/62[b]	–	56/58	44/50
(e)	66[a]/66[a]	66[a]/64[a]	60[b]/62[a]	42/42	–	30[a]/30[a]
(f)	70[a]/72[a]	72[a]/74[a]	68[a]/72[a]	48/46	64[a]/68[a]	–

[a] Significant at $p < 0.01$.

[b] Significant at $p < 0.05$.

[c] Significant at $p < 0.1$.

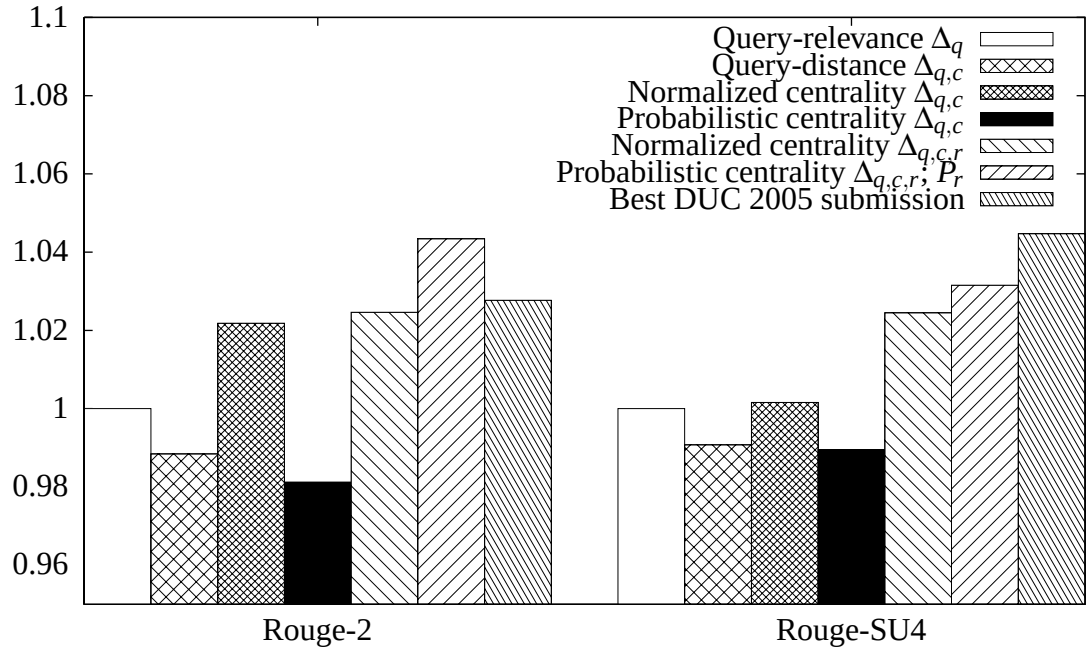

Figure 6.11: Indexed performance on DUC 2005 data: 1 indicates the performance of the query-relevance system.

Table 6.4: Percentage of DUC 2005 topics (Rouge-2/Rouge-SU4) for which one system (rows) beat another (columns). Note that percentages do not add up to 100 if both systems receive the same score for at least one topic. The compared systems are (a) query-relevance (Δ_q); (b) query-distance ($\Delta_{q,c}$); (c) normalized centrality ($\Delta_{q,c}$); (d) probabilistic centrality ($\Delta_{q,c}$); (e) normalized centrality ($\Delta_{q,c,r}$); (f) probabilistic centrality ($\Delta_{q,c,r}$; P_r).

%	(a)	(b)	(c)	(d)	(e)	(f)
(a)	–	20/26[b]	46/44	42/42	50/50	40[c]/40[c]
(b)	12/10[b]	–	44/44	40/38[b]	48/50	38[b]/36[a]
(c)	52/54	54/54	–	50/34[a]	50/54	38[b]/34[a]
(d)	54/58	56/62[b]	50/66[a]	–	58[b]/64[a]	36[c]/42
(e)	44/44	48/48	46/44	38[b]/36[a]	–	30[a]/30[a]
(f)	58[c]/60[c]	60[b]/64[a]	60[b]/66[a]	54[c]/54	60[a]/70[a]	–

[a] Significant at $p < 0.01$.

[b] Significant at $p < 0.05$.

[c] Significant at $p < 0.1$.

6.10. The design of the DUC 2005 data set is almost identical to the design of DUC 2006. The only difference in the summarization task is that a DUC 2005 topic includes a desired 'granularity' (specific or general) of the summary. I ignored this directive.

Figure 6.11 shows the average Rouge-2 and Rouge-SU4 scores achieved with the DUC 2005 corpus. Table 6.4 shows an overview of the significance tests. The redundancy-aware probabilistic centrality system significantly outperformed all other systems when Rouge-2 is used ($p < 0.1$), and all except the redundancy-aware normalized centrality system according to Rouge-SU4. This system would have ranked first (Rouge-2) or second (Rouge-SU4) if it had participated in DUC 2005. Remarkably, no system outperformed the query-relevance baseline system with a certainty of $p < 0.01$ to falsely reject the null hypothesis, despite the distinct differences in average Rouge scores.

Note that it is not guaranteed that the combination of graph weights that leads to the best performance has been found, not even on DUC 2006 data. Apart from the risk of overfitting, the number of possible graph weight combinations is infinite and a greater number of graphs makes it more difficult to find the best combination of weights. A possible future extension to the summarization system may use machine learning methods such as genetic algorithms to find the optimal solution.

6.3 Evaluation: DUC

An implementation of the summarization framework described in the previous section was evaluated in the context of DUC 2006. Because this evaluation was performed before the experiments in section 6.2, the system described here deviates from the system described previously. The data and the evaluated task are the same as used in section 6.2. In addition to Route-2 and Rouge-SU4, the summarization system was externally evaluated by means of:

- Basic Elements (see section 4.2.1.5);
- responsiveness – all summaries are rated by human assessors on a five-point scale for responsiveness with respect to the query;
- linguistic quality – all summaries are rated by human assessors on a five-point scale for five aspects of linguistic quality, i.e. grammaticality, non-redundancy, referential clarity, focus, and coherence;

- Pyramids – 22 of the 35 participant systems of DUC (including the system described here) participated in the optional Pyramids evaluation (see section 4.2.1.4).

The additional evaluations reveal aspects of the quality of discourse oriented summaries which cannot be evaluated using automatic metrics. The baseline used in DUC is a summary composed of the leading sentences of the most recent documents, up to 250 words.

6.3.1 Feature graphs

A combination of the following four feature graphs was used to find intersentential relations:

- an entailment graph, to express query-relevance;
- an entailment graph, to express redundancy between sentences of different documents.
- the cosine similarity graph δ_c as defined in eq. 6.4, to express relatedness between sentences;
- a layout graph, to express relatedness between sentences;

An entailment system recognizes whether one piece of text is entailed by another. For recognizing entailment, the dependency tree alignment algorithm described in section 3.3.1.2 is used. This algorithm detects entailment by identifying overlap in their dependency trees. If a sentence entails what is said in the query, the sentence may provide an answer to the query. The use of an alignment algorithm for finding answers is based on the observation that recognizing a question/answer relation is similar to recognizing an entailment relation, and both can be found using syntactic structure. Bouma et al. (2006) show that it is likely that a sentence answers a question if the syntactic structure of question and candidate answer sentence sentence is similar. The query-relevance entailment graph is defined as follows:

$$\begin{aligned} \delta_{qa}(q,j) &= align(q,j) && \text{, if } q \in Q;\ j \in S \\ \delta_{qa}(q,j) &= 0 && \text{, otherwise} \end{aligned} \tag{6.21}$$

where $\delta_{qa}(q,j)$ is the strength of the relation between q and j; and $align(q,j)$ is the alignment value as calculated by the dependency tree alignment algorithm.

Recognizing textual entailment is also useful for detecting redundancy across documents. If a summary sentence entails another sentence, the latter sentence is redundant and including it in the summary should be avoided. The redundancy entailment graph is defined as follows:

$$\begin{aligned} &\delta_{ra}(i,j) = align(i,j) && \text{, if } j \in S;\ i \text{ is among the 10 most query-relevant sentences} \\ &\delta_{ra}(i,j) = 0 && \text{, otherwise} \end{aligned} \tag{6.22}$$

The alignment algorithm uses lemma equivalence (van den Bosch and Daelemans, 1999) and WordNet synonymy and hyponymy (Miller, 1995) for alignment on the lexical level. Synonyms and words with the same lemma are considered equivalent. The MaltParser system (Nivre and Scholz, 2004) is used for syntactic analysis.

In addition to direct answers to questions, sentences which elaborate on answers are also included. A combination of layout and cosine similarity is used to relate sentences within a document. Cosine similarity relies on the graph δ_c (eq. 6.4). The lowest level coherence relations typically do not cross paragraph boundaries – a paragraph participates as a whole in a coherence relation with the text in which it is embedded. This knowledge may be used to derive the structure of a text, but it can also be exploited directly by a summarization system. The way layout is used here is based on the idea that the first sentence in a paragraph often contains the most important information. The layout feature graph δ_p bidirectionally connects each sentence with the first sentence of its paragraph:

$$\begin{aligned} &\delta_p(i,j) = 1 && \text{, if } i \text{ is the first sentence of the paragraph of } j\text{, or vice versa} \\ &\delta_p(i,j) = 0 && \text{, otherwise} \end{aligned} \tag{6.23}$$

To summarize, alignment of dependency trees is measured to find answers to questions, and to detect redundancy across documents. Paragraph boundaries are used as an indication of structural relations between sentences, and cosine similarity is used to find semantic relations between sentences within a document. The relevancy and redundancy graphs used for determining the salience of sentences are the following:

$$\Delta_{DUC} = \{0.15\delta_{qa}, 5\delta_{ra}, 0.1\delta_p, \delta_c\} \tag{6.24}$$

$$P_{DUC} = \{5\delta_{ra}\} \tag{6.25}$$

6.3.2 Content selection

The content selection algorithm is similar to the query distance algorithm described in section 6.2.4, but it differs in several ways. First, in the previous section, I assumed that all relations strengths are in the range $[0..1]$. The centrality based algorithms demand this, but not the query-distance algorithm. The restriction is relieved in this system.

Second, the system favors long sentences over short sentences. The reason for this is that short sentences are expected to contain more anaphoric references to other sentences. Including a sentence but not the sentences containing the antecedents of its references results in a summary with unresolvable references. In the DUC system, query-distance is calculated as follows (c.f. eq. 6.7):

$$D_{duc}(j) = 0 \qquad \text{, if } j \in Q \tag{6.26}$$

$$D_{duc}(j) = min\left\{ D_{duc}(i) + \left(\sum_{r \in \Delta_{DUC}(i,j)} r \right)^{-1} \mid i \in Q \cup S \right\} \qquad \text{, otherwise}$$

where $D_{duc}(j)$ is the distance of j to the query, using Δ_{DUC}. Distance is measured by D_{duc} as the shortest path from the query to the sentence, not taking redundancy into account.

If the last step of the path to a sentence follows a redundancy relation, the corresponding sentence is redundant with respect to another sentence which is closer to the query. This is expressed by the value $D_{ducR}(j)$, which is the distance of j to the query, measured as the shortest path of which the last step is *not* a redundancy relation:

$$D_{ducR}(j) = min\left\{ D_{duc}(i) + \left(\sum_{r \in \Delta_{DUC}(i,j) \setminus P_{DUC}(i,j)} r \right)^{-1} \mid i \in Q \cup S \right\} \tag{6.27}$$

The redundancy-aware sentence salience is the reciprocal of the distance to the query:

$$R_{duc}(j) = \frac{\sqrt{\|j\|}}{D_{ducR}(j)} \tag{6.28}$$

where $R_{duc}(j)$ is the salience of sentence j; $D_{ducR}(j)$ is the distance to the query, and $\|j\|$ is the number of characters in sentence j. The latter is to favor long sentences over shorter ones.

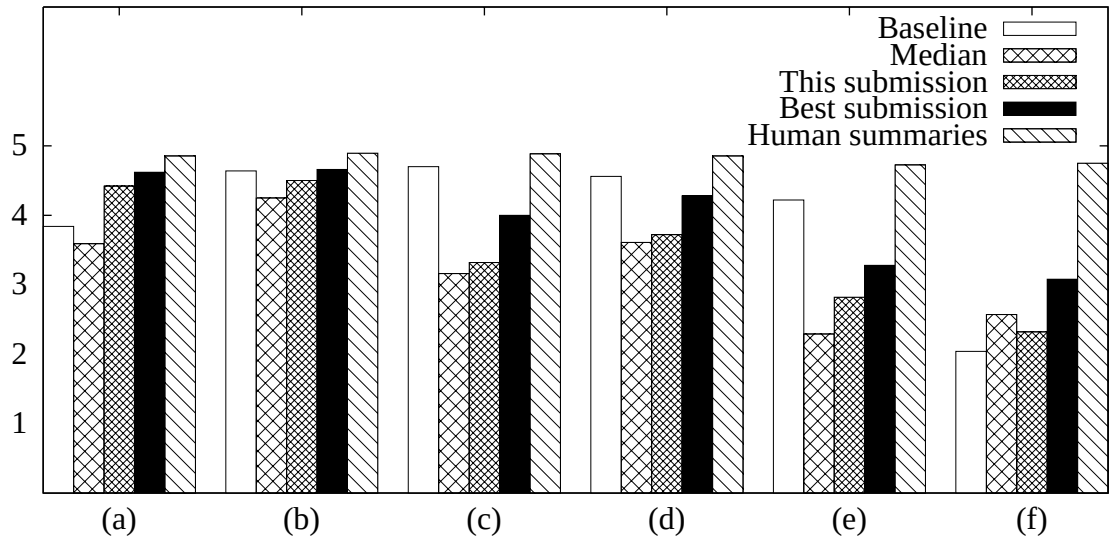

Figure 6.12: Average human assessment of various aspects of the quality of 50 summaries: (a) grammaticality, (b) non-redundancy, (c) referential clarity, (d) focus, (e) structure and coherence, (f) responsiveness as evaluated by NIST assessors on a five point Likert scale.

6.3.3 The results

Organizers of DUC 2006 at NIST hired assessors to evaluate various aspects of the quality of summaries. Figure 6.12 shows the average results of the readability assessments. On all aspects of readability, the average results are above the median of DUC participants. On average over all summaries and all evaluated aspects of linguistic quality, the system performed second-best of 35 participants.

Figure 6.12 also shows results of a human assessment of 'responsiveness', i.e. how well the summaries respond to the information need expressed in the query. In this evaluation, the system outperformed the baseline (composed of leading sentences of the most recent source documents) but not the median submission. The same is true for other content based evaluation methods. Figure 6.13 shows the results of content based evaluation methods which measure the content overlap between a candidate summary and hand crafted reference summaries.

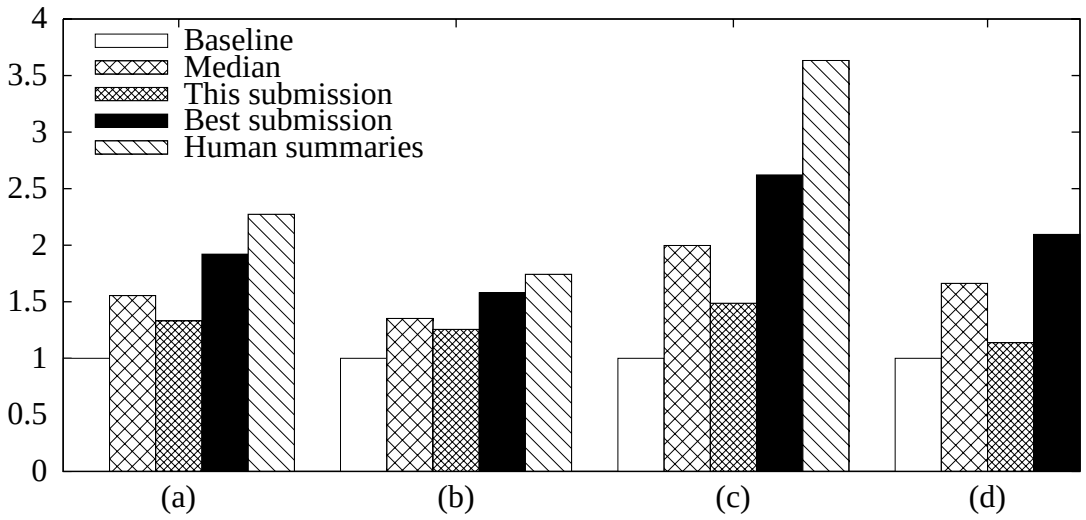

Figure 6.13: Indexed performance (baseline performance is 1) of DUC submissions as produced by NIST and Columbia University using four evaluation metrics: (a) Rouge-2, (b) Rouge-SU4, (c) Basic Elements, (d) Pyramid.

6.4 Conclusion

This chapter presented a graph-based system for query-based summarization. The system's objective is to select cohering sentences whose content is relevant to the user. Content is likely to be relevant if (1) it answers the query, or if (2) it supports answers to the query by providing background information, elaborating on specific details, or otherwise providing information which is related to the answer. The summarization system presented in this chapter models the source text by means of graphs, aiming at capturing the discourse structure and using that to produce coherent summaries containing an answer to the query, but also information supporting this answer. Previous work on query-based summarization has mainly focused on extracting the set of sentences which best match the query, not on producing coherent summaries.

The effect of two system variables on the system performance is measured: the use of different graphs, and the use of different algorithms to search those graphs. My main interest is to evaluate search algorithms, but their performance rely on the particular choice of graphs. The algorithms used are query-relevance, query-distance, normalized centrality and probabilistic centrality. The graphs express the strength of relations between sentences, measured as their cosine similarity. All evaluations are performed on DUC (2005 and 2006) corpora, using automatic performance metrics

also used in DUC, i.e. Rouge-2 and Rouge-SU4. One implementation of the system is evaluated by NIST in the context of DUC 2006.

The Rouge measurements on the DUC 2006 corpus (section 6.2) suggest that the probabilistic centrality algorithm leads to the best performance. Because the configuration of the system parameters are also based on this corpus, the results are validated on the DUC 2005 corpus. Rouge measurements on the DUC 2005 corpus show that the probabilistic centrality algorithm significantly outperforms all other algorithms ($p < 0.1$). When using the probabilistic centrality algorithm, the redundancy graph contributed significantly to the Rouge-2 system performance ($p < 0.1$) but not to the Rouge-SU4 performance, despite the difference in average performance.

Apart from Rouge-2 and Rouge-SU4, the evaluation by NIST (section 6.3) included also performance measurements using Basic Elements, Pyramids, as well as ratings of responsiveness and various aspects of readability. Because this evaluation was performed before the previously mentioned experiments and the parameters of the system were not optimized for content at this stage, it is not surprising that the content-based metrics indicate a score well below the results presented in section 6.2. Nevertheless, the average of readability scores was second best of all DUC 2006 submissions. The high level of readability is in line with my expectation that including content supporting the answer to the query (sentences with indirect evidence of relevance) increases coherence of the summary.

In sum, the graph-based approach to query-based summarization results in high quality summaries with respect to content (section 6.2) and readability (section 6.3). Unfortunately, the readability measures of DUC 2006 could not easily be repeated in experiments with the content-optimized system. As a result, further research will have to point out whether graphs can be used to combine the virtues of readability and high quality content in a single summarization system. Another possibility is that there is a trade-off between content and readability.

In section 6.2, the cosine similarity of sentences was used to measure their relatedness. Other knowledge-poor evidence of relatedness is used in addition to cosine similarity in section 6.3. These methods are computationally cheap and easy to port to other languages. On the other hand, knowledge-intensive methods may detect relations between sentences more accurately. A promising direction of further research is to use more different sources of information for detecting relations, including knowledge-intensive methods such as rhetorical relation detection or anaphora resolution.

7

Illustrating answers

The previous chapters made a case for discourse oriented text summarization. This chapter applies ideas from discourse oriented summarization on automatic illustration of answers to medical questions. Similarities between the answers and picture-related text (its caption or its section/paragraph) is used as evidence that the picture would be appropriate to illustrate the answer. In a user study, participants rated presentations consisting of a textual component and a picture. The text was manually written and the picture was automatically retrieved using either its caption or its section. The caption-based selection method gave better results than the section-based method, and the caption-based selection method could better predict the effect of the picture on the attractiveness of the presentation. Compared to manual picture selection, results of automatic pictures are similar if the manual picture is mainly decorative.

Summarization is generating content by reusing elements of existing documents. In previous chapters, corpora of text only news releases were used for evaluation. Media other than text require a different approach because non-text representations do not compare directly with text. A solution for dealing with non-textual media that has been proposed for use in multimedia summarization and retrieval is to analyze and convert the media content to a semantic representation usable by the system (Maybury and Merlino, 1997; Nagao et al., 2002; van Deemter and Power, 2003; Petrushin, 2007). However, automatic analysis of media content is difficult and often unreliable, while

manual annotation is very laborious. Another solution, which according to de Jong et al. (2007) is often overlooked, is the use of related linguistic content instead of the media items themselves. If related text adequately describes a media item, text-based retrieval methods can be used to retrieve non-textual media.

In Bosma (2005b), I proposed a method for extending the answers returned by a QA-system with appropriate illustrations by searching pictures whose related text is similar to the text of the answer. Pictures are selected by taking the best match of the answer text and a text snippet automatically associated with the picture. This method is an application of the discourse oriented summarization algorithms described in chapter 6, and has been applied in the IMIX system for answering medical questions (Boves and den Os, 2005). The purpose of the IMIX system is to answer medical questions from non-expert users, of the kind to which answers can be typically found in an encyclopedia. Questions can be typed or spoken (in Dutch), and answers are presented using speech, text and pictures. Questions can be asked in isolation, but the system is also capable of engaging in dialogs and answer follow-up questions.

This chapter presents a user evaluation of the picture selection method earlier described in Bosma (2005b). In the experiment, answer presentations with automatically selected pictures were rated by naive participants judging the attractiveness and informativeness of the text-picture combination. I also investigated the influence of the different presentations on learning. The experimental design was the same as that used by van Hooijdonk et al. (2007a), who evaluated manually created answer presentations consisting of different text-picture combinations. I repeated their experiment for answer presentations with automatically retrieved pictures, comparing two versions of the automatic picture retrieval method: one where the picture's textual annotation consists of its caption (resulting in 'caption-selected' illustrations), and one where the annotation is a part of the text near which the picture was found (resulting in 'section-selected' illustrations).

7.1 Automatic text illustration

The query-based summarization framework in chapter 6 relies on a combination of one or more feature graphs which express relations between content units. The graphs themselves are constructed using content (e.g. cosine similarity) or context (e.g. layout) to relate content units. This way, content can be presented for which there is just

indirect evidence of relevance. For instance, a sentence may be included in a summary if it is linked to the query indirectly.

This concept may also be applied to multimedia. A picture can be related to a piece of text by using layout information. A straight forward relatedness clue of text and picture is when the text is the picture's caption. But also if the picture belongs to a certain paragraph or section, the section and the picture may be considered related. When the relevance of the text is established, the relevance of the picture is established indirectly. This idea is compatible with the graph-based framework in chapter 6. If picture-text relations and text-text relations are expressed as graphs, the graph search algorithms of chapter 6 can be applied. This summarization method avoids requiring to automatically analyze the content of the picture itself (which is more difficult), or to use a manual annotation (which is more laborious).

Unfortunately, no suitable corpus is available for evaluating such a multimedia summarization system. In this chapter, instead of using a full multimedia summarization system, I focus on illustrating a given text with pictures whose relevance is based on an automatically extracted textual annotation. This may be regarded a specific instance of the approach described above.

The evaluated task is to select the best picture to illustrate a given textual answer to a medical question. The final answer presentation consists of a textual component and a picture. The textual component of the answer presentation is manually written, in order to be able to concentrate on evaluating the multimedia aspect – selecting the best picture. To find this picture, the illustration system compares the answer text with picture-associated text. The more similar the two text passages, the more likely the picture is relevant. The picture-associated text is interpreted as a textual representation of the picture. This may be either the picture's caption or the paragraph (or section if no single paragraph could be related to the picture) in which the picture was found. The relevancy of a picture is calculated as:

$$R_{picture}(i,t) = cosim(t, text(i)) \tag{7.1}$$

where $R_{picture}(i)$ is the relevancy to the text t of the picture i; and $text(i)$ is the text associated with picture i. The function $cosim(i, j)$ calculates the cosine similarity of i and j, as explained in section 6.2.3. The final answer presentation consists of the textual answer and the most relevant picture and its caption. An example of an answer presentation containing an automatic picture is given in Figure 7.1.

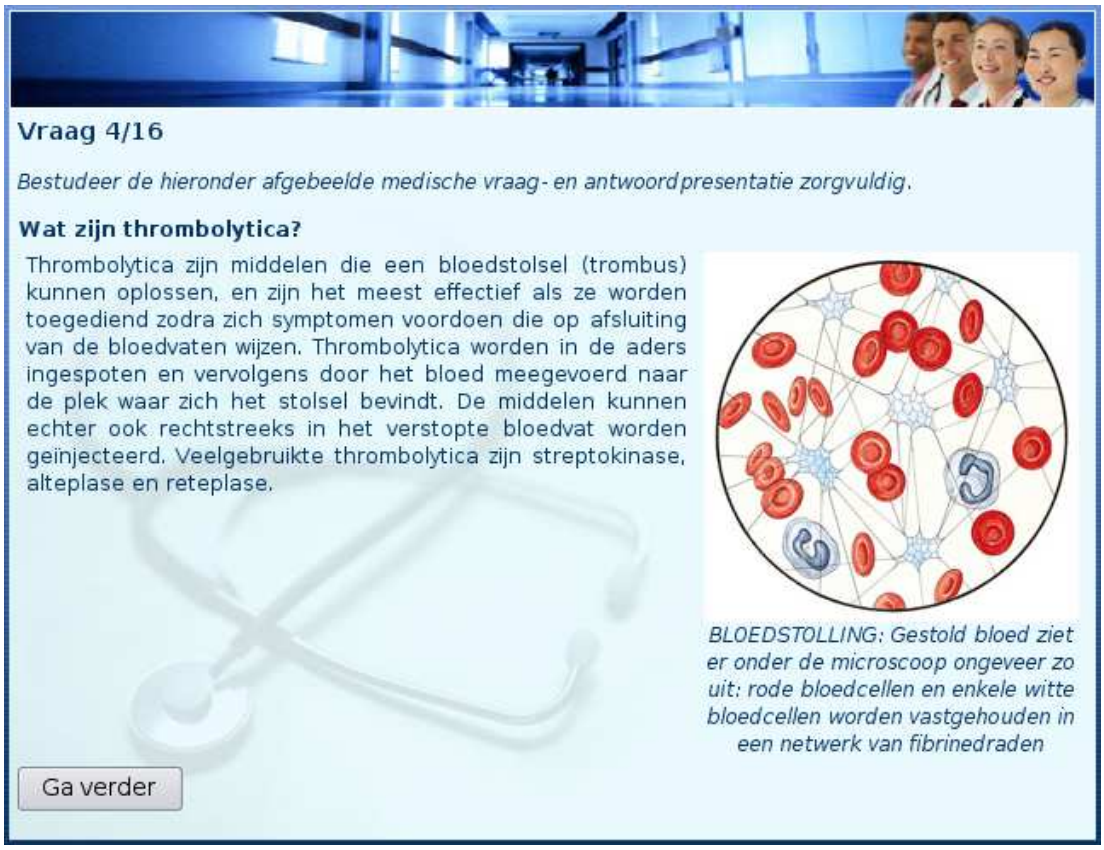

Figure 7.1: Screenshot of an answer presentation consisting of text and an automatically selected picture. The presentation answers the question *What are thrombolytics?* The text of the answer explains that thrombolytics are drugs used to dissolve blood clots. The picture depicts a schematic representation of clotted blood.

7.2 Data and methodology

The pictures as well as their textual annotations are automatically extracted from two medical sources, both intended for a general audience and written in Dutch. They provide information about anatomy, processes, diseases, treatment and diagnosis. The first source, *Merck Manual medisch handboek* (Berkow et al., 2005), Merck in short, contains 188 schematic illustrations of anatomy and treatment, process schemas, plots and various types of diagrams. The other source, *Winkler Prins medische encyclopedie* (Fiedeldij Dop and Vermeent, 1974), WP in short, contains a variety of 421 pictures, including photographic pictures, schemas and diagrams. These sources were selected because they cover the popular medical domain and they are relatively structured – paragraph boundaries are marked in the text and all 609 pictures have captions. The

Table 7.1: Examples of medical questions. Questions are equally divided in the categories of *definition questions* (Def.) or *procedure questions* (Proc.); and in questions which refer to body parts and questions which do not.

Type	Bodypart	Question
Def.	Yes	Where is testosterone produced?
Def.	No	What does ADHD stand for?
Proc.	Yes	How to apply a sling to the left arm?
Proc.	No	How to organize a workspace in order to prevent RSI?

pictures have a high information density; only few pictures are decorative. Consequently, the pictures are relatively specific to their context, which complicates their reuse in a slightly different context.

7.2.1 Questions and answers

Participants evaluated a set of answer presentations to medical questions. Apart from the presentations themselves, the study was identical to the study of manually selected answers by van Hooijdonk et al. (2007a). In van Hooijdonk et al. (2007a), we evaluated presentations consisting of a picture and text – we measured the effect of the length of the textual component of the answer (long or short) and the type of picture (no picture, a decorative picture, or an informative picture) on the participant's perception of informativeness and attractiveness. No participants took part in both the experiment described here and the experiment in van Hooijdonk et al. (2007a).

Sixteen questions in the medical domain were selected. Of the sixteen questions, half are definition questions and half are procedural questions. Of the eight questions in both groups, half refer to body parts and half do not. Table 7.1 shows examples of the questions used. References to body parts may be indirect, as is the case in the first question in Table 7.1.

For each medical question, van Hooijdonk et al. (2007a) formulated a concise and an extended textual answer. The concise answer gives a direct answer to the question, while the extended answer may also provide relevant background information (c.f. chapter 5). The average length of the concise answer and the extended answer is approximately 26 words and 66 words respectively.

In this experiment, for each of the textual answers, two presentations are generated by illustrating them using the algorithm described in section 7.1. For one of the pre-

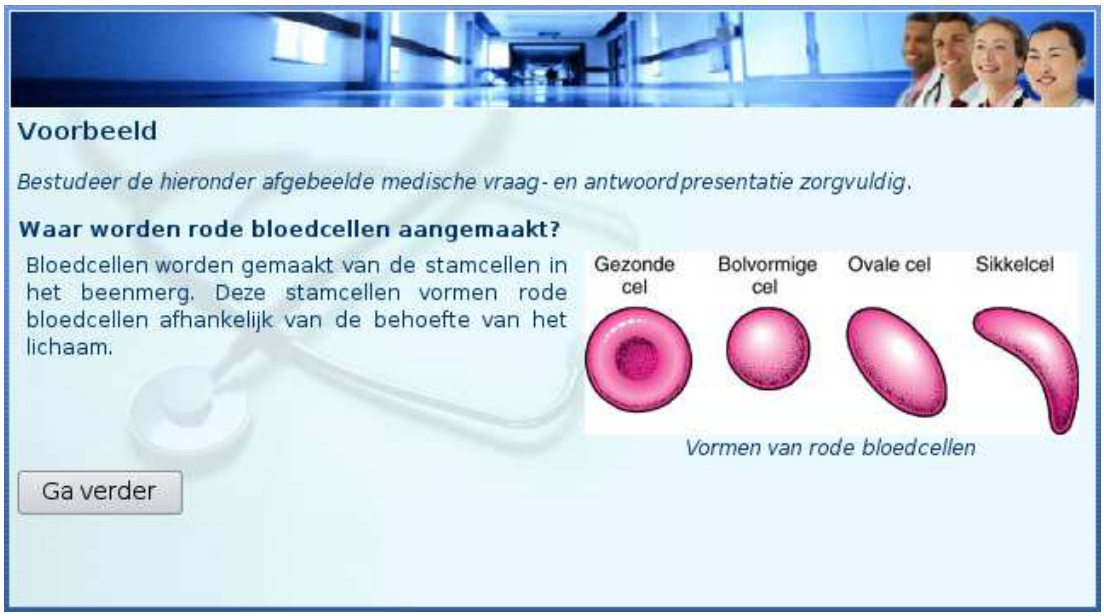

Figure 7.2: Example of a picture which is related but not complementary to the answer. The presentation answers the question: *where are red blood cells generated?*

Table 7.2: Statistics of the Merck corpus (Berkow et al., 2005) and the WP corpus (Fiedeldij Dop and Vermeent, 1974).

	Caption length (words)			Section length (words)		
	Average	SD	Range	Average	SD	Range
Merck	4.4	1.9	[1,10]	354	325	[30,1944]
WP	39.1	42.9	[1,428]	67	48	[5,336]
Combined	28.4	39.1	[1,428]	156	227	[5,1944]

sentations for each answer, the picture's caption is used as associated text, the other is associated with the smallest unit of surrounding text from the original document of the picture. This can be a section or a paragraph. Regardless which text is used for selecting the picture, the caption is considered part of the picture and is thus presented along with the picture. If the surrounding text was used for picture selection, this text is not included in the answer presentation.

The corpus did not contain an appropriate picture for all answers, which forced the illustration system to select less appropriate pictures for some of the presentations. In some cases the selected picture was plain irrelevant, but in some other cases, the picture was related to the text but had a different perspective. For instance, the picture in Figure 7.2 addresses the deformation of red blood cells rather than their generation.

Table 7.2 shows details of the distribution of lengths of associated text. Captions vary greatly in length, especially in the WP corpus. In the extreme case, the caption is as long as 428 words, while the textual component of the presentation averages 26 or 66 words (for concise and extended presentations respectively). Because some captions are presented along with pictures, this would lead to an imbalance between the amount of text in the caption and the amount of text in the textual component of the answer. In order to prevent excessive caption lengths, the caption is truncated to its first sentence *after* it is selected, so that only the caption's first sentence is presented along with the picture, rather than the caption as a whole, without affecting the picture selection process.

7.2.2 Experimental setup

Seventy five people participated in the experiment (44 female and 31 male, between 18 and 55 years old). Fifty six of them (75 percent) were students recruited from Tilburg University. None had participated in the experiments of van Hooijdonk et al. (2007a). The participants were randomly assigned to one of the four conditions (concise or extended text, selection by means of caption or surrounded text), for which they were shown all sixteen answer presentations.

The participants were invited to participate by e-mail. This e-mail shortly stated the goal of the experiment, the amount of time it would take to participate, the possibility to win a gift certificate, and the URL of the experiment. The experiment was entirely online. When the participant accessed the experiment, they first received instructions about the procedure. The participants were told that they would receive the answer presentations of 16 medical questions. They had to study these answer presentations carefully, after which they had to assess them on their informativeness and on their attractiveness. Next, the participants entered their personal data (i.e., age, gender, level of education, and optionally their e-mail to win a gift certificate).

After a participant had filled out personal data, s/he practiced the procedure of the actual experiment in a practice session: s/he was given the medical question *Where are red blood cells produced?*. First, the participant answered on a seven-point Likert scale how confident s/he was to know the answer to the medical question. Subsequently, the participant was shown the answer to the medical question corresponding to the condition s/he was assigned to. The participant studied the answer presentation until s/he thought that s/he could assess its informativeness and attractiveness. Then, the participant was shown the medical question, the answer presentation, and a questionnaire.

This questionnaire consisted of five questions, addressing:

1. the clarity of the text;
2. the informativeness of the answer presentation;
3. the attractiveness of the answer presentation;
4. the informativeness of the combination of text and picture;
5. the attractiveness of the combination of text and picture.

The participants judged the informativeness of the text-picture combination instead of directly assessing the relevance of the picture. This is because the experiment in van Hooijdonk et al. (2007a) contained manually selected pictures only, for which relevance was assumed (although a distinction was made between decorative and informative pictures). In contrast, automatic pictures may be irrelevant or somewhat relevant. However, I chose not to change the design of the experiment in order to get comparable results. (See Section 7.3.2 for a comparison between presentations with manually and automatically selected pictures.)

After completing the practice session, the participants started with the actual experiment, proceeding in the same way as during the practice session. When they finished their assessment of the answer presentations to the 16 medical questions, the participants received a post test which was the same for all participants (regardless the experimental condition). In the post test, the participants had to answer the same 16 questions of which they had rated the answer presentations in the previous part of the experiment. This was done in the form of a multiple choice test, in which each medical question was provided with four textual answer possibilities. Of these four answer possibilities, one answer was correct and the other three were plausible incorrect ones. The order in which the medical questions were presented in the post test was the same as in the actual experiment. Note that – with respect to the concise textual answer – the additional information in the extended textual answers and in the pictures was not necessary to answer the question in the post test correctly.

7.3 Results

The results of the assessments were normalized to be in the range $[0..1]$. A rating n between one and seven (inclusive) is normalized as $\frac{1}{6}(n-1)$.

For processing the results, I used the following non-standard method (c.f. section 6.2.2). For each condition and each medical question and assessment question, I calculate the average assessment. For pair-wise significance testing of differences between two experimental conditions for a particular assessment question, I measured the percentage of answer presentations for which the rating of one condition was higher than that of another. A condition that consistently received higher average ratings than the other for each medical question got a score of 100 percent; consequently, the other condition got a relative score of 0 percent. Significance is tested by means of 10^6-fold approximate randomization. A difference is considered significant if the null hypothesis (that the sets are not different) can be rejected at a certainty greater than 95 percent ($p < 0.05$), unless stated otherwise.

The reasons for using the mutual rank instead of the average judgment are similar to the reasons mentioned in chapter 6. The standard deviation of ratings of answers to some medical questions was higher than the standard deviation for answers to other medical questions. As a result, some medical questions affect the average rating more than others. This makes it less likely to find significant differences in the average rating. Using the mutual rank avoids this problem.

7.3.1 Caption or section?

Figure 7.3 shows an overview of the average assessments per condition. The level of clarity of the textual component of the answer (Figure 7.3 (a)) was judged similar. No significant differences between any two conditions was found.

Regarding the informativeness of the answer presentation as a whole (Figure 7.3 (b)), extended answers were rated significantly more informative than concise answers. However, for extended answers, the combination of picture and text (Figure 7.3 (d)) was judged less informative. This effect was the strongest for pictures which are selected using their surrounding section, although the differences were not significant.

The presentation (Figure 7.3 (c)) as well as the picture/text combination (Figure 7.3 (e)) was rated significantly more attractive if the pictures were selected by their captions than when the surrounding section was used for picture selection. The attractiveness of the presentation or the picture/text combination was not affected by the length of the textual component of the answer.

All in all, the presentations containing a section-selected picture were less informative and less attractive than the presentations containing a caption-selected picture. Apparently, captions are more representative of the content of a picture, and thus are

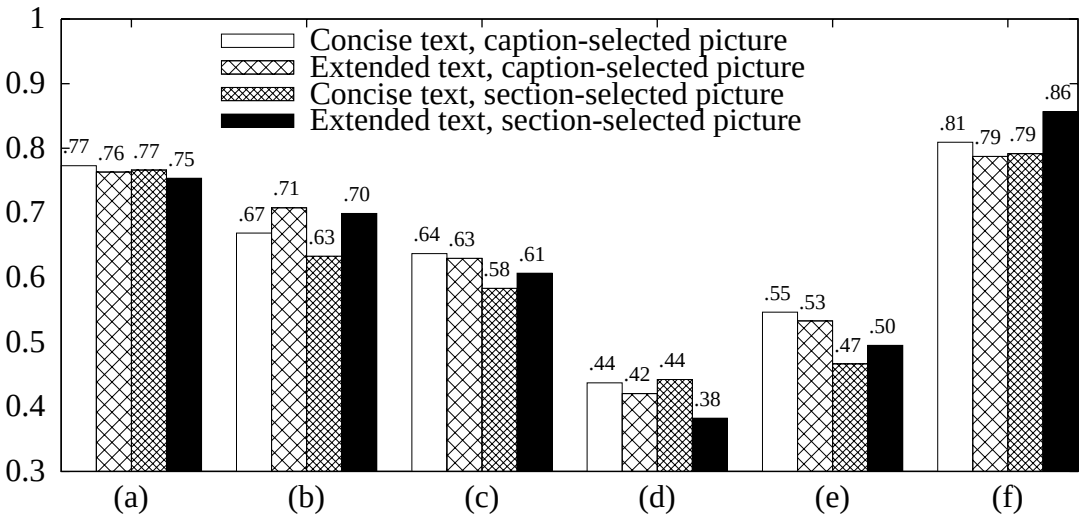

Figure 7.3: Average assessments of (a) textual clarity; (b) informativeness of the presentation; (c) attractiveness of the presentation; (d) informativeness of the text-picture combination; (e) attractiveness of the text-picture combination; and (f) the average percentage of correct answers in the post test.

more reliable indicators of the picture's relevance to the answer text. This is not entirely surprising, as the content of a caption generally describes (only) the picture, whereas the text surrounding a picture may also contain unrelated content.

In seeming contradiction with the good ratings of caption-selected pictures, in the post test where participants had to select the correct answer in a multiple choice test, participants who were shown section-selected pictures gave significantly more correct answers than other participants when the section-selected picture was included in a presentation with an extended textual component. This is a remarkable result because these pictures were rated least informative. A possible explanation for this is that the participants concentrated less on the picture (because they quickly dismissed it as less relevant) and more on the text. After all, the information in the picture was not required to answer the questions in the post test.

7.3.2 Automatic or manual?

As mentioned earlier, apart from the answer presentations themselves, the design of the experiment was identical to the experiment described in van Hooijdonk et al. (2007a). This allows us to compare the evaluation results of the automatically illustrated answer

presentations to those of van Hooijdonk et al. (2007a), who evaluated manually created answer presentations.

In the experiment of van Hooijdonk et al. (2007a), the answer presentations consisted of the same (concise or extended) textual component used in the current experiment, in combination with either no picture, a decorative picture, or an informative picture (i.e. six experimental conditions in total). These manually selected pictures can be regarded as a *gold standard* for decorative and informative pictures respectively. However, in practice, it is unlikely that this gold standard can be achieved with the set of 609 medical pictures used in the experiment for automatic picture selection, because the picture sources used by van Hooijdonk et al. (2007a) were unrestricted and thus offered far more opportunities to find a suitable illustration for a given answer text.

A large portion of participants in both experiments are students from Tilburg University who are recruited within a short time frame using the same communication channels. Therefore, I consider both groups as fully comparable. Because these students receive course credits for participation, they filled in their student registration number, which made it possible to distinguish them from other participants.

However, in both experiments, participants took part who are not part of this community. The results of the two experiments are comparable only if the group of participants in one experiment is similar to the participants of the other experiment. There are significant differences between registered students and other participants with respect to their answers to some of the assessment questions, rendering the participant groups as a whole dissimilar. The mean rating of informativeness of the presentation was rated higher by student participants than by other participants for 65 percent ($p < 0.001$) of the answer presentations of van Hooijdonk et al. (2007a). In the same experiment, students rated the text-picture combinations more informative (60 percent, $p < 0.001$) and less attractive (58 percent, $p < 0.01$) than other participants. The answers to other assessment questions were similar for both groups, or slightly different. Because students and non-students are shown to produce different results, the group of non-students are filtered out in order to ensure that the experimental conditions are the only variables over both experiments.

In total, 98 participants (70 female, 28 male) of the participants were registered students. 42 of them contributed to the experimental conditions of van Hooijdonk et al. (2007a) and 56 contributed to the conditions described in section 7.2. No one participated twice. The average assessments of the 98 participants are shown in Figure

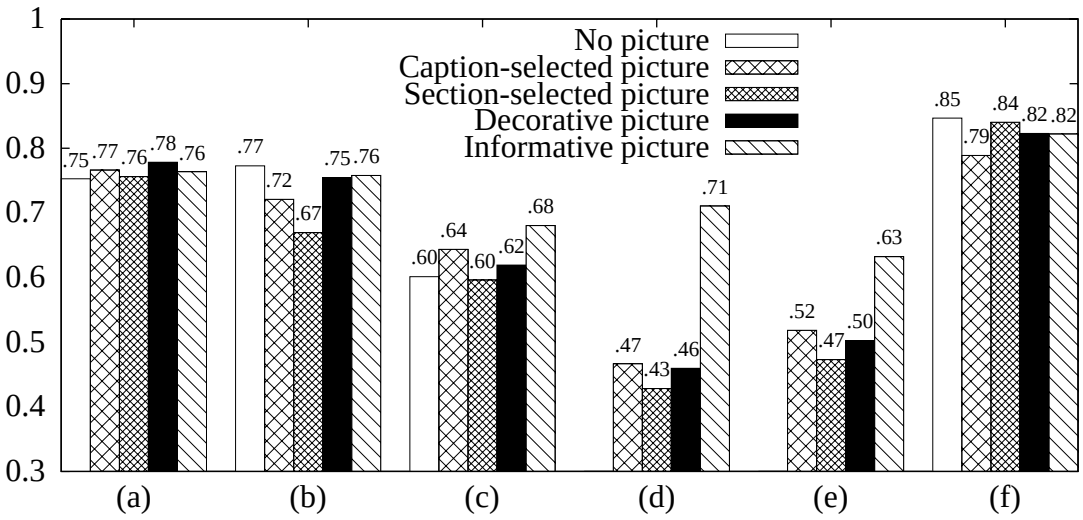

Figure 7.4: Average assessments of (a) textual clarity; (b) informativeness of the presentation; (c) attractiveness of the presentation; (d) informativeness of the text-picture combination; (e) attractiveness of the text-picture combination; and (f) the average percentage of correct answers in the post test. For comparability, these results include only registered students from Tilburg University. Therefore, the actual values may differ slightly from Figure 7.3.

7.4. These results combine the 16 concise and the 16 extended answer presentations, comprising 32 data points for each condition and assessment question.

The informativeness of text-picture combinations as well as the attractiveness of the presentation was similar when the answer contained an automatically selected picture, a manually selected decorative picture, or no picture at all. No significant differences were found. However, the text-picture combination of manually selected informative pictures was rated significantly more informative than the text-picture combination of manually selected decorative pictures and automatically selected pictures. Answer presentations were rated significantly less informative if the presentation contained a section-selected picture than if the answer contained an informative picture, a decorative picture, or no picture at all. Presentations containing caption-selected pictures are not significantly less informative than presentations with informative pictures.

Average ratings of automatic presentations may have been negatively affected by inconsistent performance of the picture selection algorithm. In some cases, the algorithm selected an irrelevant or a somewhat irrelevant picture because there was no appropriate picture in the database or simply because the algorithm failed to find it. If

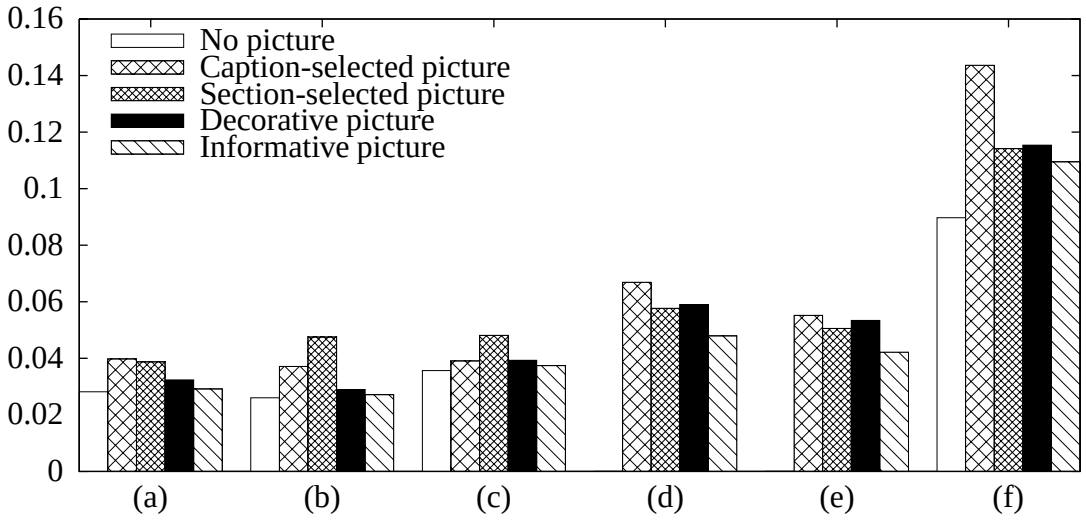

Figure 7.5: Standard deviations per answer presentation in the assessments of (a) textual clarity; (b) informativeness of the presentation; (c) attractiveness of the presentation; (d) informativeness of the text-picture combination; (e) attractiveness of the text-picture combination; and (f) the average percentage of correct answers in the post test. For comparability, these results include only registered students from Tilburg University.

the relevance of automatic pictures is less consistent than that of manual pictures, this should reflect in the variability of the results. Figure 7.5 shows the standard deviations of assessments. For automatic pictures, participants indeed show greater variability than for manual pictures in their assessments of textual clarity, informativeness and attractiveness of the answer presentation. Remarkably, the standard deviation of the number of correct answers in the post test was also greater for pictures which are selected by their captions.

7.4 The value of confidence

The selection criterion for automatic pictures was the cosine similarity of the textual component of the answer and the text associated with the picture (a caption or a section, depending on the condition). The picture with the highest cosine similarity was selected. Because cosine similarity is used as a measure of relevance, this value can be interpreted as a *confidence value*, i.e. how confident the system is that the selected picture is actually relevant. If the cosine similarity is actually a good indicator of rele-

Table 7.3: Statistics of the cosine similarity of the textual component of the answer and the text passage used for indexing the selected picture.

Condition	Average	(standard deviation)	[Range]
Brief text; caption-selected picture	0.190	(0.00788)	[0.0687,0.347]
Extended text; caption-selected picture	0.188	(0.00631)	[0.0786,0.397]
Brief text; section-selected picture	0.133	(0.00501)	[0.0295,0.311]
Extended text; section-selected picture	0.162	(0.00654)	[0.0373,0.319]

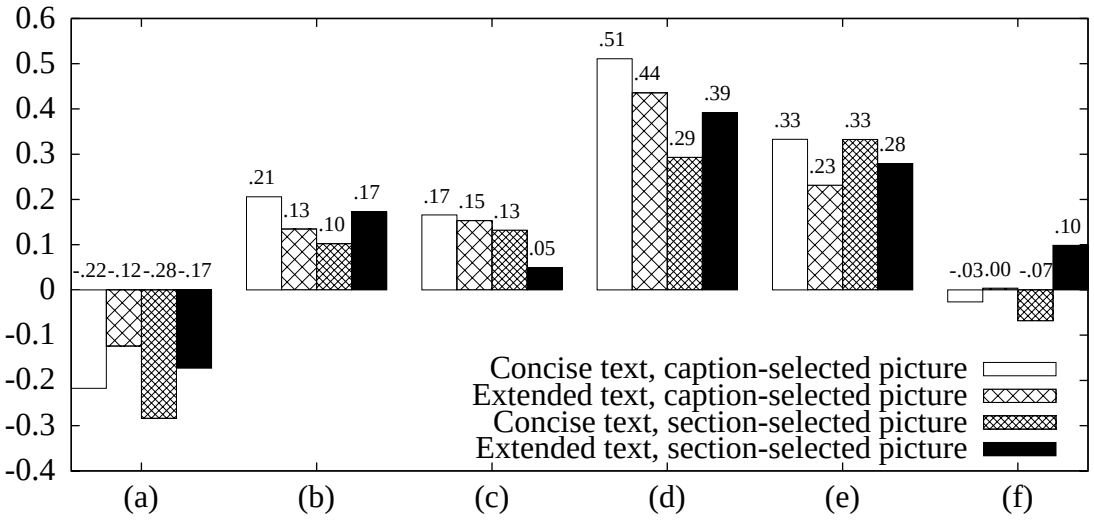

Figure 7.6: Pearson correlation coefficient between the confidence of picture selection and the assessments of (a) textual clarity; (b) informativeness of the presentation; (c) attractiveness of the presentation; (d) informativeness of the text-picture combination; (e) attractiveness of the text-picture combination; and (f) the average percentage of correct answers in the post test.

vance, one would expect a high correlation between cosine similarity and relevance. In the IMIX system (in which this picture selection method is implemented), the answer is presented text-only if no picture has a confidence (cosine similarity) above a certain (configurable) threshold. Table 7.3 shows the averages of the cosine similarity values of the pictures selected for the answers in the experiment described in this chapter.

But what is the meaning of cosine similarity as a confidence value? Cosine similarity can be used to predict the relevance of the picture if there is a correlation between the cosine similarity and the experimental participants' judgments of a presentation. Figure 7.6 shows the correlation of the confidence (cosine similarity) value and the

participant judgments. A value of 1 (or -1) indicates a perfect increasing (or decreasing) linear correlation. This correlation was greatest for the participant judgments of the informativeness of the text-picture combination (0.51 and 0.44 with concise and extended text respectively). This is an encouraging result, given that this aspect seems to correspond most closely to picture relevance. With respect to attractiveness, the correlation with confidence was significantly greater for concise answers than for extended answers. There was only a slight difference in correlation between attractiveness and confidence for different picture selection methods.

Remarkably, participants perceived the textual component of the answer as less clear when the confidence value of the picture was greater. This puzzling result suggests that relevant pictures negatively affect the clarity of the answer text rather than enhance it. A possible explanation is that any mismatches between picture and text may be more confusing when text and picture seem closely related than when the picture obviously does not fit the text, in which case it can be easily ignored and does not influence the interpretation of the text.

7.5 Conclusion

This chapter presented an algorithm for automatic illustration of answers to medical questions in Dutch. It is used in the IMIX question answering system to add appropriate illustrations to textual answers. To evaluate the algorithm, I conducted an experiment, following the same procedure as van Hooijdonk et al. (2007a) to evaluate different types of answer presentations on informativeness, attractiveness and influence on learning.

In the experiment, the answer presentations contained a textual and a visual component, of which the text was given and the visual was automatically retrieved from an offline picture database containing 609 pictures. The pictures were automatically extracted from various sources. To find an appropriate picture, the pictures were indexed by a passage of text from the document in which they were found. Two different indexing methods were compared in the experiment, either using the picture's caption for picture selection, or using the section or paragraph that contained the picture. Both selection methods were tested in combination with a concise or an extended textual answer.

Due to limitations of the corpus (i.e. for several questions it did not contain a relevant picture at all) the standard deviations of the results are quite high, which makes it

difficult to make any general claims based on them. However, some tentative conclusions can be drawn.

The results indicate that the caption-based picture selection method results in more informative and attractive presentations than the section-based method, although the difference in informativeness was not significant. Furthermore, caption-based picture selection shows a greater correlation between confidence and informativeness, which indicates that the confidence value better predicts the informativeness of the picture. A system could use this to respond by not offering any picture if no relevant picture is available (as is currently done in the IMIX system). All in all, the caption-based picture selection method offers more promising results than the section-based selection method.

An investigation of the relation between system confidence and the experimental results revealed an intriguing negative correlation between textual clarity and the predicted relevance of the selected illustration. Apparently, seeing an answer text in combination with a picture that is related to it, but not fully attuned to it, may be confusing to the user. Problems like these might be solved by the development of post-processing methods to adapt the textual and visual components of the answer presentation to each other, so that they form a more coherent whole.

When compared to manually created answer presentations, I found that answer presentations with an automatically selected picture were rated at largely the same level as presentations with a manually selected decorative picture (which did not add any information to the answer) or even no picture at all. This may be partially explained by the design of the experiment, where the visual element of the answer presentations was not needed to answer the question (since the textual element contained all the required information). Also, the results were undoubtedly influenced by the fact that the picture corpus did not contain appropriate pictures for all answers, in which case the algorithm had no choice but to select an irrelevant picture. To measure the extent of this influence, a sub-analysis could be performed on those questions for which the corpus did contain at least one appropriate picture.

8

Conclusion

This thesis is an exploration of how discourse structure can be used for query-based (multi-document) summarization systems. Summarization is a broad subject which involves many aspects of natural language processing. I highlighted a few key issues in discourse oriented summarization: utility of RST-based summaries, graph-based summarization algorithms, recognizing textual entailment, and text illustration algorithms.

A specific type of semantic relation between sentences of different documents is redundancy. This relation is of particular importance to multi-document summarization systems, but it has (to my knowledge) not been addressed in graph-based summarization algorithms. I presented a new approach to detecting textual entailment, from which redundancy can be derived. Furthermore, I proposed new performance measurements which appear to be more suitable than existing methods to detect differences in performance (chapter 3).

In order to measure the utility of discourse in query-based summarization, I designed an RST-based system for query-based summarization. A baseline system uses layout for the same summarization task. In a utility-oriented evaluation, users appear to find themselves more capable of assessing how well a summary responds to a question, if the answer is a query-based RST summary. RST summarization also reduces the amount of irrelevant information in the summary (chapter 5).

Using graphs as a mathematical concept for representing text structure, I compared a number of existing and new algorithms within a novel graph-based summarization framework. I showed that text structure can improve summarization with respect to content (Rouge2, Rouge-SU4) and readability. Furthermore, this research exposed the possible dilemma of a trade-off between coherent summarization and informative summarization (chapter 6).

The summarization framework is applied for illustrating textual answers to medical questions with pictures. An automatic system produced an answer presentation consisting of a picture and a textual component, given the text and a set of annotated candidate pictures. The relevance of the picture is based on its caption (method 1) or its

section or paragraph (method 2). Results of a comparative study suggest that a caption is more suitable than a section for determining a picture's relevance (chapter 7).

8.1 Contributions

The main contributions of this thesis are the following.

1. A quantitative analysis of evaluation methods for binary classification tasks, and recognizing textual entailment in particular (chapter 3).

2. A discourse oriented summarization method for generating coherent answers to questions (chapters 5 and 6).

3. A graph-based framework for evaluation of summarization methods (chapter 6).

4. A new graph search algorithm which beats all known graph search algorithms for summarization in Rouge-2 performance on the DUC 2005 corpus (chapter 6).

5. A text illustration method which uses circumstantial text for selecting pictures – this method is presented as a specific case of multimedia summarization (chapter 7).

8.2 Follow-up questions

Regarding features for summarization, the summarization systems described in this thesis rely on manual annotations (chapter 5) or on surface characteristics of text (chapter 6). This makes them easier to port to other languages or even language independent. However, this does not mean that more sophisticated automatic analyses cannot improve summarization. In my opinion, the most promising unexplored opportunity is to exploit other aspects of cohesion than just lexical cohesion in graph based summarization. For instance, the pronoun resolution algorithm of Lappin and Leass (1994) could be integrated in the framework described in chapter 6. Another possibility is to use methods for automatic detection of coherence relations (e.g. Marcu and Echihabi, 2002).

Regarding algorithms for summarization, the algorithms in chapter 6 used a set of feature graphs whose relative weight may be differ, depending on the configuration. I

tried to find the optimal combination of weights by varying one weight until it reached an optimum, then varying another, etc. However, there is no guarantee that the optimum is a global optimum and not a local optimum. New methods for determining the graph weights may improve the results and ease the process of finding the optimal combination of weights if the set of feature graphs changes. A possible direction is to apply genetic algorithms for weight optimization.

Progress in the field of summarization suffers from a lack of consensus on how to evaluate summaries, and in particular which content-based method is most adequate. Some content-based evaluation methods rely on a manual effort per summary (which is expensive). Both automatic and manual methods (but especially automatic methods) rely on a (possibly too) simplistic view on what 'meaning' is. They usually use phrases, n-grams, etc. as the atomic unit of meaning, failing to take their context into account. How adequate these methods are is unclear because the evaluations of these methods rely on an assumed ground truth which is disputed itself (c.f. Donaway et al., 2000; Saggion et al., 2002; Lin, 2004; Passonneau, 2005), as they focus on consistency (e.g. interannotator reliability) or on correlation with other methods. Possibly more important than the consistency of an evaluation method's judgment of the same summary (e.g. by using different annotators) is the consistency of a method's relative judgment for *different* summaries of the same system. For instance, if (by some method) system *A* is ranked consistently better than system *B*, the evaluation method measures something reproducible.

The entailment detection methods described may be useful for summarization but also for summarization evaluation. Zhou et al. (2006) show that results of a paraphrase-based evaluation method correlate better with human judgments than Rouge-1 scores.

Regarding automatic text illustration, in chapter 7, a corpus of 621 pictures was used to illustrate answers to medical questions. The quality of the final presentations then measured by means of user judgments. If the relevance of each picture to each question/answer pair was known, offline experiments could be performed to compare illustration algorithms without user judgments. This corpus might be constructed by annotating each picture manually.

A cost-effective alternative is to repeat the production experiment of van Hooijdonk et al. (2007b) in a more constrained setting, i.e. have users find the most relevant picture(s) in a fixed set of candidate pictures. In an evaluation, the resulting corpus would have a function similar to the function of reference summaries in summarization evaluation: pictures which are selected by more users are more relevant.

Finding a relevant picture may be rewarding, but knowing when *no* relevant picture can be found may be even more important. Presenting no picture is probably better than presenting an irrelevant picture. As shown in chapter 7, the cosine similarity measurement can be used to predict to a large extent the attractivity of the presentation. In addition, the original context of the picture may provide more information on the type of the picture and the type of text for which it is suitable. For instance, if the text describes a procedure, it is more likely coherent with an answer describing a procedure.

If a picture is presented with an answer, the picture is presented in another context than it is designed for. As a result, the meaning of a picture may change, possibly confusing or even misleading the user. This is potentially a risk when pictures are selected automatically and an interesting issue to investigate further.

A

Questions and answers

The user study described in chapter 5 used a set of twelve questions from the RST corpus of Carlson et al. (2002), and four semi-automatic answers for each question. This appendix lists the questions and their answers as used in this study. The questions are listed in no particular order.

Question:	How has the press been affected by Colombia's failure to take action against drug lords?
Concise answer:	The most ruthless dictatorships have not censored their press more brutally than the drug mafias censor Colombia's.
Answ. + context:	Energetic and concrete action has been taken in Colombia during the past 60 days against the mafiosi of the drug trade, but it has not been sufficiently effective. The most ruthless dictatorships have not censored their press more brutally than the drug mafias censor Colombia's.
Ext. answ. I:	Energetic and concrete action has been taken in Colombia during the past 60 days against the mafiosi of the drug trade, but it has not been sufficiently effective. The most ruthless dictatorships have not censored their press more brutally than the drug mafias censor Colombia's. The solidarity of the uncensored media world-wide against drug terrorism is the only way press freedom can survive
Ext. answ. II:	Then, when it would have been easier to resist them, nothing was done and my brother was murdered by the drug mafias three years ago. The most ruthless dictatorships have not censored their press more brutally than the drug mafias censor Colombia's. The censorship is enforced through terrorism and assassination. In the past 10 years about 50 journalists have been silenced forever, murdered.

Question:	What actions is Colombia taking to fight drug lords?
Concise answer:	Much is being done in Colombia.
Answ. + context:	Much is being done in Colombia. Luxurious homes and ranches have been raided by the military authorities, and sophisticated and powerful communications equipment have been seized.

Ext. answ. I: Colombia alone does not have the capacity. All drug-consuming countries must jointly decide to combat and punish the consumers and distributors of drugs. Much is being done in Colombia. Luxurious homes and ranches have been raided by the military authorities, and sophisticated and powerful communications equipment have been seized.

Ext. answ. II: Reduction, if not the total cessation, of drug consumption is the requirement for victory. Much is being done in Colombia to fight the drug cartel mafia. Luxurious homes and ranches have been raided by the military authorities.

Question: Does laundered drug money drive Colombia's economy?

Concise answer: In my opinion, this is not true.

Answ. + context: In my opinion, this is not true. Most of the drug money is kept in investments and in financial institutions outside Colombia.

Ext. answ. I: There has been a lot of talk that a large portion of the Colombian economy is sustained by the laundering of drug money. In my opinion, this is not true. Most of the drug money is kept in investments and in financial institutions outside Colombia.

Ext. answ. II: There has been a lot of talk that a large portion of the Colombian economy is sustained by the laundering of drug money. In my opinion, this is not true. Laundered drug money has served only to increase, unrealistically, the price of real estate.

Question: How does laundered drug money affect Colombia's economy?

Concise answer: Laundered drug money has served only to increase, unrealistically, the price of real estate.

Answ. + context: Laundered drug money has served only to increase, unrealistically, the price of real estate, creating serious problems for low-income people who aspire to own their own homes.

Ext. answ. I: Laundered drug money has served only to increase, unrealistically, the price of real estate, creating serious problems for low-income people who aspire to own their own homes. But most of the drug money is kept in investments and in financial institutions outside Colombia.

Ext. answ. II: There has been a lot of talk that a large portion of the Colombian economy is sustained by the laundering of drug money. In my opinion, this is not true. Laundered drug money has served only to increase, unrealistically, the price of real estate, creating serious problems for low-income people.

Question: What U.S. policy would benefit Colombia's economy?

Concise answer: A just price and an open market should be the policy of the U.S.

Answ. + context: A just price and an open market for what Colombia produces and exports should be the policy of the U.S.

Ext. answ. I: What is of much more importance to the Colombian economy than the supposed benefits of laundered drug money is higher prices for Colombia's legitimate products. A just price and an open market for what Colombia produces and exports should be the policy of the U.S.

Ext. answ. II: U.S. interests occasionally try to impose barriers to the import of another important Colombian export – cut flowers – into the American market. A just price and an open market for what Colombia produces and exports should be the policy of the U.S. I take advantage of this opportunity given to me by the Wall Street Journal.

Question: What is the status of Colombian coffee prices?

Concise answer: The price of coffee has gone down almost 45% since the beginning of the year, to the lowest level since the Great Depression.

Answ. + context: The price of coffee has gone down almost 45% since the beginning of the year, to the lowest level (after inflation) since the Great Depression. Market conditions point to even lower prices next year.

Ext. answ. I: What is of much more importance to the Colombian economy than the supposed benefits of laundered drug money is higher prices for Colombia's legitimate products. The price of coffee has gone down almost 45% since the beginning of the year, to the lowest level (after inflation) since the Great Depression. Market conditions point to even lower prices next year.

Ext. answ. II: What is of much more importance to the Colombian economy than the supposed benefits of laundered drug money is higher prices for Colombia's legitimate products. The price of coffee has gone down almost 45% since the beginning of the year, to the lowest level (after inflation) since the Great Depression. Market conditions point to even lower prices next year. The 27-year-old coffee cartel had to be formally dissolved this summer.

Question: What could be reasons for the dollar's weakness?

Concise answer: Analysts peg the dollar's recent weakness to an underlying slowdown in the U.S. economy.

Answ. + context: Analysts peg the dollar's recent weakness to an underlying slowdown in the U.S. economy. Narrowing interest-rate differentials between the U.S. and its major trading partners tend to make the U.S. currency less attractive to foreign investors.

Ext. answ. I: The market's strong reaction to Wall Street reflects a general uneasiness about the dollar. Analysts peg the dollar's recent weakness to an underlying slowdown in the U.S. economy. Narrowing interest-rate differentials between the U.S. and its major trading partners tend to make the U.S. currency less attractive to foreign investors.

Ext. answ. II: "The next leg could be the beginning of a longer term bearish phase." Analysts peg the dollar's recent weakness to an underlying slowdown in the U.S. economy, highlighted by recent economic data, particularly a surprisingly sharp widening in the August U.S. trade gap.

Question: How does the Fuji compare to the Red Delicious?

Concise answer: The Fuji is decidedly more dowdy.

Answ. + context: Compared to the Red Delicious, the exemplar of apple pulchritude, the Fuji is decidedly more dowdy – generally smaller, less-perfectly shaped, greenish, with tinges of red.

Ext. answ. I: The Fuji could someday tumble the Red Delicious from the top of America's apple heap. It certainly won't get there on looks. Compared to the Red Delicious, the exemplar of apple pulchritude, the Fuji is decidedly more dowdy – generally smaller, less-perfectly shaped, greenish, with tinges of red. But how sweet it is.

Ext. answ. II: It certainly won't get there on looks. Compared to the Red Delicious, the exemplar of apple pulchritude, The Fuji is decidedly more dowdy. – generally smaller, less-perfectly shaped, greenish, with tinges of red. To hear most U.S. growers tell it, we'd still be in Paradise.

Question: Why are apple growers diversifying?

Concise answer: They can protect themselves against the price vagaries of any one variety.

Answ. + context: Though growers can't always keep the worm from the apple, they can protect themselves against the price vagaries of any one variety by diversifying. "I've ripped out a lot of Delicious" and grafted the trees with many different shoots.

Ext. answ. I: The apple industry is ripe for change. Though growers can't always keep the worm from the apple, they can protect themselves against the price vagaries of any one variety by diversifying. "I've ripped out a lot of Delicious" and grafted the trees with many different shoots.

Ext. answ. II: Though growers can't always keep the worm from the apple, they can protect themselves against the price vagaries of any one variety by diversifying.

Question: What is one major difference between the Tokyo and Cannes film festivals?

Concise answer: The Tokyo International Film Festival awarded the largest cash prize of any film festival to young and first-time film makers.

Answ. + context: The Tokyo International Film Festival awarded the largest cash prize of any film festival to young and first-time film makers. By comparison, Cannes now gives $39,000 to the winner of its young director's award.

Ext. answ. I: The Tokyo International Film Festival made its mark: it awarded the largest cash prize of any film festival to young and first-time film makers. By comparison, Cannes now gives $39,000 to the winner of its young director's award.

Ext. answ. II: The Tokyo International Film Festival was no match for the Cannes Film Festival in terms of prestige, but it made its mark: it awarded the largest cash prize of any film festival to young and first-time film makers. At this year's event, the third since the festival got under way in 1985, Idrissa Ouedraogo of Burkina Faso won the Sakura Gold prize of $143,000 for "Yaaba".

Question: Who won the Young Director's award at the third Tokyo film festival?

Concise answer: At this year's event, the third since the festival got under way in 1985, Idrissa Ouedraogo of Burkina Faso won the Sakura Gold prize of $143,000 for "Yaaba".

Answ. + context: At this year's event, the third since the festival got under way in 1985, Idrissa Ouedraogo of Burkina Faso won the Sakura Gold prize of $143,000 for "Yaaba" ("Old Woman"). The Tokyo festival may become known as a major attraction for young directors because of the money as well as the recognition."

Ext. answ. I: The Tokyo International Film Festival made its mark: It awarded the largest cash prize of any film festival to young and first-time film makers. At this year's event, the third since the festival got under way in 1985, Idrissa Ouedraogo of Burkina Faso won the Sakura Gold prize of $143,000 for "Yaaba" ("Old Woman"). the Tokyo festival may become known as a major attraction for young directors because of the money as well as the recognition."

Ext. answ. II: The Tokyo International Film Festival made its mark: it awarded the largest cash prize of any film festival to young and first-time film makers. At this year's event, the third since the festival got under way in 1985, Idrissa Ouedraogo of Burkina Faso won the Sakura Gold prize of $143,000 for "Yaaba" ("Old Woman"). By comparison, Cannes now gives $39,000 to the winner of its young director's award.

Question: Why did Mahmoud Vaezi visit Paris?

Concise answer: To discuss such matters as compensation to French enterprises for contracts.

Answ. + context: To discuss such matters as compensation to French enterprises for contracts broken by the Khomeini regime.

Ext. answ. I: In Paris, Mahmoud Vaezi, Iran's vice minister of foreign affairs, began a five-day visit to discuss such matters as compensation to French enterprises for contracts broken by the Khomeini regime.

Ext. answ. II: In Paris, Mahmoud Vaezi, Iran's vice minister of foreign affairs, began a five-day visit to discuss such matters as compensation to French enterprises for contracts broken by the Khomeini regime. Toto Co., a Japanese ceramics maker, has developed a toilet that can check the user's health.

B

Sample summaries

Experiments in chapter 6 are performed on DUC 2006 data. This data set consists of fifty topics, each of which consists of a title, a query, a set of twenty five news articles, and four reference summaries. This appendix lists summaries created for topic D0650E.

The title and query for DUC topic D0650E are:

T former President Carter's international activities

Q Describe former President Carter's international efforts including activities of the Carter Center.

B.1 Human summary

One of four reference summaries written manually by NIST abstractors for DUC 2006 topic D0650E:

35A Former President Jimmy Carter has played an active role on the international stage.

35B Working in conjunction with the Carter Center that he founded and the National Democratic Institute of International Affairs, Carter has led many international teams of observers to monitor elections throughout the world.

35C By 1999 he had monitored more than 20 elections in 16 countries including Nicaragua, Liberia, Nigeria, Venezuela, China (village and township committee elections), Indonesia, Mozambique, Peru, the Dominican Republic and Mexico.

35D Carter has visited 115 countries to promote peace and human rights or to combat disease and hunger.

35E He is credited with gaining release of approximately 50,000 political prisoners, not hesitating to meet personally with such leaders as Yasser Arafat, Kim-Il-Sung, Daniel Ortega, Haitian bully Raoal Cedras and Bosnian Serb Radovan Karadzic.

35F In 1998 the Carter Center devoted a $1.5 million grant from Coca-Cola to ”Transparency for Growth in The Americas,” a program for combating corruption through o penness in government.

35G Through the Carter Center, the former President has led successful campaigns against two devastating diseases: river blindness and Guineau worm disease.

35H In 1999 the Center received grants totaling $30 million to expand programs for treatment and prevention of blindness in Africa and Latin America and to extend its anti-trachoma program from Mali to Sudan, Ethiopia, Nigeria, Niger, Ghana and Yemen.

35I In August 1999 President Clinton awarded former President Carter the Medal of Freedom for his successes in the struggle for peace and human rights and against disease and hunger.

B.2 Query-relevance summary

Query-relevance summary generated for DUC 2006 topic D0650E, Δ_q:

36A Haas said that the Coca-Cola Co. does not have any control over Carter Center projects, but he is not concerned about any backlash that could affect his company from the Carter Center’s efforts.

36B So it worked out nicely to confer the medals on the former president and his first lady at the Carter Center.

36C Founded in 1982 by former U.S. president Jimmy Carter and his wife, the Carter Center has been devoted to observing many countries in respect of democracy and development, human rights issues and public health.

36D The Carter Center, founded by former U.S. President Jimmy Carter, has been fighting blindness since 1987.

36E Former U.S. President Jimmy Carter will head a group of international observers during Peru’s elections of May 28, Barry Levitt, official of Carter Center told press on Tuesday.

36F Spokeswoman Deanna Congileo said the delegation would not include former President Jimmy Carter, despite reports in local newspapers that Carter and four other former heads of state would be observers.

36G The Central Elections Board said Wednesday it had accredited a host of former heads of state to serve as elections observers, including Carter, former Spanish Prime Minister Felipe Gonzalez, ex-President Luis Herrera Campins of Venezuela and former President Armando Calderon Sol of El Salvador.

36H The former president was invited by all of Mexico’s major political parties.

36I After Carter left, voter Antonio Arreguin, 49, said, “The former president?”

B.3 Query-distance summary

Query-distance summary generated for DUC 2006 topic D0650E, $\Delta_{q,c}$ ($w_q = 0.5$; $w_c = 1.0$):

37A Haas said that the Coca-Cola Co. does not have any control over Carter Center projects, but he is not concerned about any backlash that could affect his company from the Carter Center’s efforts.

37B So it worked out nicely to confer the medals on the former president and his first lady at the Carter Center.

37C Spokeswoman Deanna Congileo said the delegation would not include former President Jimmy Carter, despite reports in local newspapers that Carter and four other former heads of state would be observers.

37D The Central Elections Board said Wednesday it had accredited a host of former heads of state to serve as elections observers, including Carter, former Spanish Prime Minister Felipe Gonzalez, ex-President Luis Herrera Campins of Venezuela and former President Armando Calderon Sol of El Salvador.

37E The former president was invited by all of Mexico's major political parties.

37F After Carter left, voter Antonio Arreguin, 49, said, "The former president?"

37G Through the Global 2000 River Blindness Program in collaboration with the National and District Health Services, the center has been involved in the control of river blindness in Uganda since 1996.

37H The Carter Center, founded by former U.S. President Jimmy Carter, has been fighting blindness since 1987.

37I Former U.S. President Jimmy Carter will head a group of international observers during Peru's elections of May 28, Barry Levitt, official of Carter Center told press on Tuesday.

37J Levitt, director of Political Analysis of this organization, said that Carter is worried by Peru's electoral process, which has been questioned by international observers.

B.4 Normalized centrality summary

Normalized centrality summary generated for DUC 2006 topic D0650E, $\Delta_{q,c}$:

38A A 35-member mission headed by former United States President Jimmy Carter will arrive in Nicaragua next Thursday to observe the general elections to be held on October 20 in this country.

38B The Carter Center, founded by former U.S. President Jimmy Carter, has been fighting blindness since 1987.

38C Former U.S. President Jimmy Carter will head a group of international observers during Peru's elections of May 28, Barry Levitt, official of Carter Center told press on Tuesday.

38D The events that allowed a former coup plotter to become front-runner in Sunday's presidential vote indicate that Venezuelans are fed up with the status quo, said former U.S. President Jimmy Carter.

38E He is heading a 40-person team of international election monitors that includes former presidents Patricio Aylwin of Chile and Gonzalo Sanchez de Lozada of Bolivia and Nicholas Brady, an ex-U.S. treasury secretary.

38F Spokeswoman Deanna Congileo said the delegation would not include former President Jimmy Carter, despite reports in local newspapers that Carter and four other former heads of state would be observers.

38G The Central Elections Board said Wednesday it had accredited a host of former heads of state to serve as elections observers, including Carter, former Spanish Prime Minister Felipe Gonzalez, ex-President Luis Herrera Campins of Venezuela and former President Armando Calderon Sol of El Salvador.

38H So far only former Colombian President Belisario Betancur has said he would attend, the board said.

B.5 Probabilistic centrality summary

Probabilistic centrality summary generated for DUC 2006 topic D0650E, $\Delta_{q,c}$ ($w_q = 1.0$; $w_c = 0.1$):

39A So it worked out nicely to confer the medals on the former president and his first lady at the Carter Center.

39B Former U.S. President Jimmy Carter will be in Mexico on July 2 to observe this country's highly competitive presidential election, an elections monitoring groups said Tuesday.

39C Founded in 1982 by former U.S. president Jimmy Carter and his wife, the Carter Center has been devoted to observing many countries in respect of democracy and development, human rights issues and public health.

39D Former United States President Jimmy Carter is scheduled to arrive in Nigeria next Tuesday to monitor the on- going transition program in the country.

39E The Carter Center, founded by former U.S. President Jimmy Carter, has been fighting blindness since 1987.

39F Former U.S. President Jimmy Carter will head a group of international observers during Peru's elections of May 28, Barry Levitt, official of Carter Center told press on Tuesday.

39G Spokeswoman Deanna Congileo said the delegation would not include former President Jimmy Carter, despite reports in local newspapers that Carter and four other former heads of state would be observers.

39H The Central Elections Board said Wednesday it had accredited a host of former heads of state to serve as elections observers, including Carter, former Spanish Prime Minister Felipe Gonzalez, ex-President Luis Herrera Campins of Venezuela and former President Armando Calderon Sol of El Salvador.

39I The former president was invited by all of Mexico's major political parties.

39J After Carter left, voter Antonio Arreguin, 49, said, "The former president?"

B.6 Redundancy-aware normalized centrality summary

Redundancy-aware normalized centrality summary generated for DUC 2006 topic D0650E, $\Delta_{q,c,r}$:

40A President Alberto Fujimori has promised to comply with recommendations of a Carter Center delegation that said Peru does not meet international standards for fair elections.

40B Spokeswoman Deanna Congileo said the delegation would not include former President Jimmy Carter, despite reports in local newspapers that Carter and four other former heads of state would be observers.

40C So it worked out nicely to confer the medals on the former president and his first lady at the Carter Center.
40D Former U.S. President Jimmy Carter will be in Mexico on July 2 to observe this country's highly competitive presidential election, an elections monitoring groups said Tuesday.
40E The former president was invited by all of Mexico's major political parties.
40F A 35-member mission headed by former United States President Jimmy Carter will arrive in Nicaragua next Thursday to observe the general elections to be held on October 20 in this country.
40G Founded in 1982 by former U.S. president Jimmy Carter and his wife, the Carter Center has been devoted to observing many countries in respect of democracy and development, human rights issues and public health.
40H Former United States President Jimmy Carter is scheduled to arrive in Nigeria next Tuesday to monitor the on- going transition program in the country.
40I The Carter Center, founded by former U.S. President Jimmy Carter, has been fighting blindness since 1987.
40J Former U.S. President Jimmy Carter will head a group of international observers during Peru's elections of May 28, Barry Levitt, official of Carter Center told press on Tuesday.

B.7 Redundancy-aware probabilistic centrality summary

Redundancy-aware probabilistic centrality summary generated for DUC 2006 topic D0650E, $\Delta_{q,c,r}$ ($w_q = 1$; $w_c = 0.1$; $w_{rP} = 1.0$):

41A The Carter Center, founded by former President Jimmy Carter, made its conclusion in conjunction with the National Democratic Institute, another U.S. based pro-democracy group, after a five-day visit at the invitation of Fujimori's administration.
41B President Alberto Fujimori has promised to comply with recommendations of a Carter Center delegation that said Peru does not meet international standards for fair elections.
41C Spokeswoman Deanna Congileo said the delegation would not include former President Jimmy Carter, despite reports in local newspapers that Carter and four other former heads of state would be observers.
41D So it worked out nicely to confer the medals on the former president and his first lady at the Carter Center.
41E Former U.S. President Jimmy Carter will be in Mexico on July 2 to observe this country's highly competitive presidential election, an elections monitoring groups said Tuesday.
41F The former president was invited by all of Mexico's major political parties.
41G Founded in 1982 by former U.S. president Jimmy Carter and his wife, the Carter Center has been devoted to observing many countries in respect of democracy and development, human rights issues and public health.
41H Former United States President Jimmy Carter is scheduled to arrive in Nigeria next Tuesday to monitor the on- going transition program in the country.
41I The Carter Center, founded by former U.S. President Jimmy Carter, has been fighting blindness since 1987.
41J Former U.S. President Jimmy Carter will head a group of international observers during Peru's elections of May 28, Barry Levitt, official of Carter Center told press on Tuesday.

B.8 Summary submitted to DUC 2006

Summary submitted to DUC 2006 for topic D0650E:

42A The Central Elections Board said Wednesday it had accredited a host of former heads of state to serve as elections observers, including Carter, former Spanish Prime Minister Felipe Gonzalez, ex-President Luis Herrera Campins of Venezuela and former President Armando Calderon Sol of El Salvador.

42B The Carter Center, founded by former President Jimmy Carter, said Friday that Fujimori's government was abusing public funds to garner votes, had blocked opposition candidates 'access to the media, and had used other irregular practices to gain an unfair advantage in the April 9 vote.

42C Former U.S. President Jimmy Carter will head a group of international observers during Peru's elections of May 28, Barry Levitt, official of Carter Center told press on Tuesday.

42D Founded in 1982 by former U.S. president Jimmy Carter and his wife, the Carter Center has been devoted to observing many countries in respect of democracy and development, human rights issues and public health.

42E According to Gordon Streeb, leader of the delegation, the former US president Jimmy Carter would lead a delegation of about 60 observers to Liberia as a reflection of the importance the center attaches to the success of the country's post-war elections.

42F "If the voting process and the counting of the ballots and the reporting of the results are deemed to be transparent and without flaws of any significant degree," Carter said, "they indeed will accept the results of the election – even though they retain the right to criticize some previous activities during the campaign."

Bibliography

H. J. A. op den Akker, H. C. Bunt, S. Keizer, and B. W. van Schooten. From question answering to spoken dialogue - towards an information search assistant for interactive multimodal information extraction. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 2793–2796. ESCA, 2005.

R. Alterman. Understanding and summarization. *Artificial Intelligence Review*, 5(4): 239–254, 1991.

E. André. *The Generation of Multimedia Presentations*, chapter 12, pages 305–327. Marcel Dekker, 2000. ISBN 978-0-8247-9000-4.

E. André. *Ein plan-basierter Ansatz zur Generierung multimedialer Präsentationen*. PhD thesis, Universität des Saarlandes, Saarbrücken, Germany, 1995.

K. Bakshi, D. Huynh, B. Katz, , D. R. Karger, J. Lin, D. Quan, and V. Sinha. What makes a good answer? the role of context in question answering. In *Proceedings of the 9th IFIP TC13 International Conference on Human-Computer Interaction*, Zürich, Switzerland, 2003a.

K. Bakshi, D. Huynh, B. Katz, D. R. Karger, J. Lin, D. Quan, and V. Sinha. The role of context in question answering systems. In *CHI '03 extended abstracts on Human Factors in Computing Systems*, pages 1006–1007, New York, NY, USA, 2003b. ACM Press.

M. Banko, V. O. Mittal, and M. J. Witbrock. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 318–325, Hong Kong, Oct. 2000.

C. Bannard and C. Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the ACL*, 2005.

R. Bar-Haim, I. Szpektor, and O. Glickman. Definition and analysis of intermediate entailment levels. In *Proceedings of the ACL workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 2005.

R. Bar-Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor. The second PASCAL recognising textual entailment challenge. In B. Magnini and I. Dagan, editors, *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge*, Trento, Italy, Apr. 2006.

R. Barzilay. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. PhD thesis, Columbia University, 2003.

R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL workshop on Intelligent Scalable Text Summarization*, pages 10–17, Aug. 1997.

R. Barzilay and L. Lee. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the conference of HLT/NAACL*, pages 16–23, Morristown, NJ, USA, 2003.

R. Barzilay, N. Elhadad, and K. R. McKeown. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55, 2002.

M. J. Bates. The berry-picking search: user interface design. In H. Thimbleby, editor, *User Interface Design*. Addison-Wesley, 1990.

P. B. Baxendale. Machine-made index for technical literature: an experiment. *IBM Journal of Research and Development*, 2(4):354–361, 1958.

R. Berkow, M. H. Beers, and A. J. Fletcher, editors. *Merck manual medisch handboek*. Bohn Stafleu van Loghum, Houten, the Netherlands, 2nd edition, 2005. ISBN 978-90-3134300-3.

S. Blair-Goldensohn and K. McKeown. Integrating rhetorical-semantic relation models for query-focused summarization. In *Proceedings of the Document Understanding Conference*, New York, NY, USA, 2006.

R. A. Bolt. "Put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, pages 262–270, New York, NY, USA, 1980. ACM Press. ISBN 0-89791-021-4.

A. van den Bosch and W. Daelemans. Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 285–292, San Francisco, CA, USA, June 1999.

A. van den Bosch, S. Canisius, W. Daelemans, I. Hendrickx, and E. Tjong Kim Sang. Memory-based semantic role labeling: Optimizing features, algorithm, and output. In H. T. Ng and E. Riloff, editors, *Proceedings of the 8th Conference on Computational Natural Language Learning*, pages 102–105, Boston, MA, USA, May 2004.

W. E. Bosma. Extending answers using discourse structure. In H. Saggion and J.-L. Minel, editors, *Crossing Barriers in Text Summarization Research*, pages 2–9, Shoumen, Bulgaria, Sept. 2005a. Incoma Ltd. ISBN 954-909068-X.

W. E. Bosma. Image retrieval supports multimedia authoring. In E. Zudilova-Seinstra and T. Adriaansen, editors, *Linguistic Engineering meets Cognitive Engineering in Multimodal Systems*, ICMI workshop, pages 89–94, Trento, Italy, Oct. 2005b. ITC-irst.

W. E. Bosma. Query-based summarization for question answering. In T. van der Wouden, M. Poß, H. Reckman, and C. Cremers, editors, *Proceedings of the 15th Meeting of CLIN*, Leiden, the Netherlands, 2005c.

W. E. Bosma. Query-based extracting: how to support the answer? In *Proceedings of the Document Understanding Conference*, pages 202–208, New York, NY, USA, 2006.

W. E. Bosma and C. Callison-Burch. Paraphrase substitution for recognizing textual entailment. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multimodal Information Retrieval*, Lecture Notes in Computer Science, pages 502–509. Springer Verlag, Berlin, Germany, 2007.

G. Bouma, J. Mur, G. van Noord, L. van der Plas, and J. Tiedemann. Question answering for Dutch using dependency relations. In *Proceedings of the CLEF2005 workshop*, Lecture Notes in Computer Science. Springer, 2006.

G. Bouma, I. Fahmi, J. Mur, G. van Noord, L. van der Plas, and J. Tiedemann. Using syntactic knowledge for QA. In *Evaluation of Multilingual and Multi-modal Information Retrieval*, volume 4730 of *Lecture Notes in Computer Science*, pages 318–327. Springer, Berlin, Germany, 2007. ISBN 978-3-540-74998-1.

L. Boves and E. den Os. Interactivity and multimodality in the IMIX demonstrator. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1578–1581. IEEE Computer Society, 2005. ISBN 0-7803-9331-7.

R. Brandow, K. Mitze, and L. F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5): 675–685, 1995.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, New York, NY, USA, 1998. ACM Press. ISBN 1-58113-015-5.

L. Carlson, D. Marcu, and M. E. Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the second SIGdial workshop on Discourse and Dialogue*, pages 1–10, Morristown, NJ, USA, 2001.

L. Carlson, D. Marcu, and M. E. Okurowski. RST discourse treebank. Linguistic Data Consortium, Philadelphia, PA, USA, 2002.

L. Carlson, D. Marcu, and M. E. Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith, editors, *Current Directions in Discourse and Dialogue*, volume 22 of *Text, speech and language technology series*. Kluwer Academic Publishers, 2003. ISBN 1-40201-614-X.

H. Chen and T. D. Ng. An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): symbolic branch-and-bound search vs. connectionist hopfield net activation. *Journal of the American Society for Information Science*, 46(5):348–369, June 1995.

L. L. Cherry and W. Vesterman. Writing tools: The STYLE and DICTION programs. Technical report CS-91, Bell Laboratories, Murray Hill, NJ, USA, 1981.

J. M. Conroy and J. D. Schlesinger. Back to basics: Classy 2006. In *Proceedings of the Document Understanding Conference*, New York, NY, USA, 2006.

S. H. Corston-Oliver. Beyond string matching and cue phrases: Improving efficiency and coverage in discourse analysis. Technical report MSR-TR-98-66, Microsoft Research, Nov. 1998.

E. T. Cremmins. *The Art of Abstracting*. ISI Press, Philadelphia, PA, USA, 1982. ISBN 0-89495-015-0.

M. Crochemore and W. Rytter. *Text algorithms*. Oxford University Press, Inc., New York, NY, USA, 1994. ISBN 0-19-508609-0.

I. Dagan, O. Glickman, and B. Magnini. The PASCAL recognising textual entailment challenge. In *PASCAL Challenges workshop on Recognising Textual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190, Berlin, Germany, 2006. Springer. ISBN 978-3-540-33427-9.

H. T. Dang. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference*, 2005.

H. T. Dang. Overview of DUC 2006. In *Proceedings of the Document Understanding Conference*. NIST, 2006.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

J. Delin, J. Bateman, and P. Allen. A model of genre in document layout. *Information Design Journal*, 11(1):54–66, 2002.

E. Dijkstra. A note on two problems in connection with graphs. *Numerische Mathematik*, 1:269–271, 1959.

R. L. Donaway, K. W. Drummey, and L. A. Mather. A comparison of rankings produced by summarization evaluation measures. In *NAACL/ANLP workshop on Automatic Summarization*, pages 69–78, Morristown, NJ, USA, 2000.

P. Fiedeldij Dop and S. Vermeent, editors. *Winkler Prins medische encyclopedie*. Spectrum, 3rd edition, 1974. ISBN 978-90-1000997-5.

L. L. Earl. Experiments in automatic extracting and indexing. *Information Storage and Retrieval*, 6:313–334, 1970.

H. P. Edmundson. New methods in automatic extracting. *Journal of the ACM*, 16(2): 264–285, Apr. 1969.

B. Endres-Niggemeyer. Simsum: an empirically founded simulation of summarizing. *Information Processing and Management*, 36(4):659–682, 2000.

B. Endres-Niggemeyer. *Summarizing Information*. Springer, 1998. ISBN 978-3-54063735-6.

G. Erkan and D. R. Radev. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 2004.

D. K. Evans. The future of multilingual summarization: Beyond sentence extraction. In F. Gey, N. Kando, C.-Y. Lin, and C. Peters, editors, *Proceedings of New Directions in Multilingual Information Access*, Aug. 2006.

R. P. Futrelle. Summarization of diagrams in documents. In I. Mani and M. T. Maybury, editors, *Advances in Automated Text Summarization*. MIT Press, Cambridge, MA, USA, July 1999. ISBN 978-0-262-13359-3.

J. Geurts, J. R. van Ossenbruggen, and L. Hardman. Requirements for practical multimedia annotation. In *Proceedings of the workshop on Multimedia and the Semantic Web*, Heraklion, Crete, May 2005.

D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL/PASCAL workshop on Textual Entailment and Paraphrasing*, pages 1–9, June 2007.

J. Goldstein and J. Carbonell. Summarization: (1) using MMR for diversity-based reranking and (2) evaluating summaries. In *Proceedings of TIPSTER Text Program Phase III*, pages 181–195, Morristown, NJ, USA, 1996.

J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 121–128, New York, NY, USA, 1999. ACM Press. ISBN 1-58113-096-1.

Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval, pages 19–25, New York, NY, USA, 2001. ACM Press. ISBN 1-58113-331-6.

B. J. Grosz and C. L. Sidner. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12:(3):175–204, 1986.

B. J. Grosz, S. Weinstein, and A. K. Joshi. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

M. A. Halliday and R. Hasan. *Cohesion in English*. Longman, London, United Kingdom, Sept. 1976. ISBN 0-582-55041-6.

H. van Halteren and S. Teufel. Examining the consensus between human summaries: initial experiments with factoid analysis. In *Proceedings of the HLT/NAACL workshop on Text Summarization*, volume 5, pages 57–64, Morristown, NJ, USA, 2003.

A. Hämäläinen, L. Boves, J. de Veth, and L. ten Bosch. On the utility of syllable-based acoustic models for pronunciation variation modelling. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007(2), 2007.

D. Harman. How effective is suffixing? *Journal of the American Society for Information Science*, 42(1), Jan. 1999.

A. Harnly, A. Nenkova, R. Passonneau, and O. Rambow. Automation of summary evaluation by the Pyramid method. In *Proceedings of the Conference on Recent Advances in Natural Language Processing 2005*, 2005.

J. Hirschberg and D. Litman. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530, 1993.

L. Hirschman and R. Gaizauskas. Natural language question answering: The view from here. *Natural Language Engineering*, 7:275–300, 2001.

J. R. Hobbs. On the coherence and structure of discourse. Technical report CSLI-85-37, Center for the Study of Language and Information, Stanford University, 1985.

J. R. Hobbs. Resolving pronoun references. In B. J. Grosz, K. Spärck Jones, and B. Lynn-Webber, editors, *Readings in natural language processing*, pages 339–352. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1986. ISBN 0-934613-11-7.

C. M. J. van Hooijdonk, J. de Vos, E. J. Krahmer, A. Maes, M. Theune, and W. E. Bosma. On the role of visuals in multimodal answers to medical questions. In *Proceedings of the Conference of the IEEE Professional Communication Society*, 2007a.

C. M. J. van Hooijdonk, E. J. Krahmer, A. Maes, M. Theune, and W. E. Bosma. Towards automatic generation of multimodal answers to medical questions: a cognitive engineering approach. In I. van der Sluis, M. Theune, E. Reiter, and E. Krahmer, editors, *Proceedings of the workshop on Multimodal Output Generation*, CTIT Workshop Proceedings, pages 93–104, Enschede, the Netherlands, 2007b.

E. H. Hovy, C.-Y. Lin, and L. Zhou. A BE-based multi-document summarizer with sentence compression. In *Proceedings of the ACL workshop on Multilingual Summarization Evaluation*, 2005a.

E. H. Hovy, C.-Y. Lin, and L. Zhou. Evaluating DUC 2005 using Basic Elements. In *Proceedings of the Document Understanding Conference*, 2005b.

E. H. Hovy, C.-Y. Lin, L. Zhou, and J. Fukumoto. Automated summarization evaluation with basic elements. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006.

J. Jagarlamudi, P. Pingali, and V. Varma. Query independent sentence scoring approach to DUC 2006. In *Proceedings of the Document Understanding Conference*, New York, NY, USA, 2006.

J. Janoš. Theory of functional sentence perspective and its application for the purposes of automatic extracting. *Information Processing and Management*, 15:19–25, 1979.

N. Jindal and B. Liu. Identifying comparative sentences in text documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 244–251, Seattle, WA, USA, 2006. ACM Press. ISBN 1-59593-369-7.

H. Jing. Sentence reduction for automatic text summarization. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 310–315, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

H. Jing and K. R. McKeown. The decomposition of human-written summary sentences. In M. Hearst, F. Gey, and R. Tong, editors, *Proceedings of the 22nd Annual*

International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 129–136, New York, NY, USA, Aug. 1999. ACM Press. ISBN 1-58113-096-1.

H. Jing, K. McKeown, R. Barzilay, and M. Elhadad. Summarization evaluation methods: experiments and analysis. In *American Association for Artificial Intelligence Spring Symposium Series*, pages 60–68, Mar. 1998.

F. M. G. de Jong, T. Westerveld, and A. P. de Vries. Multimedia search without visual analysis: the value of linguistic and contextual information. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):365–371, 2007.

W. Kintsch and T. van Dijk. Toward a model of text comprehension and production. *Psychological Review*, (85):363–394, 1978.

G. Klein, B. Sundheim, L. Hirschman, I. Mani, T. Firmin, and D. House. The TIPSTER SUMMAC text summarization evaluation. In *Proceedings of EACL'99*, Bergen, Norway, 1999.

K. Knight and D. Marcu. Statistics-based summarization – step one: Sentence compression. In *Proceedings of the 17th National Conference on Artificial Intelligence and the 12th Conference on Innovative Applications of Artificial Intelligence*, pages 703–710, Austin, Texas, July 2000. AAAI Press. ISBN 0-262-51112-6.

A. Knott and R. Dale. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62, 1995.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*, 2005.

G.-J. M. Kruijff and I. Kruijff-Korbayová. A hybrid logic formalization of information structure sensitive discourse interpretation. In *Proceedings of the 4th International Conference on Text, Speech and Dialogue*, pages 31–38, London, UK, 2001. Springer. ISBN 3-540-42557-8.

J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, Seattle, Washington, USA, 1995. ACM Press. ISBN 0-89791-714-6.

T. K. Landauer, P. W. Foltz, and D. Laham. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

M. Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 545–552, July 2003.

S. Lappin and H. J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

A. Lascarides and N. Asher. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493, Oct. 1993.

J. R. Levin. On functions of pictures in prose. In F. J. Pirozzolo and M. C. Wittrock, editors, *Neuropsychological and Cognitive Processes in Reading*, pages 203–228. Academic Press, New York, NY, USA, 1981. ISBN 978-0-12-557360-3.

C.-Y. Lin. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the ACL workshop: Text Summarization Branches Out*, Barcelona, Spain, 2004.

C.-Y. Lin and E. H. Hovy. Automated multi-document summarization in NeATS. In *Proceedings of the 2nd International Conference on Human Language Technology Research*, pages 59–62, San Francisco, CA, USA, 2002a. Morgan Kaufmann Publishers Inc.

C.-Y. Lin and E. H. Hovy. Manual and automatic evaluation of summaries. In *Proceedings of the workshop on Automatic Summarization*, pages 45–51, July 2002b.

C.-Y. Lin and E. H. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the Conference on HLT/NAACL*, volume 1, pages 71–78, Morristown, NJ, USA, May 2003.

C.-Y. Lin and E. H. Hovy. Identifying topics by position. In *Proceedings of the 5th Conference on Applied natural language processing*, pages 283–290, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

D. Lin. Dependency-based evaluation of minipar. In *Proceedings of LREC workshop on the Evaluation of Parsing Systems*, Granada, Spain, 1998.

D. Lin and P. Pantel. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(4):343–360, 2001.

H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

N. Madnani, R. Passonneau, J. Conroy, N. F. Ayan, B. Dorr, J. Klavans, D. O'Leary, and J. Schlesinger. Measuring variability in sentence ordering for news summarization. In *Proceedings of the 11th European workshop on Natural Language Generation*, June 2007.

O. Manabu and M. Hajime. Query-biased summarization based on lexical chaining. *Computational Intelligence*, 16(4):578–585, Nov. 2000.

I. Mani. *Automatic Summarization*, volume 3 of *Natural language processing*. John Benjamins, Amsterdam, the Netherlands, 2001. ISBN 90-2724986-5.

I. Mani and E. Bloedorn. Multi-document summarization by graph search and matching. In *Proceedings of the 14th National Conference on Artificial Intelligence (AAAI'97)*, pages 622–628, 1997.

I. Mani and E. Bloedorn. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):35–67, 1999.

I. Mani, B. Gates, and E. Bloedorn. Improving summaries by revising them. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 558–565, Morristown, NJ, USA, 1999. ISBN 1-55860-609-2.

W. C. Mann and S. A. Thompson. Toward a theory of reading between the lines: An exploration in discourse structure and implicit communication. In *Proceedings of the 7th International Pragmatics Conference*, Budapest, Hungary, July 2000a.

W. C. Mann and S. A. Thompson. Two views on Rhetorical Structure Theory. In *Proceedings of the 10th Annual Meeting of the Society for Text and Discourse*, Lyon, France, July 2000b.

W. C. Mann and S. A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8:243–281, 1988.

D. Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, Nov. 2000. ISBN 978-0-262-13372-2.

D. Marcu. From discourse structures to text summaries. In *The Proceedings of the ACL/EACL workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain, July 1997a.

D. Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD thesis, University of Toronto, Dec. 1997b.

D. Marcu. To build text summaries of high quality, nuclearity is not sufficient. In *The Working Notes of the the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pages 1–8, Stanford, CA, USA, March 1998. American Association for Artificial Intelligence.

D. Marcu. Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury, editors, *Advances in Automatic Text Summarization*, pages 123–136. MIT Press, 1999.

D. Marcu and A. Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 368–375, Morristown, NJ, USA, July 2002.

E. E. Marsh and M. D. White. A taxonomy of relationships between images and text. *Journal of Documentation*, 59(6):647–672, Dec. 2003.

E. C. Marsi. Optionality in evaluating prosody prediction. In *Proceedings of the 5th ISCA ITRW on Speech Synthesis*, pages 13–18, June 2004.

E. C. Marsi and E. J. Krahmer. Explorations in sentence fusion. In *Proceedings of the 10th European workshop on Natural Language Generation*, Aberdeen, UK, Aug. 2005.

E. C. Marsi, E. J. Krahmer, W. E. Bosma, and M. Theune. Normalized alignment of dependency trees for detecting textual entailment. In B. Magnini and I. Dagan, editors, *Second PASCAL Recognising Textual Entailment Challenge*, pages 56–61, Venice, Italy, Apr. 2006. PASCAL.

E. C. Marsi, E. J. Krahmer, and W. E. Bosma. Dependency-based paraphrasing for recognizing textual entailment. In *Proceedings of the ACL/PASCAL workshop on Textual Entailment and Paraphrasing*, pages 83–88, June 2007.

R. Martinec and A. Salway. A system for image-text relations in new (and old) media. *Visual communication*, 4(3):339–374, 2005.

M. T. Maybury and A. E. Merlino. Multimedia summaries of broadcast news. In *Proceedings of the IASTED International Conference on Intelligent Information Systems*. IEEE, 1997.

K. McKeown, R. J. Passonneau, D. K. Elson, A. Nenkova, and J. Hirschberg. Do summaries help? In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and development in information retrieval*, pages 210–217, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-034-5.

S. Miike, E. Itoh, K. Ono, and K. Sumita. A full-text retrieval system with a dynamic abstract generation function. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 152–161, New York, NY, USA, 1994. Springer. ISBN 0-387-19889-X.

G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4): 235–244, 1990.

J.-L. Minel, S. Nugier, and G. Piat. How to appreciate the quality of automatic text summarization? In I. Mani and M. Maybury, editors, *EACL workshop Intelligent Scalable Text Summarization*, pages 25–31, July 1997.

M. Mitra, A. Singhal, and C. Buckley. Automatic text summarization by paragraph extraction. In *Proceedings of the workshop on Intelligent Scalable Text Summarization*, pages 39–46, Aug. 1997.

C. Monz and M. de Rijke. Light-weight entailment checking for computational semantics. In P. Blackburn and M. Kohlhase, editors, *Proceedings Inference in Computational Semantics*, pages 59–72, 2001.

T. Mori, M. Nozawa, and Y. Asada. Multi-document summarization using a question-answering engine. In *Proceedings of the 4th NTCIR workshop*, Tokyo, Japan, June 2004.

A. H. Morris, G. M. Kasper, and D. A. Adams. The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research*, 3(1):17–35, Mar. 1992.

J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991.

T. Mullen and R. Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In *Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs*, pages 159–162, 2006.

K. Nagao, S. Ohira, and M. Yoneoka. Annotation-based multimedia summarization and translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 1–7, Morristown, NJ, USA, 2002.

A. Nenkova and R. Passonneau. Evaluating content selection in summarization: the Pyramid method. In *Proceedings of the Conference on HLT/NAACL*, 2004.

NISO. Guidelines for abstracts. Technical Report 239.14-1997, ANSI/NISO, Bethesda, MD, USA, 1997.

J. Nivre and M. Scholz. Deterministic dependency parsing of English text. In *Proceedings of COLING 2004*, pages 23–27, Geneva, Switzerland, 2004.

E. W. Noreen. *Computer intensive methods for testing hypotheses: an introduction*. Wiley, New York, NY, USA, 1989. ISBN 978-0-471-61136-3.

D. P. O'Leary, M. E. Okurowski, A. Taylor, W. Wong, L. Carlson, J. M. Conroy, and D. Marcu. An empirical study of the relation between abstracts, extracts, and the discourse structure of texts. In *Proceedings of the Document Understanding Conference*, Sept. 2001.

J. Otterbacher, G. Erkan, Erkan, and D. R. Radev. Using random walks for question-focused sentence retrieval. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 915–922, Morristown, NJ, USA, 2005.

H. den Ouden. *Prosodic realizations of text structure*. PhD thesis, University of Tilburg, Dec. 2004.

P. Over and J. Yen. An introduction to DUC 2003: Intrinsic evaluation of generic news text summarization systems. In *Proceedings of the Document Understanding Conference*, 2003.

C. D. Paice. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In *Proceedings of the 3rd annual ACM Conference on Research and Development in Information Retrieval*, pages 172–191, Kent, United Kingdom, 1981. Butterworth & Co. ISBN 0-408-10775-8.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 39th Annual Meeting of the ACL*, pages 311–318, Morristown, NJ, USA, July 2001.

S. Park, J.-H. Lee, C.-M. Ahn, J. S. Hong, and S.-J. Chun. Query based summarization using non-negative matrix factorization. In *Knowledge-Based Intelligent Information and Engineering Systems*, volume 4253, pages 84–89. Springer, Berlin, Germany, 2006. ISBN 978-3-540-46542-3.

R. J. Passonneau. Evaluating an evaluation method: the Pyramid method applied to 2003 Document Understanding Conference (DUC) data. Technical report CUCS-010-06, Columbia University, 2005.

K. Pastra and H. Saggion. Colouring summaries BLEU. In *Proceedings of EACL 2003*, 2003.

A. Peñas, Á. Rodrigo, V. Sama, and F. Verdejo. Overview of the answer validation exercise 2006. In C. Peters, P. Clough, F. Gey, J. Karlgren, B. Magnini, D. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multimodal Information Retrieval*, volume 4730 of *Lecture Notes in Computer Science*, pages 257–264, Berlin, Germany, 2007. Springer. ISBN 978-3-540-74998-1.

V. A. Petrushin. *Introduction into Multimedia Data Mining and Knowledge Discovery*, pages 3–13. Springer London, 2007. ISBN 978-1-84628-799-2.

M. Pinto-Molina. Documentary abstracting: toward a methodological model. *Journal of the American Society for Information Science*, 46(3):225–234, 1995.

L. Polanyi. A formal model of the structure of discourse in cognitive aspects of language use. *Journal of Pragmatics*, 12(5–6):601–638, 1988.

J. J. Pollock and A. Zamora. Automatic abstracting research at chemical abstracts service. *Journal of Chemical Information and Computer Sciences*, 15(4):226–232, 1975.

M. F. Porter. Snowball: A language for stemming algorithms, 2001. http://snowball.tartarus.org/texts/introduction.html.

M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

D. R. Radev. A common theory of information fusion from multiple text sources step one: cross-document structure. In *Proceedings of the first SIGdial workshop on Discourse and Dialogue*, volume 10, pages 74–83, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

D. R. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the NAACL/ANLP workshop on Automatic Summarization*, pages 21–30, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

D. R. Radev, E. H. Hovy, and K. McKeown. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408, Dec. 2002a.

D. R. Radev, S. Teufel, H. Saggion, W. Lam, J. Blitzer, A. Çelebi, H. Qi, E. Drabek, and D. Liu. Evaluation of text summarization in a cross-lingual information retrieval framework. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA, 2002b.

D. R. Radev, S. Teufel, H. Saggion, W. Lam, J. Blitzer, H. Qi, A. Çelebi, D. Liu, and E. Drabek. Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 375–382, July 2003.

D. R. Radev, H. Jing, M. Styś, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938, 2004.

G. J. Rath, A. Resnick, and T. R. Savage. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2):139–143, Apr. 1961.

L. F. Rau and P. S. Jacobs. Creating segmented databases from free text for text retrieval. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 337–346, Chicago, Illinois, USA, 1991. ACM Press. ISBN 0-89791-448-1.

G. Redeker. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics*, 14(3):367–381, 1990.

Á. Rodrigo, A. Peñas, and F. Verdejo. The effect of entity recognition in answer validation. In C. Peters, P. Clough, F. Gey, J. Karlgren, B. Magnini, D. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval – 7th CLEF workshop*, Alicante, Spain, Sept. 2006.

J. E. Rowley. *Abstracting and Indexing*. Clive Bingley, 2nd edition, 1988. ISBN 978-0-85157-411-0.

D. E. Rumelhart. *Notes on a schema for stories*, pages 211–236. Academic Press, New York, NY, USA, 1975. ISBN 0-12-108550-3.

S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ, USA, 1995. ISBN 978-0-13-103805-9.

H. Saggion and G. Lapalme. Concept identification and presentation in the context of technical text summarization. In *Proceedings of Automatic Summarization workshop*, pages 1–10, 2000.

H. Saggion, D. R. Radev, S. Teufel, and W. Lam. Meta-evaluation of summaries in a cross-lingual environment using content-based metrics. In *Proceedings of the 40th Annual Meeting of the ACL*, volume 1, pages 1–7, Morristown, NJ, USA, Aug. 2002. Association for Computational Linguistics.

T. Sakai and K. Spärck Jones. Generic summaries for indexing in information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 190–198, New York, NY, USA, 2001. ACM Press. ISBN 1-58113-331-6.

G. Salton. *Automatic text processing*. Addison-Wesley Longman, Boston, MA, USA, 1988. ISBN 0-211-2278.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

G. Salton, A. Singhal, M. Mitra, and C. Buckley. Automatic text structuring and summarization. *Information Processing and Management*, 33(2):193–207, Mar. 1997.

T. Sanders and C. van Wijk. PISA: A procedure for analyzing the structure of explanatory texts. *Text*, 16(1):91–132, 1996.

H. Schauer and U. Hahn. Anaphoric cues for coherence relations. In G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, and N. Nikolov, editors, *Proceedings of the Conference on Recent Advances in Natural Language Processing*, pages 228–234, 2001.

S. E. Schwarm and M. Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 523–530, Morristown, NJ, USA, 2005.

E. F. Skorochod'ko. *Semantische Relationen in der Lexik und in Texten*, volume 10 of *Quantitative Linguistics*. Studienverlag Brockmeyer, Bochum, Germany, 1981. ISBN 3-88339171-9. German translation, original publication date ˜1971.

K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

K. Spärck Jones and J. R. Galliers. Evaluating natural language processing systems: an analysis and review. Springer, 1996.

M. Stede and S. Heintze. Machine-assisted rhetorical structure annotation. In *Proceedings of the 42nd Annual Meeting of the ACL*, Morristown, NJ, USA, 2004.

T. Strzalkowski, R. Gaizauskas, E. M. Voorhees, S. Harabagiu, R. Weishedel, D. Israel, C. Jacquemin, C.-Y. Lin, S. Maiorano, G. Miller, D. Moldovan, B. Ogden, J. Prager, E. Riloff, J. Burger, A. Singhal, C. Cardie, R. Shrihari, and V. Chaudhri. Issues, tasks, and program structures to roadmap research in question & answering (Q&A). NIST DUC Vision and Roadmap Documents, Oct. 2000.

R. Swan and J. Allan. Automatic generation of overview timelines. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–56, Athens, Greece, 2000. ACM Press. ISBN 1-58113-226-3.

E. Teich and P. Fankhauser. WordNet for lexical cohesion analysis. In P. Sojka, K. Pala, P. Smrz, C. Fellbaum, and P. Vossen, editors, *Proceedings of the Second International WordNet Conference*, Brno, Czech Republic, 2004. Masaryk University. ISBN 80-210-3302-9.

S. Teufel and H. van Halteren. Evaluating information content by factoid analysis: Human annotation and stability. In *Proceedings of Empirical Methods in Natural Language Processing*, 2004.

M. Theune, E. Klabbers, J. Odijk, J. R. de Pijper, and E. J. Krahmer. From data to speech: a general approach. *Natural Language Engineering*, 7(1):47–86, 2001.

M. Theune, F. Hielkema, and P. Hendriks. Performing aggregation and ellipsis using discourse structures. *Research on Language and Computation*, 4(4):353–375, 2006.

M. Theune, B. W. van Schooten, H. J. A. op den Akker, W. E. Bosma, D. H. Hofs, A. Nijholt, E. J. Krahmer, C. M. J. van Hooijdonk, and E. C. Marsi. Questions, pictures, answers: introducing pictures in question-answering systems. In L. R. Miyarez, A. M. Alvarado, and C. A. Moreno, editors, *ACTAS-1 of X Symposio Internacional de Comunicacion Social*, pages 450–463, Santiago de Cuba, 2007. Centro de Linguistica Aplicada. ISBN 959-7174-08-1.

E. Tjong Kim Sang, G. Bouma, and M. de Rijke. Developing offline strategies for answering medical questions. In *Proceedings of the AAAI workshop on Question Answering in Restricted Domains*, pages 41–45, 2005.

A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–10, New York, NY, USA, 1998. ACM Press. ISBN 1-58113-015-5.

R. H. Trigg and M. Weiser. Textnet: a network-based approach to text handling. *ACM Transactions on Information Systems*, 4(1):1–23, 1986.

K. van Deemter and R. Power. High-level authoring of illustrated documents. *Natural Language Engineering*, 9:101–126, 2003.

L. Vanderwende and W. B. Dolan. What syntax can contribute in the entailment task. In *PASCAL Challenges workshop on Recognizing Textual Entailment*, pages 205–216, Southampton, United Kingdom, 2005. Springer.

J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, D. Ferro, S. L. Oviatt, P. R. Cohen, and L. Wu. Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions for 2000 and beyond, 2000.

E. M. Voorhees. The TREC question answering track. *Natural Language Engineering*, 7(4):361–378, 2001.

E. M. Voorhees. Overview of the TREC 2002 question answering track. In *Proceedings of the 11th Text Retrieval Conference*, 2002.

E. M. Voorhees. Overview of the TREC 2003 question answering track. In *Proceedings of the 12th Text Retrieval Conference*, 2003.

E. M. Voorhees and D. M. Tice. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and development in information retrieval*, pages 200–207, Athens, Greece, July 2000. ACM Press.

C. van Wijk and T. J. Sanders. Identifying writing strategies through text analysis. *Written communication: a quarterly journal of research, theory and application*, 16:51–75, Jan. 1999.

M. J. Witbrock and V. O. Mittal. Ultra-summarization: a statistical approach to generating highly condensed non-extractive summaries. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and development in information retrieval*, pages 315–316, New York, NY, USA, 1999. ACM Press. ISBN 1-58113-096-1.

R. Witte and S. Bergler. Fuzzy coreference resolution for summarization. In *Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization*, pages 43–50, Venice, Italy, 2003. Università Ca' Foscari.

R. Witte, R. Krestel, and S. Bergler. Generating update summaries for DUC 2007. In *Proceedings of the Document Understanding Conference*, 2007.

F. Wolf and E. Gibson. Paragraph-, word-, and coherence-based approaches to sentence ranking: A comparison of algorithm and human performance. In *Proceedings of the 42nd Annual Meeting of the ACL*, Morristown, NJ, USA, July 2004.

F. Wolf and E. Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–288, 2005.

J.-Y. Yeh, H.-R. Ke, W.-P. Yang, and I.-H. Meng. Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing and Management*, 41(1):75–95, 2005.

D. Zajic, B. J. Dorr, and R. Schwartz. Automatic headline generation for newspaper stories. In *Proceedings of the Document Understanding Conference*, pages 78–85, Philadelphia, PA, USA, 2002.

K. Zechner. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485, 2002.

Z. Zhang, S. Blair-Goldensohn, and D. R. Radev. Towards CST-enhanced summarization. In *Proceedings of the AAAI Conference*, 2002.

Z. Zhang, J. Otterbacher, and D. R. Radev. Learning cross-document structural relationships using boosting. In *Proceedings of the 12th International Conference on Information and Knowledge Management*, pages 124–130, New York, NY, USA, 2003. ACM Press. ISBN 1-58113-723-0.

L. Zhou, C.-Y. Lin, D. S. Munteanu, and E. H. Hovy. ParaEval: using paraphrases to evaluate summaries automatically. In *Proceedings of the Conference on HLT/NAACL*, pages 447–454, Morristown, NJ, USA, 2006.

L. Zhou, N. Kwon, and E. H. Hovy. A semi-automatic evaluation scheme: Automated nuggetization for manual annotation. In *Proceedings of the Conference on HLT/NAACL, companion volume, short papers*, pages 217–220, Apr. 2007.

Abstract

The meaning of text appears to be tightly related to intentions and circumstances. Context sensitivity of meaning is addressed by theories of discourse structure. Few attempts have been made to exploit text organization in summarization. This thesis is an exploration of what knowledge of discourse structure can do for content selection as a subtask of automatic summarization, and query-based summarization in particular. Query-based summarization is the task of answering an arbitrary user query or question by using content from potentially relevant sources.

This thesis presents a general framework for discourse oriented summarization, relying on graphs to represent *semantic relations* in discourse, and *redundancy* as a special type of semantic relation. Semantic relations occur on several levels of text analysis (query-relevance, coherence, layout, etc.), and a broad range of textual features may be required to detect them. The graph-based framework facilitates combining multiple features into an integrated semantic model of the documents to summarize. Recognizing redundancy and entailment relations between text passages is particularly important when a summary is generated of multiple documents, e.g. to avoid including redundant content in a summary. For this reason, I pay particular attention to recognizing textual entailment.

Within this framework, a three-fold evaluation is performed to evaluate different aspects of discourse oriented summarization. The first is a user study, measuring the effect on user appreciation of using a particular type of knowledge for query-based summarization. In this study, three presentation strategies are compared: summarization using the rhetorical structure of the source, a baseline summarization method which uses the layout of the source, and a baseline presentation method which uses no summarization but just a concise answer to the query. Results show that knowledge of the rhetorical structure not only helps to provide the necessary context for the user to verify that the summary addresses the query adequately, but also to increase the amount of relevant content.

The second evaluation is a comparison of implementations of the graph-based framework which are capable of fully automatic summarization. The two variables in the experiment are the set of textual features used to model the source and the algorithm used to search a graph for relevant content. The features are based on cosine similarity, and are realized as graph representations of the source. The graph search

algorithms are inspired by existing algorithms in summarization. The quality of summaries is measured using the Rouge evaluation toolkit. The best performer would have ranked first (Rouge-2) or second (Rouge-SU4) if it had participated in the DUC 2005 query-based summarization challenge.

The third study is an evaluation in the context of the DUC 2006 summarization challenge, which includes readability measurements as well as various content-based evaluation metrics. The evaluated automatic discourse oriented summarization system is similar to the one described above, but uses additional features, i.e. layout and textual entailment. The system performed well on readability at the cost of content-based scores which were well below the scores of the highest ranking DUC 2006 participant. This indicates a trade-off between readable, coherent content and useful content, an issue yet to be explored.

Previous research implies that theories of text organization generalize well to multimedia. This suggests that the discourse oriented summarization framework applies to summarizing multimedia as well, provided sufficient knowledge of the organization of the (multimedia) source documents is available. The last study in this thesis is an investigation of the applicability of structural relations in multimedia for generating picture-illustrated summaries, by relating summary content to picture-associated text (i.e. captions or surrounding paragraphs). Results suggest that captions are the more suitable annotation for selecting appropriate pictures. Compared to manual illustration, results of automatic pictures are similar if the manual picture is mainly decorative.

Samenvatting

De betekenis van een stuk tekst hangt nauw samen met omstandigheden en de onderliggende bedoelingen. Deze contextgevoeligheid is beschreven in modellen van tekststructuur. Hiervan wordt echter nauwelijks gebruik gemaakt voor het automatisch samenvatten van tekst. Dit proefschrift tracht een antwoord te geven op de vraag hoe kennis van de structuur van een tekst benut kan worden voor het selecteren van inhoudelijk relevante passages, als onderdeel van het automatisch genereren van samenvattingen. In het bijzonder is het gericht op de generatie van een samenvatting van een of meerdere brondocumenten die dient als antwoord op een gebruikersvraag – een toepassing van natuurlijke taalverwerking die de laatste jaren toenemende aandacht geniet.

Eerst behandel ik theorieën die de structuur van tekst vanuit verschillende invalshoeken benaderen. Vanuit dit oogpunt presenteer ik een raamwerk voor een systeem dat samenvattingen genereert, uitgaand van een graafrepresentatie van de brontekst. Grafen worden gebruikt om semantische relaties tussen zinnen weer te geven, waarbij redundantie wordt aangemerkt als een bijzonder type semantische relatie. Verschillende eigenschappen van tekst kunnen worden gebruikt als indicatie dat bepaalde zinnen gerelateerd zijn. Voorbeelden hiervan zijn de alinea-indeling, lexicale overlap, retorische relaties, enz. Het raamwerk maakt het mogelijk alle beschikbare indicaties te benutten door de verschillende grafen die hieraan ten grondslag liggen samen te voegen in een enkel model van de brondocumenten. Het herkennen van afleidbaarheidsrelaties tussen tekstfragmenten is van belang wanneer meerdere bronteksten worden omgezet in een enkele samenvatting, ondermeer om redundantie in de samenvatting te voorkomen. In dit licht onderzoek ik welke tekstuele eigenschappen geschikt zijn om automatisch te herkennen of een tekstfragment inhoudelijk afleidbaar is uit een ander fragment.

Binnen dit raamwerk voer ik drie evaluaties uit om verschillende aspecten van het samenvatten te belichten. In de eerste evaluatie wordt potentiële gebruikers gevraagd op verschillende wijze samengestelde samenvattingen te waarderen. De resultaten laten zien dat samenvattingen de gebruiker helpen te bepalen of het antwoord bij de vraag past. Verder bevatten op de retorische analyses gebaseerde samenvattingen meer relevante informatie dan de op layout gebaseerde samenvattingen. De tweede evaluatie is een vergelijkende studie van algoritmen om in een graafrepresentatie van een tekst

te zoeken naar relevante inhoud teneinde daaruit een samenvatting samen te stellen. De kwaliteit van aldus gegenereerde samenvattingen wordt gemeten door middel van automatische evaluatiemethoden uit het Rouge softwarepakket. Het best presterende algoritme was als beste (Rouge-2) of als tweede (Rouge-SU4) geëindigd als het systeem was gebruikt voor deelname aan het DUC 2005, een evaluatieprogramma voor automatisch gegenereerde samenvattingen. Ten derde is er een meer diepgaande evaluatie van een implementatie van het raamwerk in de vorm van een deelname aan DUC 2006. Deze evaluatie omvat naast automatische evaluatiemethoden ook methoden die een menselijk oordeel vergen, waaronder verscheidene maten voor leesbaarheid. De resultaten laten zien dat het systeem goed presteert op leesbaarheid, maar ook dat er een ruime afstand is tot de best presterende deelnemer aan DUC 2006. De resultaten laten zien dat het voorgestelde raamwerk kan leiden tot samenvattingen van hoge kwaliteit, met betrekking tot zowel inhoud als leesbaarheid. Tegelijkertijd wijzen de resultaten erop dat mogelijk een afweging gemaakt moet worden tussen optimalisatie van inhoud en leesbaarheid.

Uit literatuur blijkt dat modellen die tekststructuur beschrijven vaak ook toepasbaar zijn op multimediale documenten. Dit suggereert dat het raamwerk voor het samenvatten van tekst in beginsel gebruikt moet kunnen worden om multimedia samen te vatten, mits voldoende bekend is over de interne structuur van de bron. Ik onderzoek of afbeeldingen uit multimedia-documenten in een samenvatting opgenomen kunnen worden op basis van een semantische relatie tussen de samenvatting en een tekstuele annotatie van de afbeelding. De tekstuele annotatie van de afbeelding is automatisch samengesteld uit het document waaruit de afbeelding afkomstig is (het onderschrift of de alinea waar de afbeelding oorspronkelijk bij hoorde). Resultaten van een gebruikersstudie geven aan dat het gebruik van onderschriften betere resultaten oplevert dan het gebruik van de bijbehorende tekst.

SIKS dissertation series

This is the 174th publication in the SIKS dissertation series. Below is an overview of the most recent previous SIKS dissertations.

2008-09 Christof van Nimwegen, UU. *The paradox of the guided user: assistance can be counter-effective.*

2008-08 Janneke Bolt, UU. *Bayesian Networks: Aspects of Approximate Inference.*

2008-07 Peter van Rosmalen, OU. *Supporting the tutor in the design and support of adaptive e-learning.*

2008-06 Arjen Hommersom, RUN. *On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective.*

2008-05 Bela Mutschler, UT. *Modeling and simulating causal dependencies on process-aware information systems from a cost perspective.*

2008-04 Ander de Keijzer, UT. *Management of Uncertain Data – towards unattended integration.*

2008-03 Vera Hollink, UVA. *Optimizing hierarchical menus: a usage-based approach.*

2008-02 Alexei Sharpanskykh, VU. *On Computer-Aided Methods for Modeling and Analysis of Organizations.*

2008-01 Katalin Boer-Sorbán, EUR. *Agent-Based Simulation of Financial Markets: A modular, continuous-time approach.*

2007-25 Joost Schalken, VU. *Empirical Investigations in Software Process Improvement.*

2007-24 Georgina Ramírez Camps, CWI. *Structural Features in XML Retrieval.*

2007-23 Peter Barna, TUE. *Specification of Application Logic in Web Information Systems.*

2007-22 Zlatko Zlatev, UT. *Goal-oriented design of value and process models from patterns.*

2007-21 Karianne Vermaas, UU. *Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005.*

2007-20 Slinger Jansen, UU. *Customer Configuration Updating in a Software Supply Network.*

2007-19 David Levy, UM. *Intimate relationships with artificial partners.*

2007-18 Bart Orriens, UvT. *On the development an management of adaptive business collaborations.*

2007-17 Theodore Charitos, UU. *Reasoning with Dynamic Networks in Practice.*

2007-16 Davide Grossi, UU. *Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems.*

2007-15 Joyca Lacroix, UM. *NIM: a Situated Computational Memory Model.*

2007-14 Niek Bergboer, UM. *Context-Based Image Analysis.*

2007-13 Rutger Rienks, UT. *Meetings in Smart Environments; Implications of Progressing Technology.*

2007-12 Marcel van Gerven, RUN. *Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty.*

2007-11 Natalia Stash, TUE. *Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System.*

2007-10 Huib Aldewereld, UU. *Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols.*

2007-09 David Mobach, VU. *Agent-Based Mediated Service Negotiation.*

2007-08 Mark Hoogendoorn, VU. *Modeling of Change in Multi-Agent Organizations.*

2007-07 Nataša Jovanović, UT. *To Whom It May Concern – Addressee Identification in Face-to-Face Meetings.*

2007-06 Gilad Mishne, UVA. *Applied Text Analytics for Blogs.*

2007-05 Bart Schermer, UL. *Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance.*

2007-04 Jurriaan van Diggelen, UU. *Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach.*

2007-03 Peter Mika, VU. *Social Networks and the Semantic Web.*

2007-02 Wouter Teepe, RUG. *Reconciling Information Exchange and Confidentiality: A Formal Approach.*

2007-01 Kees Leune, UvT. *Access Control and Service-Oriented Architectures.*

2006-28 Borkur Sigurbjornsson, UVA. *Focused Information Access using XML Element Retrieval.*

2006-27 Stefano Bocconi, CWI. *Vox Populi: generating video documentaries from semantically annotated media repositories.*

2006-26 Vojkan Mihajlovic, UT. *Score Region Algebra: A Flexible Framework for Structured Information Retrieval.*

2006-25 Madalina Drugan, UU. *Conditional log-likelihood MDL and Evolutionary MCMC.*

2006-24 Laura Hollink, VU. *Semantic Annotation for Retrieval of Visual Resources.*

2006-23 Ion Juvina, UU. *Development of Cognitive Model for Navigating on the Web.*

2006-22 Paul de Vrieze, RUN. *Fundaments of Adaptive Personalisation.*

2006-21 Bas van Gils, RUN. *Aptness on the Web.*

2006-20 Marina Velikova, UvT. *Monotone models for prediction in data mining.*

2006-19 Birna van Riemsdijk, UU. *Cognitive Agent Programming: A Semantic Approach.*

2006-18 Valentin Zhizhkun, UVA. *Graph transformation for Natural Language Processing.*

2006-17 Stacey Nagata, UU. *User Assistance for Multitasking with Interruptions on a Mobile Device.*

2006-16 Carsten Riggelsen, UU. *Approximation Methods for Efficient Learning of Bayesian Networks.*

2006-15 Rainer Malik, UU. *CONAN: Text Mining in the Biomedical Domain.*

2006-14 Johan Hoorn, VU. *Software Requirements: Update, Upgrade, Redesign – towards a Theory of Requirements Change.*

2006-13 Henk-Jan Lebbink, UU. *Dialogue and Decision Games for Information Exchanging Agents.*

2006-12 Bert Bongers, VU. *Interactivation – Towards an e-cology of people, our technological environment, and the arts.*

2006-11 Joeri van Ruth, UT. *Flattening Queries over Nested Data Types.*

2006-10 Ronny Siebes, VU. *Semantic Routing in Peer-to-Peer Systems.*

2006-09 Mohamed Wahdan, UM. *Automatic Formulation of the Auditor's Opinion.*

2006-08 Eelco Herder, UT. *Forward, Back and Home Again – Analyzing User Behavior on the Web.*

2006-07 Marko Smiljanic, UT. *XML schema matching – balancing efficiency and effectiveness by means of clustering.*

2006-06 Ziv Baida, VU. *Software-aided Service Bundling – Intelligent Methods & Tools for Graphical Service Modeling.*

2006-05 Cees Pierik, UU. *Validation Techniques for Object-Oriented Proof Outlines.*

2006-04 Marta Sabou, VU. *Building Web Service Ontologies.*

2006-03 Noor Christoph, UVA. *The role of metacognitive skills in learning to solve problems.*

2006-02 Cristina Chisalita, VU. *Contextual issues in the design and use of information technology in organizations.*

2006-01 Samuil Angelov, TUE. *Foundations of B2B Electronic Contracting.*

2005-21 Wijnand Derks, UT. *Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics.*

2005-20 Cristina Coteanu, UL. *Cyber Consumer Law, State of the Art and Perspectives.*

2005-19 Michel van Dartel, UM. *Situated Representation.*

2005-18 Danielle Sent, UU. *Test-selection strategies for probabilistic networks.*

2005-17 Boris Shishkov, TUD. *Software Specification Based on Re-usable Business Components.*

2005-16 Joris Graaumans, UU. *Usability of XML Query Languages.*

2005-15 Tibor Bosse, VU. *Analysis of the Dynamics of Cognitive Processes.*

2005-14 Borys Omelayenko, VU. *Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics.*

2005-13 Fred Hamburg, UL. *Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen.*

2005-12 Csaba Boer, EUR. *Distributed Simulation in Industry.*

2005-11 Elth Ogston, VU. *Agent Based Matchmaking and Clustering – A Decentralized Approach to Search.*

2005-10 Anders Bouwer, UVA. *Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments.*

2005-09 Jeen Broekstra, VU. *Storage, Querying and Inferencing for Semantic Web Languages.*

2005-08 Richard Vdovjak, TUE. *A Model-driven Approach for Building Distributed Ontology-based Web Applications.*

2005-07 Flavius Frasincar, TUE. *Hypermedia Presentation Generation for Semantic Web Information Systems.*

2005-06 Pieter Spronck, UM. *Adaptive Game AI.*

2005-05 Gabriel Infante-Lopez, UVA. *Two-Level Probabilistic Grammars for Natural Language Parsing.*

2005-04 Nirvana Meratnia, UT. *Towards Database Support for Moving Object data.*

2005-03 Franc Grootjen, RUN. *A Pragmatic Approach to the Conceptualisation of Language.*

2005-02 Erik van der Werf), UM. *AI techniques for the game of Go.*

2005-01 Floor Verdenius, UVA. *Methodological Aspects of Designing Induction-Based Applications.*