# TWLT 14

# Language Technology in Multimedia Information Retrieval

Proceedings of the fourteenth
Twente Workshop on Language Technology

December 7-8, 1998
Enschede, The Netherlands

**Djoerd Hiemstra, Franciska de Jong**
**and Klaus Netter (eds.)**

CIP GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Hiemstra, D., De Jong, F.M.G., Netter, K.

*Language Technology in Multimedia Information Retrieval*
Proceedings Twente Workshop on Language Technology 14
D. Hiemstra, F.M.G. de Jong, K. Netter (eds.)
Enschede, Universiteit Twente, Faculteit Informatica

ISSN 0929-0672

trefwoorden: information retrieval, natural language processing, speech, multilinguality, language engineering, multimedia information disclosure



Book orders:
Ms. A. Hoogvliet
University of Twente
Dept. of Computer Science
P.O. Box 217
NL 7500 AE Enschede
tel: +31 53 4893680
fax: +31 53 4893503
Email: hoogvlie@cs.utwente.nl

Druk- en bindwerk: Reprografie U.T. Service Centrum, Enschede

# Preface

TWLT is an acronym of Twente Workshop(s) on Language Technology. These workshops on natural language theory and technology are organised by the Parlevink Project, a language theory and technology project of the Department of Computer Science of the University of Twente, Enschede, The Netherlands. For each workshop proceedings are published containing the papers that were presented.

TWLT 14, has been organised together with the German Research Center for Artificial Intelligence, DFKI Saarbrücken, Germany. The idea for this workshop grew out of a longstanding cooperation between the University of Twente, TNO-TPD in Delft and DFKI. This co-operation manifested itself for the first time in the Twenty-One project, which inspired a whole series of other projects, such as Pop-Eye and Olive, but which also led to a close contact and exchange with independently established projects such as Mulinex and MIETTA for which DFKI was responsible. All of these projects had in common that they were funded by the Telematics Application Programme of the European Commission, all, except for Twenty-One, by the Language Engineering Sector.

Beyond this formal feature, the projects mentioned also had in common that part of their agenda was and is to explore and prove the use and usefulness of language technology in the area of information retrieval. The application domains of these projects are quite different. Twenty-One concentrated on the environmental domain of sustainable development, Pop-Eye and Olive are situated in the area of video retrieval on the basis of linguistic material, Mulinex is a rather general purpose search engine, and MIETTA belongs to the domain of travel and tourism. What distinguishes these domains is clearly the type of media involved: Twenty-One's document base consists mostly of printed papers which had be scanned in and processed. Pop-Eye and Olive deal almost exclusively with video material, where Olive brings in the additional feature of speech processing. Mulinex is concerned with retrieval in the world wide web, and MIETTA combines the search in structured databases with searching the web for all kinds of information. However, there is again one common feature which plays quite a crucial role in all of these projects viz. multi-linguality. As the individual projects all take somewhat different tacks on this issue, one could almost claim that between them they cover the whole range of possible approaches to multi-lingual information retrieval.

Our objective when organising this workshop was to bring together researchers from these projects at a more formal conference to present their different approaches. However we also wanted to initiate a more lively exchange with other prominent scientists in the field. The work they are representing here shows some intriguing overlap with our own work, but in addition some contributions address the issues we feel should play a more prominent role in the next generation of projects that given the success of the previous co-operation we are likely to start in the future. In particular we foresee that the need for more personalised information delivery will urge for advanced tools such as information extraction and filtering, concept-based retrieval, domain modeling, image processing, etc. to be integrated in what we established thusfar. With funding from the Telematica Instituut (initiative of the Dutch government and a number of industrial organisations) TNO and the University of Twente recently started the DRUID project for the exploration of some of these themes and for setting up new collaborative links. We expect that TWLT14 will contribute to the aims of this initiative, and more generally to a better understanding of the potential and needs of this particular research area.

A workshop is the concerted action of many people. We are grateful to the authors and the organisations they represent, for their efforts and contributions. In addition we would like to mention here the people whose efforts have been less visible during the workshop proper, but whose contribution was evidently of crucial importance. Alice Hoogvliet took care of the administrative tasks (registration, hotel reservations, etc.), Thijs Westerveld supported us in disseminating the workshop announcements and Michiel Visser helped with the lay-out of the proceedings.

Djoerd Hiemstra, Franciska de Jong and Klaus Netter December, 1998

## Previous TWLT workshops

Previous TWLT workshops were

| | |
|---|---|
| TWLT1, | *Tomita's Algorithm: Extensions and Applications*. 22 March, 1991. |
| TWLT2, | *Linguistic Engineering: Tools and Products*. 20 November, 1991. |
| TWLT3, | *Connectionism and Natural Language Processing*. 12 and 13 May 1992. |
| TWLT4, | *Pragmatics in Language Technology*. 23 September, 1992. |
| TWLT5, | *Natural Language Interfaces*. 3 and 4 June, 1993. |
| TWLT6, | *Natural Language Parsing*. 16 and 17 December, 1993. |
| TWLT7, | *Computer-Assisted Language Learning*. 16 and 17 June 1994. |
| TWLT8, | *Speech and Language Engineering*. 1 and 2 December 1994. |
| TWLT9, | *Corpus-based Approaches to Dialogue Modelling*. 9 June, 1995 |
| TWLT10, | *Algebraic Methods in Language Processing*. 6-8 December, 1995 |
| TWLT11, | *Dialogue Management in Natural Language Systems*. 19-21 June, 1996 |
| TWLT12, | *Automatic Interpretation and Generation of Verbal Humor*. 11-14 Sept. 1996 |
| TWLT13, | *Formal Semantics and Pragmatics of Dialogue, Twendial'98*. 13-15 May 1998 |

For the contents of the previous proceedings, please consult the last pages of this volume.

# Contents

## Papers

## Posters and demonstrations

# Sponsors and support

We gratefully acknowledge help from:

- University of Twente
- Telematics Institute
- Nederlandse Spoorwegen
- Comsys Call Center Automation





# Organisation

TWLT 14 is organised by:

- Research Project Parlevink
- DFKI GmbH Saarbrücken
- Centre for Telematics and  Information Technology

# Cross-Language Information Retrieval:
## From Naive Concepts to Realistic Applications

Hans Uszkoreit
DFKI Language Technology Lab
& Saarland University
Stuhlsatzenhausweg 3,
66123 Saarbrücken, Germany
uszkoreit@dfki.de

ABSTRACT

In this paper I combine an overview of the goals and major approaches in cross-language information retrieval with some observations of current trends and with a report on a CLIR project that differs in many respects from most research activities in the fast growing area. In the overview, I will start from a generic model of an information retrieval system. Then the necessary extensions will be introduced that are needed for allowing queries in a language different from the document language. Several options for adding translation technology will be contrasted. I will then report on the research strategy followed in the EU-funded international project Mulinex. In this project a complete modular CLIR system was developed and integrated as the core software for a number of applications and as a plattform for research and technology development.

**Keywords:** Cross Language Information Retrieval, Translingual Information Management, Machine Translation, Query Translation, Query Expansion

## 1 INTRODUCTION

Cross language information retrieval (CLIR) is a language technology that extends the well established technology of information retrieval (IR) by adding facilities for the search of documents in a language different from the language of the query. As IR is a complex technology made up of several component technologies, so is CLIR. The added functionality of CLIR requires additional component technologies for the mapping between languages.

As with many new technologies, it is difficult to trace the origins of the underlying idea. Nifty ideas often start out from a dream that seems totally unrealistic. Then some innocent experiments demonstrate what would happen if one tried to achieve the goal by simply combining readily available components. Usually these experiments help to understand which technologies are still missing. The next step is less innocent experiments trying to show how close one could come to the desired functionality by putting together carefully selected technology in the best possible way.

Such first serious attempts often disclose that there is more than one way to accomplish the task. They may even suggest additional or alternative functionalities. Now the race is on. Different researchers explore the options by trying to fill the recognized gaps. During the course of competitive research more options are discovered. Specialized research on the open problems will continue until all preconditions for a commercially viable application are met.

However, the course of events is quite different if the dream is very powerful, if the need is so urgent that application building will not wait for the research to come to a successful end. In such a

situation, application building will start before the conditions are fulfilled. The result may be an imperfect application that tests and prepares the market for better things to come. Such an application will boost research. It will yield valuable insights for the researchers and may even provide them with additional tools and reference measures for technology development.

Yet the premature exploitation may as well have the adverse effect. A lousy application can temporarily ruin the market and harm the reputation of a good idea.

In this paper, I will try to sketch out this scenario for CLIR. Starting with the dream, I will first pay tribute to the first reported experiments. Then I will concentrate on the wide variety of alternative approaches to achieve the original goal. New functionalities will be discussed that were discovered in different research contexts. Some shortcomings of available component technologies will be pointed out.I will then try to structure the space of proposed and possible architectures and relate them to the range of proposed applications. The result will be a condensed map of approaches without claiming completenes.

Next I will argue that we have a reached the point where several applications are being built before the desired functionality can fully be achieved. Some of these "premature" systems can indeed be viewed as realistic applications for restricted purposes. However, the strong market pull and ambitious industrial announcements suggest that the first commercial general purpose CLIR systems will appear soon regardless of badly needed breakthroughs in technology development.

The CLIR system of the project Mulinex is a special case of a "premature" application. It can serve as a basis for both realistic and unrealistic applications. Its modular setup makes it an ideal platform for experimentation. The paper concludes with an outlook on future opportunities for CLIR systems, forecasting them a central place in knowledge management.

## 2 THE DREAM AND FIRST EXPERIMENTS

The dream behind CLIR technology is easily told. It is the same dream that stands behind all efforts to make machines translate and interprete. It is the dream to learn and communicate without being hindered by language barriers. The dream surfaces in modern fairy tales: C3PO, the babel fish and the universal translator on USS Enterprise reflect the ancient desire.

However, in the application domain of CLIR, this dream is combined with another powerful vision which has become the promise of information society: virtually unlimited access to information. For members of the IR community, MT is a useful extension to their technology. For their colleagues in MT, IR is an efficient facility for accessing the texts to be translated.

Whichever way we look at it, CLIR does not only bring together two powerful visions, it also combines two unsolved problems.

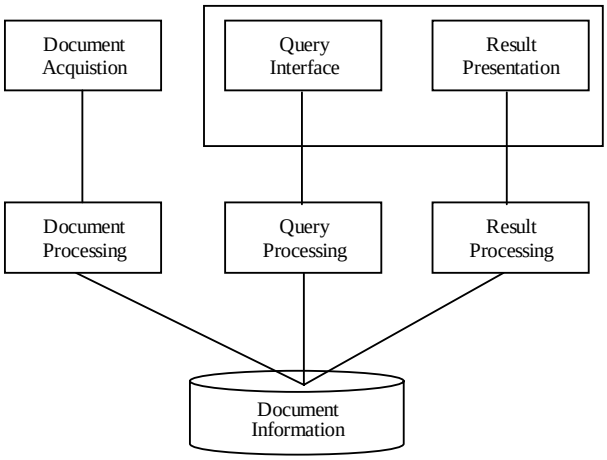
We cannot safely trace back the idea of retrieving documents accross language barriers, but we should point out that the first experiments with the CLIR approach were performed as early as 1969, when Gerard Salton used a thesaurus with hand-translated terms to search document-bases in two languages, English and German. Salton later repeated this experiment with a small base of 52 documents, this time using a hand-translated thesaurus of English and French terms [1]. On the basis of these small experiments and the achieved retrieval results, Salton concluded that the task was feasible in general. With his carefully handcrafted dictionaries, he could get results very close to monolingual retrieval. In order to appreciate these results, we should remember that the experiments were conducted at a time when nobody thought of the WorldWideWeb and an emerging information society.

I will spare the reader with the intermediate steps in the history of CLIR and refer the interested reader to Doug Oards highly informative survey paper [2] and to Fluhr [3].

Since most of the serious experiments and technology develoments only happened in the last few years and since a number of the relevant results and systems will be presented in this volume, I will immediately proceed to the overview of approaches that have been followed and refer to specific systems only in order to exemplify directions taken.

## 3 MAJOR APPROACHES

On an abstract level, the task of CLIR is not much different from traditional IR. A set of documents is acquired that are of potential interest to some user. The documents are processed in order to obtain information that may




be relevant in a query situation. We may call this information simply document information. The document information is stored in a format that permits the matching of queries against this information. Addresses of the documents are stored together with the document information to enable the actual retrieval. The addresses may be URLs, locations on a storage medium or data base identifiers.

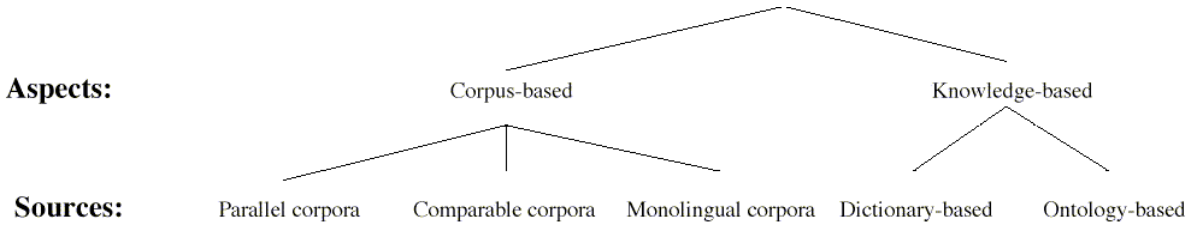
In a search situation a query is issued containing query items: words, phrases, category terms, documents or boolean expressions made up of such items. The query is processed in order to obtain the relevant query information to be matched against the document information. Query processing may add information to the query information that is not part of the issued query items such as preselected preferences, user profile information, or query history information.

The matching selects a subset of the documents by determining the degree of match, the relevance of the documents with respect to the query. The document information of this subset is passed to result processing which will present the all or parts of this information (or the documents themselves) to the user as requested.

Whereas in the abstract IR modell no difference is made between the documents and the document information, this difference can be important when we extend the model to CLIR.

The central part of the document information is a representation of the text parts of the documents. This information is usually stored as an index of words, lemmas, multiword terms, phrases, or concepts in some interlingua representation. Many IR systems also store summaries that are extracted from the documents during document processing. If the system employs an additional categorizer, the category information also becomes part of the document information. CLIR systems that acquire documents in several languages need to identify the language of each document. The language tag will be saved together with other metainformation such as date of creation.

In a CLIR system, the language of the acquired documents is different from the language of the query. If no mapping between the languages is provided, a query could only be successful if the query items happen to be also in the document language. Yet the query could only succeed if the items also have the same meaning in both languages.



However, there have been experiments with CLIR systems that do not translate at all [4]. Such systems exploit the closeness between the vocabularies of two languages. The matching between query information and document information needs to be relaxed or, better even, adapted to regular morphological and orthographic differences between corresponding words in the two languages. This type of matching is called cognate matching.

All other approaches use some kind of translation. However they differ in what they translate and how they translate. Let me start with the first distinction. Most systems are based on query translation. The translated query is matched against the documents. The major problem with this approach is the limited context in a search query. Although the advanced user will learn by frustrating experience that longer queries usually yield better results, the average query length on major search machines is still around two words. No automatic translation system can reliably select the right translation from a set of alternatives if not context is given. If the translation cannot make a decision on the basis of frequencies, query history, selected categories or users profiles, only the user can help. However this adds another interaction to the search process. We will return to this problem.

An alternative option is the automatic translation of the acquired texts. This approach offers several advantages. The translation engine has a better basis for disambiguation if the words are embedded in a larger context. The user will never be bothered with a selection request since the search uses his language. Off-line translations of the documents are immediately available if the user demands such a translation.

A third possibility is the translation of summaries. This method is rather efficient since summaries will be stored anyway. However, the search will be restricted to those aspects of the document that are identified as relevant by the off-line summarizer. Any combination of the three options is possible.

So far we have only discussed the translation between query language and document language that is needed to achieve the matching. In applications where the user cannot be expected to have at least a passive knowledge of the document language or languages, this will not suffice. If CLIR systems cannot provide at least indicative translations of the retrieved documents, their use will be limited to selected user communities and a few language combinations.

Therefore, some CLIR systems integrate automatic translation also in the third branch of the retrieval model, the processing of returned results.

I will now turn to the question of the employed translation methods. Since fully automatic high quality translation is still an unsolved problem, the selection of a suitable translation method depends on the function of translation in the CLIR system and on the envisioned application. For practical reasons, the choice also depends on the languages to be covered since MT systems of sufficient power only exist for a small number of language pairs. We should not be surprised to find that the range of translation methods utilized in CLIR systems reflects the space of approaches in automatic translation.

Oard [2] distinguishes between-knowledge based and corpus-based approaches. He sets up the a hierarchy which follows the common classification in MT.

Quite a few CLIR Systems employ corpus-based translation. It is claimed that well trained corpus-based MT is better suited for dealing with specialized terminology, multi-word expressions and technical jargon. Corpus-Based MT shows advantages in applications where the documents come from a single domain. The performance degrades after domain shifts. Carbonell et al. [5] report on impressive results that could be obtained by training a systems on sufficiently large parallel corpora.

Most corpus-based translation methods work with aligned parallel corpora. Since parallel corpora are hard to find for most language pairs and subject domains, methods have been developed for aligning so called comparable corpora [6]. Ballesteros and Croft utilized monolingual corpora for query extension and pseudo-relevance feedback [7].

Translation by bilingual or multilingual dictionaries is only applied in systems with query

translation, else the advantages of document translation would be lost. Simple bilingual dictionaries will not be sufficient for most applications, as they would miss multi-word terms.

Grammar-based MT systems can be employed for document translation in document processing and for providing the user with indicative translations of retrieved results. They also exhibit satisfactory results when they are utilized for the translation of summaries.

## 4 THE CLIR SYSTEM MULINEX

In the international project Mulinex [8], funded by the Language Engineering sector of the Telematics Applications Programme, we decided to combine high-quality available language technology to build a modular system that integrates a number of technologies currently used in CLIR systems. Although we acknowledge that available technologies are far from being perfect and that intensive efforts in basic research are needed before the dream of CLIR can be realized, we wanted to demonstrate that even with today's imperfect tools a useful application can be achieved.

The partners in this project are;

- Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Coordination)
- TRADOS Germany GmbH
- Bertelsmann Telemedia GmbH
- Grolier Interactive Europe
- DATAMAT - Ingegneria dei Sistemi S.p.A

The main goal of the project was to develop a cross-language Internet search engine that supports selective information access, navigation and browsing in a multilingual environment. During the phase of document gathering by the web spider, documents are analysed in order to obtain useful information about documents in addition to the traditional keyword-based indices. The project emphasises a user-friendly interface, which supports the user by presenting search results along with information about language, thematic category, automatically generated summaries, and allows the user to sort results by multiple criteria. A commercial machine translation system by Logos is used to provide translations of foreign-language documents on demand.

The project relies on basic NLP technology listed below:

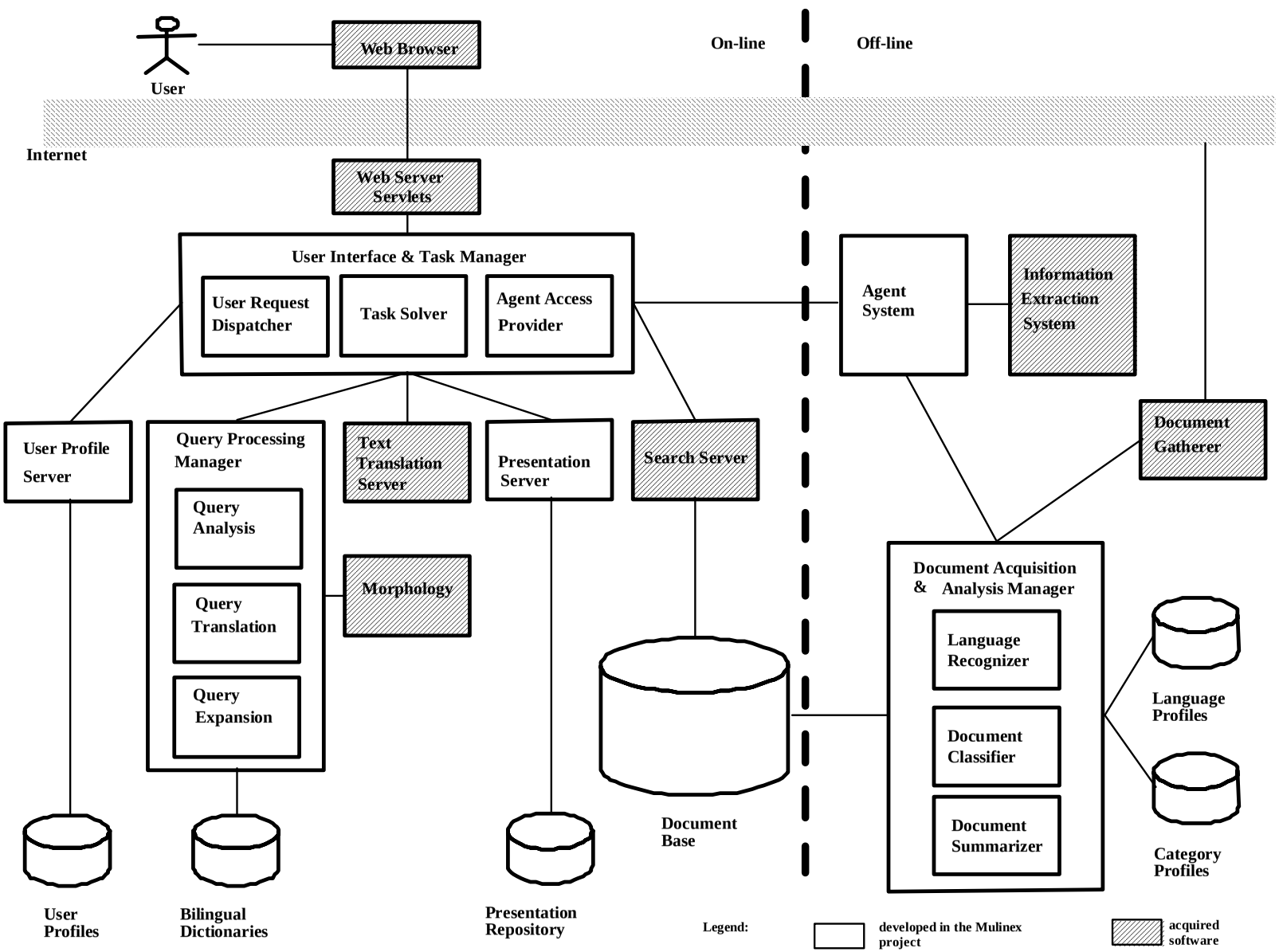
- Morphological analyser, Morphix from DFKI (German), and mmorph (English, French) from the MULTEXT project.
- Shallow parsing, phrase extraction, SMES (Saarbrücken Message Extraction System) from DFKI,
- Full text retrieval, Fulcrum Search Server.
- Multilingual lexicons and thesauri from DFKI.
- Multilingual terminology database MultiTerm plus a web interface from TRADOS.

The search application consists of two sub-systems for intelligent indexing and intelligent retrieval:

Intelligent Indexing: extracts indexing expressions, the language and thematic category from a document and generates a document summary.

The users queries are translated into the languages supported by Mulinex. Users are be able to select documents by language and thematic category. Automatically generated summaries help users in the selection of relevant documents.

The Mulinex System Architecture



## ACKNOWLEDGEMENTS

I am grateful to Gregor Erbach, Joanne Capstick, Abdel Kader Diagne, to our colleagues at the Mulinex Partner companies, and to our wonderful research assistants for their invaluable contributions to the Mulinex system.

## REFERENCES

[1] Gerard Salton. Automatic processing of foreign language documents. Journal of the American Society for Information Science, 21(3).pp.187-194, May 1970.

[2] Douglas W. Oard, "Alternative Approaches for Cross-Language Text Retrieval," in Cross-Language Text and Speech Retrieval, AAAI Technical Report

[3] Christian Fluhr, Multilingual Information Retrieval in Ronald A Cole and Joseph Mariani and Hans Uszkoreit and Annie Zaenen and Victor Zue (Eds.) Survey of the State of the Art in Human Language Technology, Cambridge University Press, 1997.

[4] Chris Buckley, Mandar Mitra, Janet Walz and Claire Cardie, "Using Clustering and SuperConcepts Within SMART: TREC 6," in The Sixth Text Retrieval Conference (TREC-6). National Institutes of Standards and Technology, 1997.

[5] Carbonell, J., Yang, Y., Frederking, R., Brown, R. D., Geng, Y., and Lee, D. Translingual information retrieval: A comparative evaluation. In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence.

[6] M. Wechsler and P. Sheridan and P. Schäuble. Multi-Language Text Indexing for Internet Retrieval. In: Proceedings of the 5th RIAO Conference on Computer-Assisted Information Searching on the Internet, 1997

[7] Lisa Ballesteros and W. Bruce Croft, "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval," in Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, July 1997, pp. 84-91.

[8] Gregor Erbach, Gunter Neumann and Hans Uszkoreit. MULINEX: Multilingual Indexing, Navigation and Editing Extensions for the World-Wide Web. Proceedings AAAI Spring Symposium on Cross-Language Text and Speech Retrieval. Menlo Park CA, 1997.

# Integrating Different Strategies for Cross-Language Information Retrieval in the MIETTA Project

Paul Buitelaar, Klaus Netter, Feiyu Xu
DFKI Language Technology Lab
Stuhlsatzenhausweg 3,
66123 Saarbrücken, Germany
{paulb,netter,feiyu}@dfki.de

ABSTRACT

In this paper we describe an integrated approach to cross-language retrieval within the MIETTA project, whose objective is to build a special purpose search engine in the tourism domain that covers information from a number of geographical regions. MIETTA is designed to enable users to search and retrieve information on the regions covered in their own language preferably. In order to facilitate the user with such functionality, the system includes document translation, cross-language query translation, multilingual generation from information extraction templates and document classification. In addition, query expansion is offered to identify proper query translation and enable template matching for information extraction purposes.

**Keywords:** Cross Language Information Retrieval, Information Extraction, Natural Language Generation, Machine Translation, Query Translation, Query Expansion

## 1 INTRODUCTION

MIETTA (Multilingual Information Extraction for Tourism and Travel Assistance) is a project in the Language Engineering Sector of the Telematics Application Program of the European Union, that combines amongst others technologies from information retrieval with the areas of shallow natural language processing and information extraction (see e.g. [1]). The main objective of the project is to facilitate cross-language retrieval of tourist information in several languages (English, Finnish, French, German, Italian) and on a number of different geographical regions (the German federal state of Saarland, the Southwestern Finnish region centered around Turku and the Italian city of Rome).

Tourism is an application domain, which is almost by its very nature multilingual and which is highly dependent on providing easy access to information on target regions. The world wide web is a growing resource for such information next to other data bases, and one of the main objectives of MIETTA is to collect and disclose such information through a specialized server, helping the user to find and navigate through this information preferably in his own language.

Approaches to cross-language information retrieval, which become relevant in this context, typically include the translation of the user query, or alternatively a full translation of the document base (see e.g. [2] [3] [4]). In the MIETTA project, a third type of multilingual information access is added, where the system can produce the relevant information in different languages by generating from a set of filled-in language independent templates which are obtained through information extraction technology. Information extraction can be used as a restricted, but goal directed search strategy that supplies the user with a fixed set of query options (templates) from which the system can generate natural language representations in preferred languages. Such a strategy presupposes domain specific natural language processing, term extraction and term translation for all languages involved.

Also, in connection to this template-based search the user will be offered the possibility to navigate through a classification hierarchy, which on the surface is language specific, but in its underlying form is language independent, thereby enabling access to documents in multiple languages. Classification-based navigation is meant to allow the user to browse through a refined conceptual classification tree or graph with categories (some of them corresponding to templates) in his own language that will take him to documents in any language. This, like query translation, requires some passive knowledge of the foreign languages.

## 2 THREE STRATEGIES FOR CROSS-LINGUAL INFORMATION ACCESS

In the following we discuss the strategies on cross-lingual information retrieval in more detail as we briefly presented them above, concentrating above all on possible strengths and weaknesses. The first two strategies are geared more towards standard information retrieval approaches, the third one is based on information extraction.

### 2.1 DOCUMENT TRANSLATION

Full document translation can be applied offline to produce translations of an entire document. The function of this translation is twofold, viz., to provide the basis for constructing an index for information retrieval or to offer the user the possibility to browse through a translated version of an original translated in his own language or in a language which he understands. The ideal scenario is that the user enters a query term in his own language, this term is matched against an index constructed from the translation of a document, which is then presented to the user. If he is satisfied with the relevance of the document he can then also access the original document and verify the content in this version.

In a genuine multilingual document base, offline document translation can result in a multiplication of the entire sets of documents in all languages covered. In such a scenario, every monolingual index is constructed from such a set of translated and original documents and covers the content of the entire multilingual original document set. Such an approach is realistic if storage space to account for the multiplication of documents and indices is not a relevant factor, i.e., above all in those cases when a specialised and limited subject domain is addressed.

Document Translation can be the preferred strategy in cross language retrieval, if the purpose is to allow the users to search for foreign documents in their own language and receive results back in that language. In this sense it is clearly a superior option which does not even require passive knowledge of the foreign language from the user.

Machine or (large scale) human translation, however, is not always available as a realistic option for every language pair. Typically machine translation systems only translate between language pairs which involve one of the major languages, such as English, German or Spanish, and ever so often only English will be the common language paired with all other languages. Also, without careful adaptation to a particular domain, machine translation may not always fulfill the necessary quality standards, and will not be sufficiently satisfying for a user even as a purely informative translation which is meant to give him only a rough indication of the relevance of the foreign language document.

Still, it must not be forgotten that even in those cases offline document translation can be useful to construct a 'translated' index, which can be matched against the user's query. Even if it is not foreseen that a user inspects the translations, the 'translated' index can point directly to the corresponding original documents. Since indexing is typically on the basis of words, terms or maximally phrases and since machine translation systems normally offer quite satisfactory quality at this level, such an approach can be a viable or even superior alternative approach to online query translation. One of the main differences between the two alternatives is of course that the user can still select and determine the correct translations in an additional interactive step (see below)

Document translation in the MIETTA project can be implemented on the basis of the LOGOS

machine translation server. The translation directions which can be provided are:

- German ⇒ English, French, Italian
- English ⇒ French, German, Italian, Spanish

Thus, a translation from Italian or Finnish documents into other languages will be impossible for the moment, unless this document is already manually translated into other languages, preferably English.

## 2.2 QUERY TRANSLATION + EXPANSION

Online translation can be applied to the query terms entered by the user. Online query translation will help the user to formulate his query in another language than his own by offering possible translations as options for him to choose from. It makes sense provided that the user either has at least some passive knowledge of the foreign language or that he can have the retrieved document(s) translated automatically or by hand. Thus, the degree of passive knowledge required will very much depend on the additional facilities available.

If a user has no understanding of the foreign language, a retrieved document will be of no use without translation facilities. If he is not able to narrow down the scope by means of disambiguation or determine the relevance of retrieved foreign language documents, even online automatic translation could be of no avail for sheer reasons of volume. The possibility to disambiguate or expand a query in the process or for the purpose of query translation is therefore quite essential.

In MIETTA, we therefore provide a close interaction with query expansion as a support for query translation, offering to the user only semantically adequate translations. Such an approach originates from the believe that the user can be more easily asked about the *meaning* of a search term in his own language, than about the proper *translation*.

Query expansion can add semantic knowledge to the original query by narrowing down the meaning through introducing related terms, synonyms or hyponyms. This extension will exclude those documents in which the related terms do not occur or rather where terms related to alternative readings are prominent and thereby raise precision. Conversely, query expansion can also raise recall by broadening the search space with semantically related terms, synonyms and hyponyms.

For instance, the query “rabbit” can be expanded in a number of ways. By adding the related terms “pork” and “meat”, the search will go in the direction of “cooking and recipes”, whereas adding ”deer” and “wildlife” will look into things like “nature and preservation”.

## 2.3 INFORMATION EXTRACTION

A third method for overcoming the language barrier is to use an interlingual representation as provided by information extraction technology. In the scenarios above the content of the documents is disclosed through an index and a corresponding query and retrieval process, which takes the user to the appropriate document(s). Information extraction on the other hand is based on the assumption that a targeted partial analysis of the texts will provide the possibility to extract selected relevant information and to store this information in the form of templates (frames with predefined slots). Additionally, information extraction combined with natural language generation offers the potential to present selected information in other languages.

One problem with this approach is of course the limited and predefined amount of information that can be disclosed, as it presupposes a partial understanding of the original text. A closely related issue is whether the extracted information can be represented by an interlingua or language neutral format. Clearly, this is not a problem for the kind of data that will be more or less independent from human languages anyway, e.g., numeric information or most kinds of named entities. However, terms in general, corresponding to all kinds of nominal phrases, other than dates and names, will be difficult to represent in a truly interlingual format unless extensive multilingual domain modelling is

achieved before hand, identifying important terms and their possible translations.

Thus, in some sense, the employment of information extraction for multilingual information handling can be interpreted as an attempt to build a highly intelligent machine translation system, which not only translates (relatively close to the surface of the language pairs), but which attempts to understand and isolate some important relations and to present them in different target languages.

As such, it is quite clear that the approach will be less robust and also less flexible than the considerably more shallow text retrieval methods. However, it also offers a substantially more structured access to the content of a document base than standard text retrieval methods do.

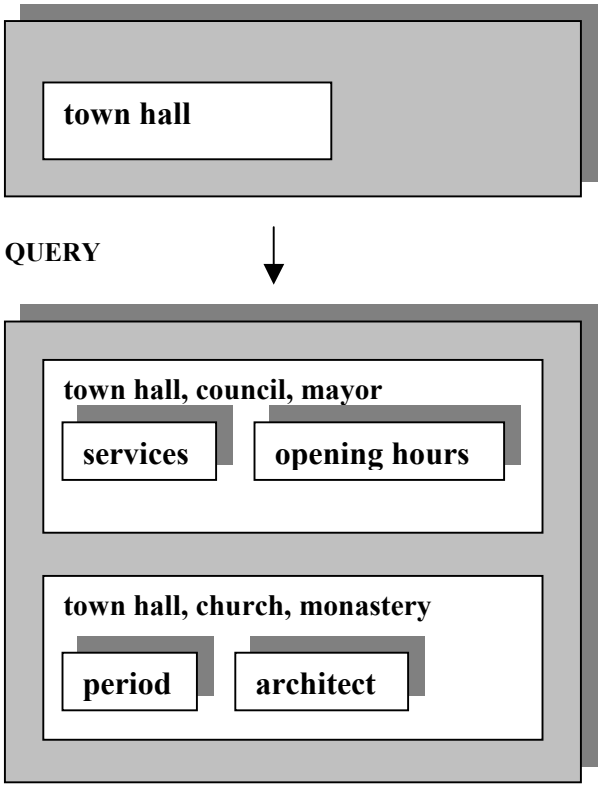
It should be pointed out that the assumption of a full fledged natural language generation component is of course not mandatory, but that the information can be equally well presented to the user in a tabular format employing the slot labels of the template. The generation and translation part in this case would then be reduced only to processing those terms that fill the slots.

## 2.4 A SIMPLE EXAMPLE

The following example serves to illustrate the combination of query expansion and information extraction. It shows how the user can navigate along different meaning dimensions of a search term and how he can exploit information extraction technology to obtain access to deeper and more structured information related to such meaning dimensions.

The search term "town hall" in English is ambiguous between a BUILDING and GOVERNMENT interpretation. Both for clarification within the query language and for query translation into other languages, we need to disambiguate between these. By giving the user a set of related terms, corresponding to one of the interpretations (semantic classes in the classification hierarchy) matching templates can be found.

So, for instance, the related terms "council" and "mayor" corresponding to the GOVERNMENT class will lead to a matching template that includes slots for "services offered" and "opening hours". On the other hand, related terms like "church" and "monastery", which correspond to BUILDING, will lead to a template that includes slots describing "building period" and "architect".




**QUERY EXPANSION**

# 3 LANGUAGE TECHNOLOGY

In the MIETTA project a broad range of language technology methods will be employed to support and enable the above mentioned approaches to information retrieval and information extraction. Most of these are relatively standard technologies that have proven useful in many projects targeting cross-language information retrieval, others are more specific to information extraction and its use in cross-lingual information access.

## 3.1 NATURAL LANGUAGE ANALYSIS

Information extraction in particular, but also the construction of full phrasal indices for information retrieval presumes robust natural language analysis as a pre-processing step. This

will enable filling out templates with appropriate data from natural language texts and identifying suitable phrases as index terms.

Natural language analysis will take place in four stages:

- Tokenisation: recognition of sentence boundaries, proper names (named entities should be recognised before morphological processing) and abbreviations

- Morphological Analysis: lemmatisation, part of speech (POS) and other morphological information

- Part of Speech Tagging: disambiguation of POS by using contextual information

- Shallow Parsing: phrase recognition using POS tagged word stems and the corresponding morpho-syntactic information

For the morphological and grammatical analyses, language specific data as well as additional language independent techniques and tools are needed:

- Language specific monolingual dictionaries

  This includes the use of gazetteers on company names, countries, person names, etc. At least partly, these will be language independent.

- Language specific morphology tools for lemmatisation and decomposition

  In addition to language specific morphological processing, the MIETTA system will also use socalled "fuzzy matching" for IR indexing purposes. Fuzzy matching simply tries, through string manipulation, to match the longest substring or combinations thereof. In some sense this could be seen as language independent morphological processing.

- Language specific NP grammars

  NP grammars in MIETTA need to be dependency grammars, because we consider head-modifier combinations as a basic term unit, instead of for instance a string, keyword or phrase. See below for more details on this.

## 3.2 TERM EXTRACTION + TRANSLATION

One option to facilitate a precise matching of query terms on to the requested documents is to analyze the document set and identify relevant terms. This is the profiling part of information retrieval, in which documents are indexed by the terms that occur in them. The big question then is what should constitute a term.

Traditionally, any string (or parts thereof) in a document will be considered a term. This secures a high recall, but lowers precision considerably. Therefore a more extensive definition of what constitutes a term is needed. For instance, experiments with terms based on (nominal) phrases have shown promising results (e.g. [5]).

**Head-Modifier Terms**

In MIETTA we take this idea even a step further by normalizing phrases into head-modifier constructions that are semantic abstractions over a number of variations of phrases (using [6]). An example will clarify this more clearly. The following nominal phrases:

Bill Clinton's computer terminal
the terminal of Bill Clinton's computer
the terminal of my 20 year old computer
the terminal of my 20 year old son's computer

can be all reduced to the same head-modifier construction:

HEAD: terminal MODIFIER: computer

By taking head-modifier constructions as the basic unit, terms become much more closer to the abstract idea of a "concept" than is possible with just phrase recognition, or simple strings for that matter.

Independent from what a term looks like, is the question what constitutes a relevant term, where relevance is measured relative to the document it occurs in. For this purpose we use standard TF/IDF techniques, combined with a measure of mutual information between heads and modifiers. Mutual information is well known in statistical approaches to natural language processing and is used in a range of methods for determining the strength of connection between words or terms. For more details on mutual information see for instance [7].

**Multilingual Thesaurus**

Identifying relevant terms in a document set can be used to construct a thesaurus of such terms, that can then be used for efficient query expansion. Additionally, by performing term extraction in parallel for each language involved in MIETTA, we can construct a multilingual thesaurus that can also assist in query translation.

Importantly, by synchronizing thesaurus construction with class-based navigation, template definition and query expansion, we arrive at a completely homogeneous approach to concept-based search that allows the user to browse freely and consistently through a network of concepts that take shape in either classes, templates or related terms.

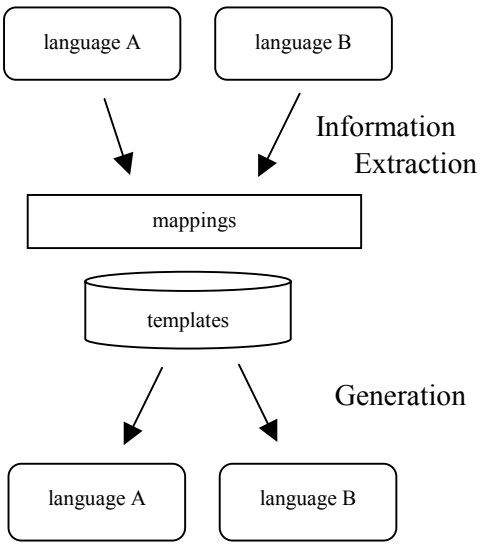
A multilingual thesaurus can be constructed in a number of different ways. First, by a direct transfer from terms in one language onto those in other languages. Secondly, through the use of abstract concept labels that are language independent and map to all languages as an interlingua or, thirdly, by choosing one language (e.g. English) as an interlingua and mapping all other languages onto this one. In MIETTA we take the latter approach, because it is the most straightforward and practical one.

## 3.3 NATURAL LANGUAGE GENERATION

Templates and sets of templates represent an ideal input to natural language generation. In the simplest case they can feed a system of canned text generation. In more complex cases some variations on the syntactic structure might be required. Finally, the most interesting situation is raised by cases when the system has to deal with *sets* of templates that jointly provide an answer to the query posted by the user. Then there is room for a full-fledged system, exploiting sophisticated text planning strategies (see e.g. [8], describing the natural language generation system we use in MIETTA).

**Multilingual Generation**

The figure below (due to [9]) shows how language independent templates interact with generation into individual languages on the one side and mappings between information extracted in these languages on the other.




These mappings are crucial to the multilinguality of the information extraction and natural language generation systems and will be, as mentioned before, organized in a multilingual thesaurus that maps terms in one language onto those of another.

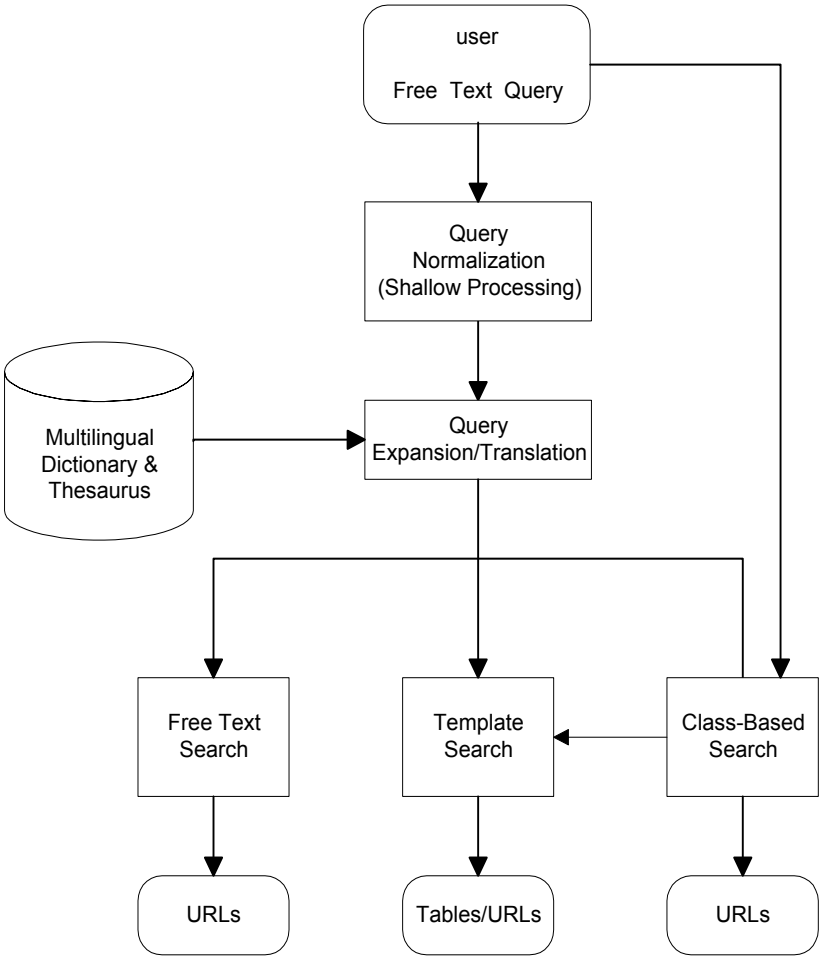
# 4 INTEGRATION

To the user, the different strategies (Document Translation; Query Translation combined with Query Expansion; Information Extraction combined with Multilingual Natural Language Generation and Classification Navigation) should be completely transparent, except for the input

and output languages, which can be set by options.

Such transparency, however, requires an integrated approach to cross-language retrieval that combines these strategies, depending on availability and need. The following figure gives an overview of how query expansion is

In the class-based navigation strategy the user can then browse through the conceptual tree, whose nodes are either directly associated with sets of URLs or document titles, or alternatively with templates related query forms (also hierarchically ordered) that can lead to URLs or document titles indirectly.




envisioned to take a central role in this.

We assume that the user has basically two choices for searching; one is on the basis of free text queries, the other on the basis of concept-oriented classifications. Free text queries can take the user either to different kinds of full text indices and/or it can take the user to a classification hierarchy, if the search term happens to map onto on of the conceptual classes.

The query forms are meant to support the user by guiding his search, i.e., making it clear to him on what aspects of a certain class or category he can pose more specific queries. To further support a precise formulation of the (free text) query, the user can submit it to the query expansion component. Here, the query is first processed and normalized, where appropriate, and

can then be enriched or made more precise on the basis of a thesaurus.

While these scenarios so far can be implemented also in a monolingual environment, multilingual search adds another layer of complexity. Thus, depending on the input and output languages, that is, query and document language, we can envision a number of different scenario's. The most simple one is where the query and document language are the same. This is for instance the case when a German user asks for information about the Saarland. This region only has information available in German, so for this scenario no translation is necessary at all. The German user will simply present his query to the system that will then normalize it into a head-modifier construction as far as possible and allow the user to expand his query with additional German terms from the thesaurus. Finally, given the user's preference, the simple or expanded query will be transmitted to one out of the sub-processes of query processing: free text query or template query. In this scenario both of these are possible.

A somewhat more complicated case is the situation where the query language is different from the document language, although translations of these documents exist in the query language. This is for instance the case when a German user requires information about Rome. Documents in Rome are available in five languages (English, French, German, Italian and Spanish), in parallel human translation. This allows the system to access term indexes in the query language (German) and the whole process will be equal to the previous one. Where no human translation exists, but an automatic translation of the document into the query language has been carried out, the user should be given the choice to see the translated document as a response to his query or to be taken to the original because of inadequate translation quality.

Finally, the really complicated cases are where the query language is different from the document language, and no translations of these documents exist in the query language. This is the case when a German user requires information about Turku, or when a Finnish user asks for information about the Saarland. In both cases, no translation from or into Finnish is available and the system has to perform a combination of query translation and query expansion in order to facilitate template query: information extraction in the document language and natural language generation in the query language.

Here, for instance, a German user submits a query to the system that will be normalized to a German head-modifier construction and then translated into Finnish by the query expansion and translation component. This translation, however, will be only available in the background, because it is not assumed that the German user will be able to judge the translation on correctness anyway. Also, the expansion component, given the information available in the multilingual thesaurus, already expanded this translation into a template that can be matched on Finnish documents. The results thereof are then synchronized with their German translations that again are available through the multilingual thesaurus, and these are used in generating a German text or table.

All of this, of course, depends on the coverage of the terms in the thesaurus, which should therefore be quite extensive and to the point, that is, covering many domain specific (simple or complex) terms. Nevertheless, if a query term is not matched by the thesaurus, the fall back option is a free text query with the translated query. This, however, has two big disadvantages. First, the translation is not semantically guided and will therefore be less accurate. Secondly, even if good results are found these can be only presented in Finnish, which will largely be useless to the German user.

## 5 CONCLUSION

We presented a cross-language retrieval strategy that combines document translation, query translation, query expansion, information extraction, natural language generation and class-based navigation. By integrating all of these into a coherent approach, the different strategies for cross-language retrieval that are needed can be left transparent to the user.

## REFERENCES


[1] J. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel and M. Tyson. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In: E. Roche and Y. Schabes, editors, *Finite State Devices for Natural Language Processing*, pages 383-406, MIT Press, 1997.

[2] G. Grefenstette. (ed.) Proceedings of the SIGIR96 Workshop on Cross-Linguistic Information Retrieval, 1996.

[3] W.G. ter Stal, J-H Beijert, G. de Bruin, J. van Gent, F.M.G. de Jong, W. Kraaij, K. Netter, G. Smart (1998) Twenty-One: Cross-language disclosure and retrieval of multimedia documents on sustainable development. In: Computer Networks and ISDN Systems 30(13), pages 1237-1248, Elsevier Science BV

[4] MULINEX: Multilingual Indexing, Navigation and Editing Extensions for the World-Wide Web. Proceedings AAAI Spring Symposium on Cross-Language Text and Speech Retrieval. Menlo Park CA, 1997.

[5] D. Evans. Lessons from the CLARIT Project In: Proceedings of SIGIR93, Pittsburgh, PA, USA, 1993.

[6] R. Backofen, J. Baur, M. Becker, C. Braun and G. Neumann. An Information Extraction Core System for Real World German Text Processing. Proceedings of the 5th ANLP, Washington DC, 1997.

[7] K. W. Church and P. Hanks. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, pages 22-29, 16, 1990.

[8] S. Busemann and H. Horacek. A Flexible Shallow Approach to text Generation. In: E. Hovy (ed.) Proceedings of the Nineth International Natural Language Generation Workshop (INLG98), Niagara-on-the-Lake, August 1998.

[9] L. Dini. Parallel Information Extraction Systems for Multilingual Information Access. Paper presented at Euriscon, 1998.

# Cross-language information retrieval in Twenty-One: Using one, some or all possible translations?

Djoerd Hiemstra and Franciska de Jong

University of Twente, CTIT
P.O. Box 217, 7500 AE Enschede
The Netherlands
{hiemstra,fdejong}@cs.utwente.nl

## Abstract

This paper gives an overview of the tools and methods for Cross-Language Information Retrieval (CLIR) that were developed within the Twenty-One project. The tools and methods are evaluated with the TREC CLIR task document collection using Dutch queries on the English document base. The main issue addressed here is an evaluation of two approaches to disambiguation. The underlying question is whether a lot of effort should be put in finding *the* correct translation for each query term before searching, or whether searching with more than one possible translation leads to better results? The experimental study suggests that in terms of average precision, searching with ambiguities leads to better retrieval performance than searching with disambiguated queries.

**Keywords:** Cross-Language Information Retrieval, Statistical Machine Translation.

## 1 Introduction

Within the project Twenty-One a system is built that supports Cross-language Information Retrieval (CLIR). CLIR supports the users of multilingual document collections by allowing them to submit queries in one language, and retrieve documents in any of the languages covered by the retrieval system. For this type of functionality the translate option offered by some web search engines does not suffice, because it does not help the users to identify material that they might want to have translated. Since this approach presupposes that the users have already found the relevant document in its original foreign language, it fails to support exactly that part of a search in a multilingual environment which is the most difficult one, viz., to formulate a query which will then take the user to the foreign language document in whose content he might be interested in. In order to support the retrieval of documents irrespective of the document language, either off-line document translation (DT), off-line index translation, or on-line query translation (QT) is required. From a practical point of view QT is enforced in environments where it would be impossible to produce translations for all documents in the document base and/or to produce translated indices for all languages. However, it has the disadvantage or at least restriction that the user must know the foreign languages at least up to the degree of a passive understanding of this language. The alternatives document translation (DT), or index translation do not necessarily presuppose even a passive knowledge of the foreign language.

Systems operating in a comparatively controlled environment, where the documents are limited either to a specific domain or to a limited number of data and document bases are likely to use DT. Twenty-One is a representative of this category. It has a clear target domain. viz. sustainable development, and with its strong focus on the disclosure of paper documents, which have to be scanned and OCRed, heavy preprocessing and storage has to be reckoned with anyhow. Off-line translation rather than translating query terms during retrieval has important advantages for the way the most critical part of the translation task is dealt with: disambiguation. As all index terms (NPs) are kept in their original position, contextual information is accessible for the disambiguation algorithms that are part of the translation software. Currently disambiguation in Twenty-One can be pursued in three ways:

- selection the dictionary preferred translation
- the use of domain specific dictionaries that are automatically generated on the basis of statistically processed parallel corpora (suited for specific applications only)
- disambiguation on the basis of the frequency of noun phrases in the document collection

This paper is organised as follows. Section 2 explores possibilities for the comparison of the DT approach with the QT approach. Section 3 introduces three basic methods for the QT approach to CLIR. Section 4 addresses heuristics and statistics for translation. Section 5 discusses the setup of our experiments and experimental results. Finally, section 6 contains concluding remarks.

## 2 Empirically comparing DT to QT

As said in the introduction DT has important advantages. Firstly, it can be done off-line. Secondly, if a classical machine translation is used, it is possible to present the user a high quality preview of a document. Thirdly, there is more context available for lexical disambiguation. This might lead to better retrieval performance in terms of precision and recall. For several types of applications, the first and second advantage may be a good reason to choose for DT. The third advantage however is more hypothetical. Does the DT approach to CLIR using classical machine translation really lead to better retrieval performance than the QT approach using a machine readable dictionary?

For a number of reasons it is very difficult to answer this question on the basis of empirical evidence. A first problem is that in the QT approach searching is done in the language of the documents while in the DT approach searching is done in the language of the query. But it is a well known fact that IR is not equally difficult for each language. A second problem is that, for a sound answer to the question, we need a machine translation system and a machine readable dictionary that have exactly the same lexical coverage. If the machine translation system misses vital translations that the machine readable dictionary does list, we end up comparing the coverage of the respective translation lexicons instead of the two approaches to CLIR. Within the Twenty-One project we have a third, more practical, problem that prevents us from evaluating the usefulness of the used translation system (LOGOS) against the usefulness of the machine readable dictionaries available within the project (Van Dale). The Van Dale dictionaries are entirely based on Dutch head words, but translation from and to Dutch is not supported by LOGOS. All these considerations urge us to rephrase the the issue into a more manageable question.

A first, manageable, step in comparing DT with QT might be the following. What is, given a translation lexicon, the best approach for QT: using one translation for each query term or using more than one translation? Picking one translation is a necessary condition of the DT approach. For QT we can either use one translation for searching, or more than one. The question one or more translations also reflects the classical precision / recall dilemma in IR: picking one specific translation of each query term is a good strategy to achieve high precision; using all possible translations of each query term is a good strategy to achieve high recall.

## 3 Methods for QT

As said in the previous section this paper compares CLIR using one translation per query term with CLIR using more than one translation per query term. We will report the results of retrieval experiments using the Dutch queries on the English TREC CLIR task collection. A Dutch query will be referred to as the source language query; the English query will be referred to as the translated query. The experiments can be divided into three categories:

1. QT using one translation per source language query term
2. QT using unstructured queries of all possible translations per source language query term
3. QT using structured queries of all possible translations per source language query term

### 3.1 Using one translation per query term

If only one translation per query term is used for searching, the translation process must have some kind of explicit disambiguation procedure. This procedure might be based on an existing machine translation system, or alternatively, on statistical techniques or heuristics. After disambiguation, the translated query can be treated the way a

query is normally treated in a monolingual setting. A 'normal' monolingual setting in this context is retrieval on the basis of a statistical 'bag-of-words' model like e.g. the vector space model [10] or the classical probabilistic model [9]. In the next section, the use of a bag-of-words model will be referred to as the *unstructured queries*-option.

In section 4 a number of heuristics and statistics for disambiguation will be explored. As explained in section 2 we will not be able to actually use machine translation for disambiguation. It is however possible to define an upper bound on what is possible with the one-translation approach by asking a human expert to manually disambiguate the output of the machine readable dictionary. We hypothesise that QT using a machine translation system with the same lexical coverage as the machine readable dictionary will not result in better retrieval performance than QT using the manually disambiguated output of the same dictionary.

### 3.2 Using unstructured queries

If more than one translation per source language query term is used for searching we might still treat the translated query as a bag-of-words. As we will see in section 5 the way of weighting the possible translations is crucial for unstructured queries. In particular it is important to normalise the possible translations in such a way that for each source language query term the weights of possible translations sum up to one. Not using normalisation will make source language query terms with a lot of possible translations unintentionally more important than source language query terms that have only less possible translations.

$$\text{similarity}(Q, D) = \sum_{k=1}^{l} w_{qk} \cdot w_{dk}$$
$$w_{qk} = tf(k, q)$$
$$w_{dk} = \log(1 + \frac{tf(k, d)}{df(k) \sum_t tf(t, d)} \cdot \frac{0.15 \sum_t df(t)}{0.85})$$

Figure 1: vector product weighting algorithm

Instead of using one of the bag-of-words models mentioned above, we will use a weighting algorithm based on a new model of information retrieval: the linguistically motivated probabilistic model [2, 5]. Figure 1 lists the weighting algorithm that was used to rank the documents given a translated query. In this formula $tf(t, d)$ is the term frequency of the term $t$ in the document $d$ and $df(t)$ is the document frequency of the term $t$.

### 3.3 Using structured queries

If all possible translations are treated as one bag-of-words we ignore the fact that a document containing one possible translation of each source language query term is more likely to be relevant than a document containing all possible translations of only one source language query term. The boolean model or weighted boolean models (see e.g. [10]) can be used to retrieve only those documents that contain a translation of all or most of the source language query terms [6]. Disjunction can be used combine possible translations of one source language query term. Conjunction can be used in a way that the translated query reflects the formulation of the source language query.

Our structured query approach is based on the linguistically motivated model. A structured query has to be formulated in conjunctive normal form, which is the form in which it is automatically produced after dictionary based translation. The definition of the conjunction is simply the definition of the probability ranking function as introduced in [2] where $T_1, T_2, \cdots, T_n$ is a query of length $n$ and $D$ is a document id.

$$P(T_1, T_2, \cdots, T_n|D) = \prod_{i=1}^{n} P(T_i|D)$$

Disjunction of $m$ possible translations of the source language query term on position $i$ is defined as follows.

$$P(T_{i1} \cup T_{i2} \cup \cdots \cup T_{im}|D) = \sum_{j=1}^{m} P(T_{ij}|D)$$

The structured query weighting algorithm implicitly normalises the possible translations in a disjunction. Explicit normalisation as done for unstructured queries is no longer necessary. If there are no disjunctions in the query (that is, if there is only one translation per source language query term) then the structured ranking formula will produce exactly the same results as the weighting algorithm of figure 1. Structured queries are generated automatically by the translation module and may take relative frequencies of possible translations into account. A more detailed description of the algorithm will be published in the near future.

### 3.4 An example

Figures 2 and 3 give an example of an English query {*third*, *world*} that is used to search a French collection. It is assumed that the English term *third* has two possible French translations: *tiers* and *troisième* and that the English term *world* has three possible translations: *monde*, *mondial* and *terre*.

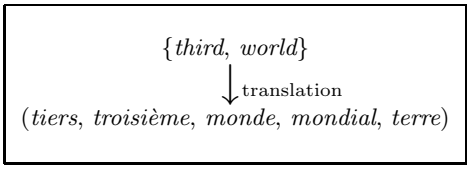



Figure 2: translation using an unstructured query

The result of figure 2 could be used directly for searching the French collection (see **run2a** in section 5), but this would make the term *world* in the source language query more important (because it has more possible translations) than the word *third*. The query weights of the weighting algorithm of figure 1 might therefore be used to make the contribution of *third* as high as the contribution of *world* by reweighting (normalising) the possible translations of *third* to 0.5 and the possible translations of *world* to 0.33 (see **run2c** in section 5). If one of the possible translations of one source language query term is more probable than the other(s), this possible translation might be weighted higher than the other(s) while keeping the normalisation in tact.

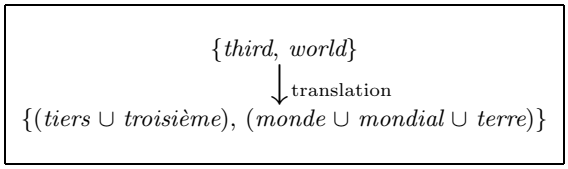



Figure 3: translation using a structured query

The structured query of figure 3 reflects the possible translations of the source language query terms in an intuitive way. Possible translations of one original query term might be weighted differently. Normalisation is an implicit feature of the weighting algorithm.

## 4 Heuristics and statistics for QT

This section lists a number of information resources that can be used to identify the proper translation or proper translations of a query term. The section briefly describes information that is explicitly or implicitly in the dictionary and information from other sources like parallel corpora and the document collection itself.

### 4.1 Dictionary preferred translation

The VLIS lexical database of Van Dale Lexicography list for each entry explicitly one *preferred* translation which is considered the most commonly used one. Replacing each query term with the preferred translation is a simple, but possibly effective, approach to CLIR.

### 4.2 Pseudo frequencies

The Van Dale database contains also explicit information on the sense of possible translations. Some Dutch head words carry over to the same English translation for different senses. For example the Dutch head word *jeugd* may be translated to *youth* in three senses: the sense of 'characteristic', 'time-frame' and 'person'. The 'person' sense has a synonym translation: *youngster*. As *youth* occurs in the dictionary under three senses and *youngster* under one sense, we assign *youth* a weight that is three times as high as the weight for *youngster*. The assumption made by weighting translations is that the number of occurrences in the dictionary may serve as rough estimates of actual frequencies in parallel corpora. In other words: the number of occurrences in the dictionary serve as *pseudo frequencies*. Ideally, if the domain is limited and parallel corpora on the domain are available, weights should be estimated from actual data as described in section 4.3.

### 4.3 Frequencies from parallel corpora

The Twenty-One system contains documents on the domain of sustainable development. Translation in Twenty-One is done using a general purpose dictionary (Van Dale) and a general purpose MT-system (LOGOS), but these resources are not very well suited for domain-specific jargon. Domain-specific jargon and its translations are implicitly available in parallel corpora on sustainable development. Translation pairs can be derived from parallel corpora using statistical co-occurrence by so-called word alignment algorithms. Within the Twenty-One project word

alignment algorithms were developed that do the job in a fast and reliable way [3, 4]. Domain specific translation lexicons were derived from Agenda 21, a UN-document on sustainable development that is available in most of the European languages including Dutch and English.

For the experiment we merged the automatically derived dictionary with the Van Dale dictionary in the following way. For each entry, we added the pseudo frequencies and the real frequencies of the possible translations. Pseudo frequencies are usually not higher than four or five, but the real frequencies in the parallel corpus may be more than a thousand for frequent translation pairs. Adding pseudo frequencies and real frequencies has the effect that for possible translations that are frequent in the corpus the real frequencies will be important, but for translations that are infrequent or missing the pseudo frequencies will be important.

Translation pairs that have a frequency of one or two in the parallel corpus may-be erroneously derived by the word alignment algorithm. If, however, such an infrequent translation pair is also listed in the machine readable dictionary, then the pair was probably correct. Therefore we added a bonus frequency of three to each possible translation that is both in the corpus and in Van Dale.

## 4.4 Context for disambiguation

The techniques introduced so far do not resemble techniques that are actually used in machine translation systems. Traditionally, disambiguation in machine translation systems is based on (syntactic) context of words. In this section a statistical algorithm is introduced that uses context of the original query words to find the best translation. The algorithm uses candidate noun phrases (NPs) extracted from the document base to disambiguate the NPs from the query. NPs were extracted using the standard tools as used in the Twenty-One system: the Xerox morphological tools and the TNO parser. The NPs were sorted and then counted, resulting in a list of unique phrases with frequency of occurrence.

The introduction of NPs (or any multi-word expression) in the translation process leads to two types of ambiguity: sense ambiguity and structural ambiguity. Figure 4 gives an example of the French translation chart of the English NP *third world war*. Each word in this NP can have several translations that are displayed in the bottom cells

| | | |
|---|---|---|
| - | | |
| tiers monde | guerre mondiale | |
| troisième<br>tiers | monde<br>mondiale<br>terre | guerre<br>bataille |
| third | world | war |

Figure 4: translation chart of *third world war*

of the chart, the so-called sense ambiguity. According to a list of French NPs there may be two candidate multi-word translations: *tiers monde* for the English NP *third world* and *guerre mondiale* for *world war*. These candidate translations are displayed in the upper cells of the chart. Because the internal structure of NPs was not available for the translation process, we can translate a full NP by decomposing it in several ways. For example *third world war* can be split up in the separate translation of either *third world* and *war* or in the separate translation of *third* and *world war*. The most probable decomposition can be found using techniques developed for stochastic grammars (see e.g. [1]). The probabilities of the parse trees can be mapped into probabilities, or weights, of possible translations. A more detailed description of the algorithm can be found in [8].

## 4.5 Manual disambiguation

The manual disambiguation of the dictionary output was done by a native speaker of English. She had access to the Dutch version of the topics and to the English dictionary output consisting of a number of possible translations per source language (Dutch) query word. For each Dutch word, one of the possible English translations had to be chosen, even if the correct translation was not one of them.

## 4.6 Other information

In the experiments described in this paper we ignored one important source of information: the multi-word entries in the Van Dale dictionaries. Multi-word expressions like for instance *world war* are explicitly listed in the dictionary. For the experiments described in this paper we only used word-by-word translations using the single word entries.

## 5 Experimental setup and results

In section 3 we identified three methods for QT: using one translation per query term, using a unstructured query of all translations per source language query term and using a structured query of all translations per source language query term. Each method is assigned a number 1, 2 or 3. In section 4 five sources of information were identified that may be used by these methods: dictionary preference, pseudo frequencies, parallel corpora, context in noun phrases and human expertise. Given the five information sources we identified seven (two experiments were done both with and without normalisation) basic retrieval experiments or runs that are listed in table 1. Each experiment is labelled with a letter from *a* to *g*.

| run name | technique to weight translations / pick the best translation |
|---|---|
| run?a | no weights used / dictionary preferred translation. |
| run?b | weight by pseudo frequencies. |
| run?c | normalise weights of possible translations (run?a) |
| run?d | weight by normalised pseudo frequencies |
| run?e | normalised 'real' frequencies estimated from the parallel Agenda 21 corpus. |
| run?f | weight by using noun phrases from documents (including normalisation) |
| run?g | disambiguation by a human expert |

Table 1: information to weight translations and / or pick the best translation

The combinations of seven information sources and three methods define a total number of 21 possible experiments. After removing combinations that are redundant or not informative 15 experiments remain.

In the remainder of this section we will report the results of 15 experiments on the TREC CLIR task test collection [11] topics 1-24 (excluding the topics that were not judged at the time of TREC-6 leaving 21 topics). The Dutch topics will be used to search the English documents. Experiments will be compared by means of their non-interpolated average precision, average precision in short. Additionally, the result of each experiment will be compared with the result of a monolingual base line run, which is the result of queries based on the English version of the TREC topics. The monolingual run performs at an average precision of 0.403. All experiments were done with the linguistically motivated experimental retrieval engine developed at the University of Twente.

### 5.1 One translation runs

Table 2 list the results of the one translation runs. Normalisation of translation weights is not useful for picking the best translation. Therefore the table does not list **run1c** and **run1d**. (**run1d** would give exactly the same results as **run1b**.)

| run name | average precision | relative to baseline (%) |
|---|---|---|
| run1a | 0.262 | 65 |
| run1b | 0.231 | 57 |
| run1e | 0.282 | 70 |
| run1f | 0.269 | 67 |
| run1g | 0.315 | 78 |

Table 2: results of 'one-translation' runs

Not surprisingly, the manual disambiguated run outperforms the automatic runs, but it still performs at 78 % of the monolingual run. Translation ambiguity and missing terminology are the two primary sources of error in CLIR [7]. We hypothesise that the loss of performance is due to missing terminology and possibly errors in the translation scripts. The 78 % performance of the monolingual base line is an upper bound on what is possible using a one-translation approach on the TREC CLIR collection.

The best automatic run is the run using corpus frequencies **run1e**. This is a surprise, because we used a relatively small corpus on the domain of the Twenty-One demonstrator which is *sustainable development*. Inspection of the topics however learns us that a lot of topics discuss international problems like air pollution, combating AIDS, etc. which fall directly in the domain of sustainable development.

The dictionary preferred run **run1a** performs reasonable well. The run using context from noun phrases **run1f** performs only a little better. Pseudo frequencies **run1b** are less useful in identifying the correct translation.

### 5.2 Unstructured query runs

Table 3 list the results of the unstructured query runs using all possible translations of each original query term. We experimented with all information sources except for the human expert.

| run name | average precision | relative to baseline (%) |
|---|---|---|
| run2a | 0.180 | 45 |
| run2b | 0.162 | 40 |
| run2c | 0.268 | 67 |
| run2d | 0.308 | 76 |
| run2e | 0.305 | 76 |
| run2f | 0.275 | 68 |

Table 3: results of 'unstructured query' runs

A first important thing to notice is that the normalisation of the term weights is a prerequisite for good performance if all possible translations per source language query term are used in an unstructured query. Not using the normalisation, as done in **run2a** and **run2b** will drop performance to a disappointing 40 to 45 per cent of the monolingual base line.

More surprisingly, the pseudo frequency run **run2d** and the real frequency run **run2e** now perform equally well and both approach the upper bound on what is possible with the one translation approach (**run1g**). Although the pseudo frequencies are not very useful for identifying the best translation, they seem to be as realistic as real frequencies if used for weighting the possible translations.

### 5.3 Structured query runs

Table 4 lists the results of the structured query runs. Normalisation of term weights is implicit in the structured query, so **run3a** and **run3b** will give exactly the same results as **run3c** and **run3d** respectively.

| run name | average precision | relative to baseline (%) |
|---|---|---|
| run3c | 0.311 | 77 |
| run3d | 0.330 | 82 |
| run3e | 0.335 | 83 |
| run3f | 0.323 | 80 |

Table 4: results of 'structured query' runs

The four runs do not differ as much in performance as their unstructured equivalents, which suggests that the structured queries are more robust than the unstructured queries. Again, the pseudo frequency run **run2d** and the real frequency run **run2e** perform almost equally well. Three out of four runs perform better than the manually disambiguated 'one translation' run **run1g**.

## 6 Conclusion

This paper gives an overview of methods and information sources that can be used for CLIR. Evaluation of these methods on the TREC cross-language collection indicates that using all possible translations for searching leads to better retrieval performance in terms of average precision than using just one translation. The results of the manually disambiguated run suggest that not much can be gained by putting a lot of effort in explicit disambiguation of possible translations. If proper weighting of possible translations is used, disambiguation is done implicitly during searching.

This paper briefly introduced a new method to rank document using structured queries. Mathematical details of the method will be published in the near future. In the cross-language experiments reported on here, structured queries outperform the unstructured queries.

## Acknowledgements

The work reported here was developed in close co-operation with Wessel Kraaij from TNO-TPD Delft as a preparation for the TREC-7 experiments. It was Wessel's idea to add the manually disambiguated run to our experiments. (A similar -unpublished- experiment with manually disambiguated queries was conducted at TNO for English-German cross-language retrieval.) We are very thankful to Wessel for his advice and support on setting up these experiments. Furthermore, we like to thank Lynn Packwood for the manual disambiguation of the Van Dale dictionary output and Thijs Westerveld for implementing the interface on the corpus dictionary.

## References

[1] Rens Bod. *Enriching Linguistics with Statistics: Performance Models for Natural Language*. Academische Pers, 1995.

[2] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In C. Nicolaou and C. Stephanidis, editors, *Proceedings of the Second European Conference on Research and Advanced Technology*

*for Digital Libraries, ECDL-2*, pages 569–584, 1998.

[3] D. Hiemstra. Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus. In *Proceedings of eightth CLIN meeting*, 1998.

[4] D. Hiemstra, F.M.G. de Jong, and W. Kraaij. A domain specific lexicon acquisition tool for cross-language information retrieval. In *Proceedings of RIAO'97 Conference on Computer-Assisted Searching on the Internet*, pages 255–266, 1997.

[5] D. Hiemstra and W. Kraaij. Trec-7 working notes: Twenty-One in ad-hoc and clir. In *Proceedings of the seventh Text Retrieval Conference, TREC-7*, (draft, to appear).

[6] D.A. Hull. Using structured queries for disambiguation in cross-language information retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence, 1997.

[7] David Hull and Gregory Grefenstette. A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[8] W. Kraaij and D. Hiemstra. Cross-language retrieval with the Twenty-One system. In E. Voorhees and D. Harman, editors, *Proceedings of the 6th Text Retrieval Conference TREC-6*, pages 753–761. NIST Special Publication 500-240, 1998.

[9] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976.

[10] G. Salton and M.J. McGill, editors. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[11] E.M. Voorhees and D.K. Harman. Overview of the 6th text retrieval conference. In *Proceedings of the 6th Text Retrieval Conference TREC-6*, pages 1–24. NIST Special Publication 500-240, 1998.

# Information Extraction from Bilingual Corpora for Machine-Aided Translation

**David A. Hull**

Xerox Research Centre Europe
6 chemin de Maupertuis, 38240 Meylan France
hull@xrce.xerox.com

## Abstract

There is a large and growing market for software tools to help human translators. The best way to speed the translation process is to leverage off of existing translated texts. Researchers have recently developed powerful tools for the extraction of translation equivalents from databases of aligned sentences, often known as translation memories. In this paper, we will describe the TRINITY alignment system which uses linguistic tools and statistical alignment algorithms to automatic generate bilingual lexicons and other resources useful for translation.We will show how this technology can improve translation memory systems, which should directly increase the productivity of translators.

**Keywords:** Machine-Aided Translation, Translation Memory, Word Alignment, Terminology Extraction

## 1 Introduction

Driven by the forces of globalization, there is a large and growing market for technology which aids in the creation and distribution of commercial and technical documentation. High quality translation is a critical part of that technology. Current machine translation systems are not capable of meeting this need, and the prospects for automatic high-quality machine translation in the near term are not bright. However, translation by hand is a labor intensive process, so software systems which can improve the quality, consistency, and efficiency of the translation process will be in high demand. Previously existing translations represent an important resource which can be exploited to achieve these goals. Current systems rely on translation memories to provide translators with quick access to existing translations.

A translation memory (TM) is an example database of sentences and their translations. These systems improve translator productivity by finding close or exact matches between a sentence which needs to be translated and existing sentences in the source language part of the database. The linked target language sentence is then proposed as the starting point for human translation. If the match is close enough, it will take less time to edit the existing translation than to translate from scratch. This technology is particularly valuable for performing updates of existing documentation, where much of the text will often be identical to the previous version. In industries where many similar products are developed, there is also a real opportunity to reuse existing translations. In one case, MKMS, a new Xerox small business responsible for the development of translation aid tools, reported that their TM system found exact matches for 70% of the sentences in an automobile repair manual, using a translation memory consisting of the repair manuals

for other models. Exact sentence matching leads to a direct savings in translation cost and time, but the real challenge is improving the TM system's ability to find partial sentence matches.

While translation memories can be extremely valuable, they are also quite fragile. Slight variations in mark-up or syntactic structure can often lead them astray. The sentence match must be quite close, otherwise it takes less time and effort to translate from scatch. This means that it is often difficult to obtain high coverage without a huge sentence database. Therefore, there is an important opportunity to improve the coverage of translation memory systems by recognizing and exploiting partial matches. This paper will present TRINITY[1], our information extraction algorithm for parallel sentenced-aligned corpora[2], which can be used to match text at the sub-sentence level and automatically suggest a possible translation. In the next section, we will demonstrate how the TRINITY system can be used to provide interactive help in translation. The third and fourth section will introduce the linguistic methods used for sentence chunking and the statistical word alignment algorithms which automatically extract phrasal translations.

## 2 Subsentential Translation Matching

Table 1 shows an example translation memory interface. The sentence to be translated appears at the top and three similar sentences from the translation memory are listed at the bottom. The similarity score between sentences (given in the left column) is computed using string matching or information retrieval search strategies. The differences between the sentence to be translated and the TM source sentence are shown in brackets (text from the TM on the bottom). Note that the second and third TM source sentence are identical but their translations are different. This does not necessarily mean that the translations were inconsistent. It may be that different translations of the English sentence are appropriate depending on the context or the discourse structure in the regions where the sentences were extracted. This implies that the translator may find it necessary on some occasions even to edit exact matches from the translation memory. If the first TM target sentence is used as the starting point, only two words: release and left-hand, need to be manually translated. The target sentence after translation appears in the second box.

The replace and edit strategy works well for the highest scoring example sentence, but it would not be advisable for the other two TM sentences where most of the text is not the same. Let us assume for the moment that the first sentence did not exist in the TM database. What can we do to help with the translation process? In order to answer this question, we look at the data which can be extracted by the TRINITY term alignment system. Table 2 presents an example sentence pair and the different levels of term alignment which the system can generate. It can align words, phrases, and dependency relation derived from shallow parsing. The granularity of phrase chunking and alignment can be different, depending on which language is using as the starting point. For example, in English, the system treats *wheel suspension* and *right-hand side* as independent units, while the French chunker identifies *élément de suspension droit* as a single unit. The parsing is performed independently in each language and the relation pairs are subsequently aligned. Details of the chunking and alignment algorithms will be presented in the next two sections.

After applying the TRINITY system to a translation memory database, we have a de-

[1] TRINITY = TRanslation INduction via ITerative probabilitY estimation.

[2] A parallel sentence-aligned corpus is a collection of documents where each source language sentence is linked to a target-language translation. It could easily serve as a translation memory, although the sentence alignment is usually constructed by automatic methods, meaning that there will be errors.

rived database of aligned words, phrases, and dependency relations along with frequency counts and probability estimates for their alignment. The phrase alignment can be from English to French or French to English, depending on which language is going to be the source and which the target for the current translation exercise. The same linguistic processing and chunking is applied to each new sentence as it arrives and the component units are looked up in the derived database. The matches are presented in the following preference order:

(1) Dependency relation match

(2) Exact phrase match

(3) Fuzzy phrase match

(4) Word match

The preference order reflects the precision of the match. An agreement between the words and their dependency relation is more exact and more powerful than a simple phrase match, which is in turn more powerful than a match on the component words. Word-level matches are often highly ambiguous. A fuzzy phrase match is match between a word a different word from the same noun compound. See Table 3 below for an example.

Table 3 demonstrates how the system would process the sentence presented in Table 1 when the closest match is removed from the translation memory system. First, the sentence is chunked into three phrases and three dependency relations are extracted. For each unit, the closest matches from the alignment database are extracted, ordered, and presented. Note that for the direct object relation: DOBJ(release, suspension), no match is found. Furthermore, no sentence is found which contains *release* (used as a verb) followed by *suspension*. However, the system notes that *wheel* is in the same noun compound as *suspension* and looks for a fuzzy phrase match based on *release* and *wheel*, finding two different dependency pairs. These relations come from to the second and third sentence appearing in the translation memory interface example (Table 1). Finally, if forced to rely on the word *release* alone, we find that there are 14 possible translations in the translation memory database! These translations could be ordered and presented with frequency information although this isn't done in Table 3.

In order to speed the translation process, this information needs to be presented to the translator in an effective manner. We don't have any special solution other than to suggest that one could duplicate the existing translation memory interface, but also include pointers back to the original sentence pair from which the aligned units were extracted. It might suffice to look at the full noun phrase rather than the entire sentence. For example, the translator might be curious about the translation of *left-hand side* by *gauche*, which might be more appropriate in the present circumstances than *côté gauche*. When asked for more context, they are given:

radiator left-hand side end tank $\Longleftrightarrow$

boîte à eau gauche du radiateur

In this case, side is not the head noun of the full noun phrase, so the shallow parser did not find a dependency relationship, which was why it only matched to the phrase operator. We should mention that all examples and results presented in this paper are derived from a real technical corpus which could potentially be used in a translation memory system.

## 3 Sentence Chunking

The first step in word and phrase alignment is to divide the sentence into linguistically meaningful sub-units. We focus on the two most important phenomena for terminology extraction, noun phrases and verb sequences. The system follows a series of steps of linguistic text processing. This processing is performed in each language independently, which allows us to take advantage of our existing suite of linguistic tools and enhances

the modularity of the system. The parallel structure of the text is only exploited in the alignment algorithm, described in the next section. The linguistic pre-processing consists of:

(1) Tokenization: divides text into words

(2) Lexicon look-up: identify all potential parts of speech

(3) Part of speech (POS) disambiguation: select most likely POS

(4) Lemmatization: reduce word to its inflectional root

We then perform a text chunking step which consists of dividing the sentence into noun phrases and verb sequences. We follow the common approach used for terminology extraction and describe a chunk by morpho-syntactic pattern matching on sequences of part-of-speech tags. For English, a noun phrase is any consecutive sequence of nouns and adjectives and verb sequence is a consecutive sequence of verbs and adverbs. For French, the definition of noun phrase is slightly more complicated, allowing the prepositions *de* and *à* and an optional determiner (l',le,la,les) to occur between noun/adjective sequences.

These rules are very simple and do not capture all potentially valid phenomena which could lead to interesting terminology units, such as coordination. They represent a reasonable trade-off between over- and under-generation of terminology units. Our approach works much better for English than it does for French, so the system is likely to work better when English serves as the source language for translation. We do some post-processing of the French chunks to identify and dispose of expressions from a stoplist of unrelated initial nouns (e.g. en vue de, au sujet de, ensemble de, etc.). For parsing, we use a shallow parser called IFSP, developed at Xerox by Ait-Moktar and Chanod [5], which generates among other things a set of dependency relations. It is a rule-based system, implemented as a cascade of finite state transducers.

## 4 Statistical Terminology Alignment

Recent research has demonstrated that statistical alignment models can be highly successful at extracting word correspondences from corpora which consist of sentences and their translations (such as translation memories) [1, 4]. These algorithms use global word cooccurrence data to derive links between words at the local sentence level. The TRINITY term alignment system is based on a probabilistic translation model developed by Hiemstra [3]. In this algorithm, each sentence pair is represented by a matrix. The rows represent source language words, the columns represent target language words, and the elements of the matrix are the expected alignment frequencies for the words appearing in the corresponding row and column. The expected alignment frequencies are unknown and must be estimated. There exists a classic algorithm in statistics, known as the Iterative Proportional Fitting Procedure (IPFP), for estimating missing values in a contingency table. By alternatively renormalizing the cell counts by the ratio of observed and estimated row and column sums, the algorithm forces the estimated sums to converge to their observed values. The cell counts after convergence provide reasonable estimates of the expected alignment frequencies.

The IPFP is not quite ideal for estimating expected counts for several reasons. First, it assumes that the row and column sums are the same. This is equivalent to forcing the source and target sentences to be the same length, which is often not the case. Hiemstra addresses this problem by adding a sufficient number of NULL tokens to the shorter sentence to equalize the length. This operation enforces a one-to-one alignment constraint on the word pairs. While the IPFP almost always converges (assuming non-negative ini-

tial cell counts which lead to positive initial row and column sums), there is no unique solution for the cell counts. In fact, there are an infinite number of solutions! For example, any set of values which satisfies the observed term frequency counts converges immediately to a valid solution (e.g. uniform cells counts). Therefore, the solution is highly dependent on the starting values and reasonable initial estimates are required.

The Hiemstra model alternates between computing local expected alignment frequencies for each sentence and summing these frequencies to generate global expected counts. A mathematical description of the algorithm is provided below. The notation $C^i_{n,[jk]} = C^i_n(s_j \Longleftrightarrow t_k)$ refers to the expected count at iteration $i$ of the alignment between source term $s_j$ and target term $t_k$ in sentence $n$, $L(S_n)$ is the length of the longest of the aligned pair of sentences $n$, $L$ is the sum of $L(S_n)$ over all sentences, and $I$ is an indicator function (1 if condition is true, 0 otherwise).

(0) $C^0_{[jk]} = \sum_{n=1}^{N} I(s_j \epsilon S_n) * I(t_k \epsilon S_n)$

(1) For each $s_j$ and $t_k$ in sentence $n = 1 \ldots N$ :

  (a) $C^i_{n,[jk]} = C^i_{[jk]}$

  (b) $C^i_{n,[jk]} \rightarrow$ IPFP $\rightarrow C^{i+1}_{n,[jk]}$

(2) $C^{i+1}_{[jk]} = \sum_{n=1}^{N} C^{i+1}_{n,[jk]}$

(3) If $\sum_{n=1}^{N} |C^{i+1}_{[jk]} - C^i_{[jk]}| < K$ end, else (1)

From the global counts, we can derive bidirectional translation probability estimates in the following manner:

- $P(s_j \rightarrow t_k) = \frac{C(s_j \Longleftrightarrow t_k)}{\sum_m C(s_j \Longleftrightarrow t_m)}$
- $P(t_k \rightarrow s_j) = \frac{C(s_j \Longleftrightarrow t_k)}{\sum_i C(s_i \Longleftrightarrow t_k)}$

Table 4 shows the IPFP at work on a typical sentence pair. Decimal points have been removed to reduce the table size and all cells with value less the 0.05 have been eliminated to make the table more readable. Unlike Hiemstra, we restrict the alignment to content words only (i.e. noun, verb, adjective, adverb). Correct alignments between function words and content words are extremely rare, and this step reduces the noise in the model substantially. The numbers in bold are the largest elements of a row or column and the underlined entries represent the correct alignment as determined by a human assessor. The data presented here show the IPFP prior to convergence. The expected counts correspond relatively well to the correct alignment with a few notable exception. There are several two-to-one and two-to-two alignments which cannot be correctly captured by the IPFP.

Once the Hiemstra algorithm has converged, we can approximate the most likely alignment on a sentence by sentence basis using the simple greedy algorithm given below. It is nearly equivalent to the one suggested by Melamed [4], but is applied to each sentence individually.

(1) Define score:
$S_{jk} = P(s_j \rightarrow t_k) * P(t_k \rightarrow s_j) * P(o_{jk})$

(2) Sort pairs in descending order of their score

(3) Take largest $S_{jk}$ and align $s_j$ and $t_k$

(4) Remove $S_{j.}$ and $S_{.k}$ from the score pool

(5) Repeat steps (3) and (4) until all terms are aligned

As the most likely words are aligned, the parameter space is reduced so there are fewer and fewer possibilities for the terms which are more difficult to align. Because the one to one alignment assumption (reflected by step (4)) limits the generality of the model, we do not actually repeat step (5) until all terms are aligned. Rather we stop when the score reaches a threshold T, as many word pairs with low alignment scores are actually incorrect alignments. The best choice of threshold

T depends on the relative importance of high coverage and high accuracy. The $P(o_{jk})$ term is an offset probability which reflects the relatively likelihood the at term in position j aligns to a term in position k. They are estimated in a fashion similar to the offset probabilities used by Dagan et al. [2] to improve IBM Model 2 [1].

We apply a very similar strategy to align terminology units. Define a score $S'_{jk}$, $S'_{jk} = S_{jk}$ if $s_j$ and $t_k$ are aligned as described above and $S'_{jk} = 0$ otherwise. Let $\sigma_m$ be a phrasal unit in the source language and $\tau_n$ be a phrasal unit in the target language.

(1) Define score $T_{mn} = S(\sigma_m \Longleftrightarrow \tau_n) = \sum_{s_j \epsilon \sigma_m, t_k \epsilon \tau_n} S'(s_j, t_k)$

(2) Sort pairs in descending order of their score

(3) Take largest $T_{mn}$ and align $\sigma_m$ and $\tau_n$

(4) Remove $T_{uv}$ only if both $\sigma_u$ and $\tau_v$ are already aligned

(5) Repeat steps (3) and (4) until all terms are aligned

Note that phrasal units are not forced into a one-to-one alignment. This is important because the sentence chunking is often inconsistent between languages. For example, the French noun phrases that we extract tend to be much longer than the English noun phrases. We do however add the restriction that for phrases which have a one-to-many or many-to-many alignment, all subcomponents must be adjacent, or separated only by function or unaligned content words. This cleaning step eliminates a number of incorrect phrase alignments.

## 5 Commentary

This paper has focused on the application of TRINITY to translation memory systems. It should be stressed that the system also has the ability to automatically construct bilingual terminology lexicons with a relatively high degree of accuracy. The system's biggest weakness is correctly identifying term boundaries, but most of these errors are relatively easy to fix by hand. The existing sentence chunking algorithms are crude and should eventually be augmented by more sophisticated terminology recognition algorithms. In the future, we can imagine aligning more sophisticated term representations such as tree-based structures.

In order to judge the effectiveness of word alignment systems, it is important to perform large-scale quantitative evaluations. The TRINITY system participated in the recent ARCADE evaluation[3]. It aligned 60 highly ambiguous content words with nearly 80% accuracy. Taken in isolation this number may seem low, but the task was extremely difficult. The words sampled for evaluation were among the most ambiguous in their frequency class, since they were also being used for a series of sense disambiguation experiments. For example, the French verb arrêter had 42 occurrences and 18 possible translations (including NULL) after normalizing for inflectional variation. By reducing coverage to only 60% of the test cases, performance could be improved to 90-95% accuracy.

In order to be deployed in an operational system, the language coverage must be substantially increased. The current implementation has only been tested on English and French. However, the TRINITY system is designed so that that the linguistic processing and the statistical alignment components are modular and independent. Xerox already linguistic tools for a large number of European languages. In theory, the statistical alignment algorithm is language independent, although some modifications would be required. For example, the constraint that words must align one-to-one is likely to cause problems in compounding languages like German which tend to conflate entire noun phrases into a single compound. In gen-

[3]For more details on ARCADE, see: http://www.lpl.univ-aix.fr/projects/arcade/index-en.html.

eral, however, we anticipate that it should be relatively easy to extend the TRINITY term alignment system to new language pairs.

## References


[1] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[2] Ido Dagan, Kenneth W. Church, and William A. Gale. Robust Bilingual Word Alignment for Machine-Aided Translation. In *Proc. of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, 1993.

[3] Djoerd Hiemstra. Using Statistical Methods to Create a Bilingual Dictionary. Master's thesis, Universiteit Twente, 1996.

[4] I. Dan Melamed. Word-to-Word Models of Translational Equivalence. Technical Report IRCS Technical Report #98-08, University of Pennsylvania, 1998. Available at: http://www.cis.upenn.edu/ melamed.

[5] Salah Aït-Moktar and Jean-Pierre Chanod. Incremental Finite-State Parsing. In *Proc. of ANLP '97*, 1997.

| Buttons | Text Windows | |
|---|---|---|
| Prev<br>Next | Release the wheel suspension on the left-hand side. | |
| Accept<br>Edit | **Desserrer** l'élément de suspension **gauche**. | |
| Score | Translation Memory | |
| 0.67 | Release / Connect } the wheel suspension on the { left-hand / right-hand } side. | Monter l'élément de suspension droit. |
| 0.44 | Release the { wheel suspension on the left-hand side. / front wheels. | Desserrer les écrous de roues. |
| 0.44 | Release the { wheel suspension on the left-hand side. / front wheels. | Débloquer les écrous de roue avant. |

Table 1: A translation memory example interface

**En:** Connect the wheel suspension on the right-hand side.

**Fr:** Monter l'élément de suspension droit.

| | | |
|---|---|---|
| connect | $\Longleftrightarrow$ | monter |
| wheel | $\Longleftrightarrow$ | élément |
| suspension | $\Longleftrightarrow$ | suspension |
| right-hand | $\Longleftrightarrow$ | droit |
| side | $\Longleftrightarrow$ | NULL |
| wheel suspension | $\Longleftrightarrow$ | élément de suspension |
| right-hand side | $\Longleftrightarrow$ | droit |
| wheel suspension on the right-hand side | $\Longleftrightarrow$ | élément de suspension droit |
| DOBJ(connect,suspension) | $\Longleftrightarrow$ | VMODOBJ(monter,de,suspension) |
| NN(wheel,suspension) | $\Longleftrightarrow$ | NNPREP(élément,de,suspension) |

Table 2: Granularity and type of sub-sentence alignments generated by TRINITY.

**En:** Release the wheel suspension on the left-hand side.

| | | |
|---|---|---|
| (1) | left-hand side | ADJ(left-hand,side) |
| (2) | wheel suspension | NN(wheel,suspension) |
| (3) | Release | DOBJ(release,suspension) |

| | | | |
|---|---|---|---|
| (1) | ADJ(left-hand,side) | → | PADJ(côté,gauche) |
| | PHRASE(left-hand side) | → | gauche |
| (2) | NN(wheel,suspension) | → | NNPREP(élément,de,suspension) |
| (3) | DOBJ(release,suspension) | → | No match |
| | PHRASE(release...suspension) | → | No match |
| | FUZZY(release...wheel) | → | VMODOBJ(desserrer,de,roue) |
| | FUZZY(release...wheel) | → | VMODOBJ(débloquer,de,roue) |
| | release_V | → | défaire, desserrer, relâcher, dégager, déposer, détendre, débloquer, retirer, déverrouiller, rejeter, faire chuter, libérer, désolidariser, désenfiler |

Table 3: Decomposing an example sentence for translation.

| | | Lors de | travaux de | soudage à l' | arc sur un | véhicule, toujours | débrancher le | câblage de l' | alternateur afin d' | éviter de | provoquer une | décharge de | courant | susceptible d' | endommager les | éléments | internes de l' | alternateur. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | - | |
| When electric | 1 | 09 | 08 | 06 | – | – | – | – | – | 07 | **16** | 06 | **17** | – | 05 | 14 | | |
| arc | 1 | – | 28 | **34** | – | – | – | – | – | – | 07 | – | 08 | – | – | 06 | | |
| welding | 1 | 05 | **36** | 24 | – | – | – | – | – | – | 08 | – | 08 | – | – | 07 | | |
| on a vehicle, | 1 | 05 | – | – | **93** | – | – | – | – | – | – | – | – | – | – | – | | |
| always disconnect | 1 | – | – | – | – | **96** | – | – | – | – | – | – | – | – | – | – | | |
| the alternator | 2 | – | – | – | – | – | – | **1.8** | – | – | – | – | 05 | – | – | – | | |
| wiring | 1 | – | – | – | – | – | **81** | – | – | – | – | – | – | – | – | 05 | | |
| to prevent | 1 | – | – | – | – | – | – | – | **68** | __ | – | – | – | 09 | – | – | | |
| the possibility | 1 | **37** | – | – | – | – | – | – | 13 | __ | 09 | – | 09 | – | – | 07 | | |
| of a surge | 1 | 09 | 08 | 07 | – | – | – | – | – | 07 | **16** | 06 | **18** | – | 05 | 14 | | |
| of current | 1 | – | – | – | – | – | – | – | – | – | 05 | **67** | 06 | – | – | 05 | | |
| causing | 1 | – | – | 14 | – | – | – | – | – | **58** | 13 | – | __ | – | – | – | | |
| damage | 1 | – | – | – | – | – | – | – | 16 | – | – | – | – | **77** | – | – | | |
| to the internal | 1 | 07 | 06 | 05 | – | – | – | – | – | 06 | 13 | 05 | 14 | – | – | **32** | | |
| components | 1 | 12 | – | – | – | – | – | – | – | – | – | – | 06 | – | **74** | – | | |
| of the alternator. | - | | | | | | | | | | | | | | | | | |

Table 4: Estimating expected counts via the Iterative Proportional Fitting Algorithm

# MiЯRor: Multimedia Query Processing in Extensible Databases

Arjen P. de Vries
Centre for Telematics and Information Technology
University of Twente, The Netherlands
arjen@cs.utwente.nl

## Abstract

The miЯRor project investigates the implications of multimedia information retrieval on database design. We assume a modern extensible database system with extensions for feature based search techniques. The multimedia query processor has to bridge the gap between the user's high level information need and the search techniques available in the database. We therefore propose an iterative query process using relevance feedback. The query processor identifies which of the available representations are most promising for answering the query. In addition, it combines evidence from different sources. Our multimedia retrieval model is a generalization of a well-known text retrieval model. We discuss our prototype implementation of this model, based on Bayesian reasoning over a concept space of automatically generated clusters. The experimentation platform uses structural object-orientation to model the data and its meta-data flexibly, without compromising efficiency and scalability. We illustrate our approach with some first experiments with text and music retrieval.

**Keywords:** Multimedia Information Retrieval, Digital Libraries, Multimedia Query Processing, Inference Network Retrieval Model

## 1 Introduction

Large archives of digitized multimedia data are set up today, and more and more digitized data will become available online. Digitized multimedia data cannot be searched directly on its binary content. Content-based access to multimedia data therefore requires meta-data about the objects. Meta-data may be manually added descriptions, but can also consist of automatically extracted **features**. Such features are low-level representations of multimedia data, like color distribution and texture [10].

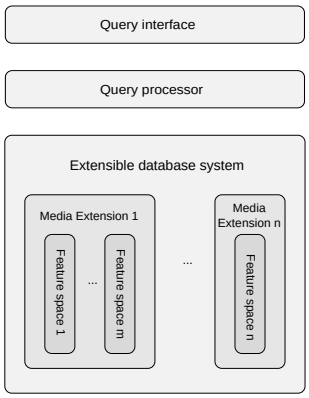



Figure 1: Multimedia database architecture

Traditional database technology, mainly developed for administrative applications, has severe shortcomings with respect to the support of multimedia digital libraries. The access to the digitized multimedia objects, the extraction of meta-data from these objects, and the management of the objects and the meta-data, all have characteristics very different from administrative applications. In the miЯRor project, we study these different requirements on database support, with the purpose to design multimedia database management systems accordingly.

The miЯRor database architecture is especially targeted to support application development in the multimedia digital library environment. It consists of three layers, corresponding to the three light gray boxes in figure 1. At the bottom, we assume an extensible database system, with extension modules (also known as 'data blades' or 'data cartridges') that provide abstract data types (ADTs) encapsulating feature spaces and their distance measures. Our research concentrates on the query processor in the middle box. At the top, we assume a user interface that supports the interaction between the user and the multimedia database.

In [7] and [5], we describe our view on the

bottom layer. We introduce an open distributed architecture for the management of multimedia data and its associated meta-data. Using this architecture, many independent parties can easily cooperate in the construction of a digital library. The extraction of meta-data from the objects in the library is a transparent process and takes place automatically when new data becomes available. A very important aspect of the architecture is modular extensibility. New data formats and new meta-data extraction software can be easily 'plugged in'.

Users typically do not know how to express their information needs in database queries, making the support of multimedia retrieval a tough problem. As we argued in more detail in [8], textual queries cannot capture the full semantics of multimedia data. Content-based retrieval techniques may provide the 'missing' semantics. Querying multimedia data using feature models is performed using example objects; a distance measure between the feature representations of two multimedia objects expresses the similarity between those objects. However, the gap between the meta-data used in the content-based retrieval techniques and the concepts in the users' minds is too big. We term this the **query formulation problem**.

The query processor in the middle layer bridges this gap between user and extensible database system. In the remainder of the paper we focus on its design and implementation. We start with an informal example in section 2, illustrating the query formulation problem in multimedia databases. Next, we introduce in section 3 our approach to multimedia query processing. We discuss the design and implementation of our prototype multimedia database management system in section 4. We are especially concerned with the issues of efficiency and scalability of the architecture. In section 5, we demonstrate the functionality supported in our system with some (small-scale) experiments in text and music retrieval.

## 2 The problem of query formulation

Imagine a journalist writing an article on *the effects of the recent economical crisis in Asia*. Part of the journalist's task is to illustrate the article with photos that hopefully attract readers and increase the sales of the magazine or news paper. A study of journalists at work, reported in [15], made clear that for such 'feature articles', journalists have more freedom than with normal news items. For example, the function of the photo may also be to evoke associations. Also, there is more time to find a 'good' photo.

A journalist usually considers more than one concept for a single illustration task. For the economical crisis example, a possible concept could be a very crowded stock market. Another illustration idea is a photo demonstrating that normal people do not have much money left to spend, for example by showing an empty shopping street in otherwise crowded Hong Kong. In both cases, a photo expressing despair or panic is probably preferred over photos without explicit emotions. Furthermore, constraints like overall page layout may affect the choices made while performing the illustration task.

Assume now that the journalist has access to a video archive of news bulletins originating from various broadcasters. In the archive, the time, date, and source are maintained for each news bulletin. The video data itself is modelled with a sequence of key-frames, and a text version of the audio track. The content of the key-frames is indexed using color and texture features. For comparison, a news archive storing similar meta-data is described in [13].

Searching for 'stock market' in the subtitles may be rather succesful as an initial query. The precision of the results is probably high, meaning that most key-frames with matching subtitles really show stock market scenes. However, the recall may be low: many scenes at stock markets may not have been labelled with an explicit annotation mentioning 'stock market'. Note that this problem will be much worse for the second illustration idea, using 'Hong Kong shopping street' as a text query.

Emotional aspects of the images searched are especially hard to capture in a textual query. Searching for 'despair' in subtitles will probably not retrieve many useful results. These aspects of the illustration task may be captured more easily in terms of feature representations of the images. However, the journalist cannot possibly be expected to express a high level concept like 'despair' in a combination of color and texture features. Conversely, the internal representation of the video with its meta-data should be completely invisible to the users.

## 3 Multimedia query processing

### 3.1 Desiderata

The query formulation problem leads to a different view on query processing than common in the database community. Instead of a one step process with a single query, and the database simply retrieving its matching objects, the interaction between a multimedia database and the user should be a *dialogue*. The query processor should iteratively interpret the user's judgements on the results of the previous step, and adapt the initial query such that it will better reflect the observed but unknown information need. It derives database queries against the meta-data, using information from the interaction with the user.

An iterative approach to query processing is already common in information retrieval (IR) systems [24]. We therefore base the miЯRor query processor on the theory and techniques developed in the IR research field [6]. However, a multimedia database management system differs significantly from a special purpose text retrieval system. The management of multimedia data requires extensible systems [7, 5]. IR systems are not designed for extensibility. The implementations assume detailed knowledge about the structure of the indexed documents and the meta-data that models the content. In an extensible system however, we do not know beforehand what representations of the multimedia objects will be available as meta-data at run-time.

A somewhat related difference between IR and multimedia databases is the number of sources of evidence used in the retrieval process. In IR, only a small number of different sources is considered, e.g. abstract, full text, citations, and maybe hypertext links. On the other hand, the combination of evidence from *many* different sources is crucial for multimedia retrieval. Experiments with the Foureyes learning agent for the Photobook image retrieval system demonstrated the advantages of a collection of data-dependent and task-dependent feature spaces over a universal similarity measure defined on a generic feature space [16, 17]. Different feature spaces capture different aspects of the data. Typically, a feature space performs only well at a small set of tasks, on only a subset of the data. Rather than a carefully selected 'society' of models as envisioned in Foureyes, 'anarchy' seems however a more appropriate metaphore in our context; indeed, in miЯRor the collection of feature spaces changes dynamically as new meta-data extraction software is added or removed.

### 3.2 Retrieval model

Figure 2 proposes the design of the miЯRor query processor. An IR system is described by its **retrieval model**, which defines the document representation, the query formulation, and the ranking function [26]. These three aspects are reflected in the design of our multimedia query processor, in subsequently the **concept layer** (document representation), the **evidential reasoning layer** (ranking function), and the **relevance feedback layer** (query formulation).

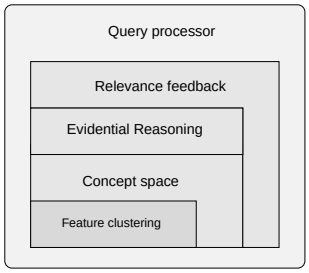



Figure 2: The multimedia query processor

#### 3.2.1 Concept layer

The concept layer defines the basic units representing the content of the multimedia objects. IR literature usually refers to these units as the **indexing features**; to avoid confusion with the features used in content-based multimedia retrieval, we prefer to call these **concepts**. The concepts are input to the evidential reasoning layer, which selects the objects in the database that best match the user's query.

Most IR systems use the words occuring in the document as concepts. In text documents, words naturally refer to classes of objects in the real world. For example, the word 'street' occuring in an English text is the same, whether that particular street is located in Cambridge or Oxford. Sometimes, words occuring in the text are first clustered, using **stemming** algorithms and **thesauri**. This may alleviate the problems with ambiguity in natural language.

In multimedia retrieval, the content representation of objects is a (usually unique) point in multi-dimensional feature space. Therefore, an important task of the concept layer is **feature clustering**. The feature representation of a

street in Cambridge will be different from the representation of a similar street in Oxford. To complicate matters, the representation of one and the same street in two different images will usually be different as well. Hence, before we can develop a theory for multimedia retrieval similar to the retrieval models in IR, we have to cluster these points, based on their relative positions in feature space.

The concept layer uses unsupervised clustering algorithms to identify clusters in feature space. Of course, we realize that not no algorithm will automatically cluster all streets in a single concept. Nor do we expect to construct concepts that only occur in a subset of the streets but in no other classes of objects. However, the assumption underlying the content-based retrieval techniques is that proximity between points in feature space corresponds to some sort of similarity in the real world. Thus, the proximity of the clusters' feature points is likely to reveal an implicit underlying concept that captures some of the semantics of the objects.

#### 3.2.2 Evidential reasoning layer

The responsibility of the evidential reasoning layer is to identify the multimedia objects in the database that may fulfill the user's information need as expressed in the query. The evidence is based on the presence or absence of concepts, very similar to traditional IR. The evidential reasoning process combines the evidence from different sources into a single judgement. It should take into account the structural composition of objects from their component objects. We discuss the evidential reasoning layer in more detail in section 3.3.

#### 3.2.3 Relevance feedback layer

The relevance feedback layer has two tasks. First, it is responsible for query (re-)formulation. It controls the dialogue between the user and database, analyzing the user's feedback information and changing the query such that it (hopefully) better reflects the user's information need. We term this online processing **query-space modification**. Second, the relevance feedback layer maintains a history for offline processing, logging the interaction between users and database. Supervised clustering techniques may use these logs to improve the initial clustering constructed in the concept layer. Also, statistical tests may identify dependencies between feature spaces. We refer to this task as **object-space modification**. Although we regard both types of feedback as important, we currently focus on query-space modification.

### 3.3 Reasoning layer

The 'probability ranking principle' states that an object ranking is optimal when the objects are ranked by their probability of relevance to the user [24, p. 113]. Many competing IR theories can be used to estimate these probabilities. We base our theory for multimedia retrieval on the inference network retrieval model, introduced by Turtle and Croft [22, 23]. It has been shown that this probabilistic model can also express other common retrieval models, such as the Boolean and the vector space model. The model is based on the theory of Bayesian belief networks. A Bayesian belief network is a graph representation of probabilistic knowledge. In a belief network, nodes represent random variables, and arcs reflect relationships between the linked variables. The direction of an arc between parent node and child node represents causality. The strength of this causal influence is expressed by a conditional probability. A belief network encodes a joint probability distribution. The advantage of the network representation of this distribution is that inference procedures exist to compute the value of any conditional probability in the network given the available evidence, without having to derive a closed form formula for the complete distribution. The reader is referred to [19] for more details.

Turtle and Croft claim advantages of their model over different retrieval models because of its theoretical foundation in Bayesian belief networks. Unfortunately, due to the simplifications made to the inference procedure and the network structure (trading mathematical correctness for efficiency), it is hardly possible to take advantage of theoretical developments in the more general theory of Bayesian networks. Nevertheless, we take this model as a starting point for the development of a theory for multimedia information retrieval [6]. It is very suited for our purpose, since it has been introduced in IR to combine evidence from different sources more easily. Also, it has a modular structure that reflects the architecture's extensibility. Most importantly, its implementation in the InQuery retrieval system has been very succesful in many IR evaluation experiments.

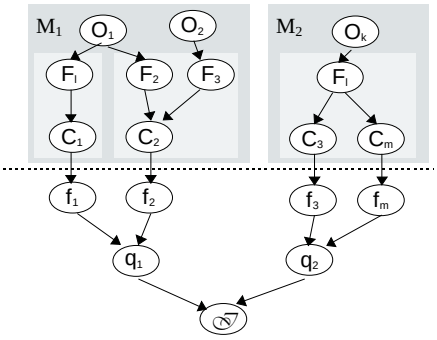



Figure 3: A multimedia retrieval model based on Bayesian inference networks

### 3.3.1 The network structure

Using figure 3, we first explain the general idea behind miЯRor's version of the inference network retrieval model. Each base type, e.g. image or audio, has its own media extension $\mathcal{M}_i$. A media extension, depicted as a dark gray box in the figure, manages a collection of content representations $\mathcal{F}_j$, shown as light gray boxes. The nodes in the network represent binary random variables. The top part of the network is called the **object network** and is static for a given data collection. The bottom part, the **query network**, is dynamically created by the relevance feedback layer, based on interaction with the user.

At the roots of the network, we find the object nodes $O_i$. For now, we will ignore the internal structure of the multimedia objects; all objects are considered **atomic**. In section 5.2, we discuss the retrieval of compound documents to illustrate a possible approach to modelling structured objects. The objects $O_i$ are connected to their meta-data representations of content $F_j$. The concept nodes $C_p$ represent the concepts identified by clustering in the concept layer. The model allows concept clusters to overlap. Thus, a single representation node may be connected to several concept nodes. Node $\mathcal{I}$ in the query network represents the user's information need. The information need is expressed by the example objects provided by the user in the interaction process. The query nodes $q_i$ model these example objects. The meta-data extracted from these objects is represented by the $f_j$ nodes. These nodes are connected to their corresponding concept nodes in the static object network. In the dialogue between database and user, the relevance feedback layer adapts the structure of the query network by adding or removing nodes.

Let us take a closer look at the example instantiation of the network model given in figure 3. Assume that $\mathcal{M}_1$ is an image media extension. It manages feature spaces $\mathcal{F}_1$ for color and $\mathcal{F}_2$ for texture. Image object $O_1$ has a color feature $F_1$ and a texture feature $F_2$. Color feature $F_1$ has been clustered into concept $C_1$, and texture features $F_2$ and $F_3$ into concept $C_2$. Color representation $f_1$ and texture representation $f_2$, extracted from the example image $q_1$, are part of the same clusters in feature space, hence also connected to $C_1$ and $C_2$ respectively.

### 3.3.2 Ranking objects

The inference network is used to compute $\Pr(\mathcal{I}|O_i)$, which corresponds to the chance that the information need as expressed in the query network is fulfilled when presenting this object to the user. The random variables associated to the objects and their meta-data represent observations. In the ranking process, each object $O_i$ is considered in isolation: its node is set to true, and all other nodes to false. This evidence is propagated through the network until it reaches $\mathcal{I}$, when we have computed the desired $\Pr(\mathcal{I}|O_i)$.

The joint probability distribution encoded in the object network is independent of the query. In our current model, observing $O_i$ always implies observing its meta-data $F_j$. We assume the feature spaces independent and equally important. In later revisions of the retrieval model, we may use the conditional probability distribution $\Pr(F_j|O_i)$ to represent knowledge about how reliably each feature space describes an object. $\Pr(C_p|F_j)$ expresses the belief that concept $C_p$ is observed when we observe feature $F_j$. This probability should be estimated in the feature clustering process. Similarly, $\Pr(f_j|C_p)$, specified at the arcs connecting the object network with the query network, describes our belief that feature $f_j$ in query space is described by the concept $C_p$ in object space.

Instead of first computing these probabilities independently, and then propagating the belief to the nodes $f_j$ in the query network, the implementation of the inference network retrieval model computes $\Pr(f_j|O_i)$ directly. In InQuery, this probability is estimated using term frequency $tf$, inverse document frequency $idf$, and default belief $\alpha$:

$$\Pr(f_j|O_i) = \alpha + (1-\alpha) \cdot tf \cdot idf \qquad (1)$$

In a multimedia feature space, we have to define a procedure to estimate this probability using the

relative position of that point in a cluster and the distribution of other points in the cluster. An unsupervised clustering algorithm like AutoClass provides such an estimate [4]. As an alternative, we plan to investigate the cluster-based probability model that has been proposed in [20].

### 3.3.3 Propagation of evidence

To explain the propagation of evidence from the $f_i$ through the query network to $\mathcal{I}$, we introduce a formal description of the inference network adapted from [21]. Let $x_i$ be a node in a Bayesian network $G$, and $\Gamma_{x_i}$ be the set of parents of this node. Since $G$ is a Bayesian belief network, the influence of $\Gamma_{x_i}$ on $x_i$ is specified by the conditional probability distribution $\Pr(x_i|\Gamma_{x_i})$. Let the cardinality of $\Gamma_{x_i}$ be $n$, and the random variables be binary like in our retrieval model. Then we have to specify $2^n$ different probabilities to describe this conditional distribution. Obviously, this is problematic for the computational tractability of the inference. Therefore, we have to find an approximation of the real probability table (also known as link matrix).

Note that, for a node $x_i$, the influence of $\Gamma_{x_i}$ on $x_i$ can be specified by *any* function $F(x_i, \Gamma_{x_i})$ that satisfies:

$$\sum_{y \in \mathrm{Y}} F(y) = 1 \tag{2}$$

$$0 \leq \mathrm{F}(y) \leq 1 \tag{3}$$

where Y is defined as $x_i \times \Gamma_{x_i}$. In the general theory of belief networks, functions approximating $\Pr(x_i|\Gamma_{x_i})$ have been used to model **causal independence** efficiently: the case when multiple causes contribute independently to a common effect. A famous example is the 'noisy-or' model [19]. In his thesis, Turtle gives closed-form expressions for a limited subclass of functions $F(x_i, \Gamma_{x_i})$, that are useful in IR and can be evaluated in $\mathcal{O}(n)$. Greiff gives a larger class of functions, described by so-called PIC-matrices, for which the evaluation depends on the number of parents that are true but not on their ordering [12]. He first provides an evaluation procedure in $\mathcal{O}(n^2)$, and then gives an algorithm in $\mathcal{O}(n)$ for a subclass of these PIC-matrices. Functions in these classes are 'sum', probabilistic versions of logical operators 'and' and 'or', as well as variations of these usually referred to as **'pnorm'-operators**. These functions are all part of InQuery's language to describe the structure of the query network.

Of course, an approximation of $\Pr(x_i|\Gamma_{x_i})$ with a different function $F(x_i, \Gamma_{x_i})$ is only semantically valid if this function behaves similar to the true probability distribution. The succes of the retrieval system InQuery, that is based on the inference network retrieval model, is often given as 'proof' that these functions really model the true probabilistic dependencies between for example the concepts and the document's relevance. We do not agree with this line of reasoning. The experiments with InQuery demonstrate *only* that the computed value for $\Pr(\mathcal{I}|\mathrm{O}_i)$ may be interpreted as a good approximation of the probability of relevance of the $\mathrm{O}_i$. The distribution captured by the complete network apparently reflects its desired interpretation in the real world. However, we should not deduce that the probability estimates for the nodes $x_i$ and their parents also have an interpretation regardless of the choice of $F(x_i, \Gamma_{x_i})$. This observation is confirmed by the difficulties with chosing an optimal value for default belief $\alpha$ (cf. equation 1) in the experiments with 'pnorm'-operators reported in [12]. Despite of these limitations, the inference network retrieval model is a very powerful model because of its ability to flexibly model varying approaches to the combination of evidence from different representations. Also, the original Bayesian belief network underlying the retrieval model, without its approximations used to achieve tractability, can still be used as a reference when we want to understand why some operator combined with some formula estimating the concept probabilities does or does not work well.

## 4 Design and implementation

The implementation of miЯRor's multimedia query processor requires the integration of IR and databases. Integration of IR and databases has historically led to impractically slow systems; the efficient execution of IR techniques required special purpose software systems. We believe that IR and databases *can* be integrated in a single system, but only if this integration is complete, and neither a layer on top of, nor a black box inside a database system. Therefore, our prototype implementation is based on **structural object-orientation**. A detailed discussion of the benefits of structural object-orientation for IR processing in a database system can be found in [9].

Figure 4 shows the design of our research prototype. The design is focused on the development of a system that will scale up to very large

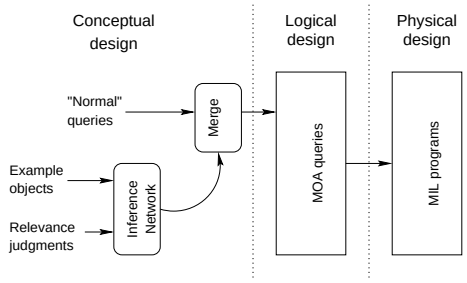



Figure 4: Design

data collections. Its main characteristic is the strict separation between the logical and physical databases. This separation provides data independence, and allows for algebraic query optimization in the translation from expressions at the logical level to queries executed in the physical database. Also, parallellization of the physical algebra is orthogonal to the logical algebra, such that we can transparently distribute the data over different database servers by changing only the mapping between the two views. In this paper, we only discuss query processing at the logical level. The interested reader is referred to [9] for a discussion of the implementation in the physical database.

**MOA** is an object algebra for the logical level, being developed by our research group. It provides an extensible nested object data model and an algebra on this model. The prototype implementation does not yet provide a query language at the conceptual level; queries can only be specified using MOA expressions. The MOA Tools translate the query expressions specified in MOA into efficient MIL programs[1] that are executed in the **Monet** database system [1]. Monet is an extensible parallel database kernel that is intended to serve as a backend in various application domains [2]; e.g., image retrieval is supported by an extension module defining the 'Acoi' algebra [18]. Monet has also been used succesfully for geographic information systems as well as commercial data mining applications.

MOA's data model is based on **base types** and **structuring primitives**. Base types are ADT-style types. They are inherited from the physical database schema, including common types such as `int` and `str`, but also large object types like `Image`. A structuring primitive combines known types to create a **structured type**. Common examples in object-oriented data models are bag, set, and tuple. To demonstrate the specification of multimedia data collections in MOA, we give in example 1 the definition of a structured data type for the video archive mentioned in section 2.

[1]MIL stands for Monet Interface Language

**Example 1**

```
BAG<
  TUPLE<
    time:       Atomic<Time>,
    date:       Atomic<Date>,
    keyframes:  LIST<
                  Atomic<Image>
                >,
    audiotrack: Atomic<Audio>
    transcript: Atomic<Text>
  >
>;
```

In the implementation of the query processor, we perform the evidential reasoning process as database queries. For this purpose, we extend MOA with structures for components of the inference network. Operations on these structures model the propagation of beliefs within a component. The resulting language allows us to specify many different network topologies, by simply choosing varying operators to combine different sources of evidence. The relevance feedback layer can thus adapt the network structure by simply generating different MOA expressions.

For the integration of content-based querying in MOA, we first define a structure that encapsulates the object network. The `CONTREP` structure is defined as the content representation of object $O_i$ in feature space $\mathcal{F}$. If an object has meta-data representations in several feature spaces, then each combination of object and feature space is modelled in a distinct instantiation of this structure. The concept layer constructs a `CONTREP` from the output of the feature clustering process. Recall that the $Pr(f_j|O_i)$ are estimated directly from the statistical distribution of occurrences of $C_p$ in $O_i$ and in the collection. Therefore, we can sufficiently describe the object network for $O_i$ by the $C_p$ present in the object. Thus, a `CONTREP` stores the connections from node $O_i$ to its associated nodes $C_p$ in $\mathcal{F}$. In the current prototype, the clustering of a set of features is performed outside the database, and the `CONTREP` structures are bulk-loaded from files describing the identified concepts.

We also extend MOA with two other structures, that allow us to specify the propagation of evidence through the query network. The `INFNET` structure models a node $x_i$ with its parents $\Gamma_{x_i}$.

It can be constructed from a set of probabilities, in which each value corresponds to the belief in a node of $\Gamma_{x_i}$. The structure defines operators for the class of functions $F(x_i, \Gamma_{x_i})$ that is expressed by PIC-matrices [12]. `DOCNET` is a specialization of `INFNET` that is optimized for the assignment of default beliefs $\alpha$ to nodes that do not occur in the content representation of an object.

The three structure extensions interact as follows in the computation of $\Pr(\mathcal{I}|\mathrm{O}_i)$. The relevance feedback layer constructs a query network, based on the example objects provided by the user. In the first step of belief computation, `CONTREP`'s operation `getBL` connects the query network to the object network. Its operands are the $\mathrm{f}_j$ nodes of the same feature space as the `CONTREP`, and a structure representing global statistics of the feature space. This operation computes estimates of $\Pr(\mathrm{f}_j|\mathrm{O}_i)$, returning a `DOCNET` structure capturing the instantiation of the nodes at the top level of the query network. A belief operator $F(q_i, \Gamma_{q_i})$ then computes an estimate of $\Pr(q_i|\Gamma_{q_i})$. Next, we repeat constructing an `INFNET` from these estimated probabilities, and computing the belief in the nodes at the next level of the query network, until we reach node $\mathcal{I}$. We then have computed $\Pr(\mathcal{I}|\mathrm{O}_i)$ using the joint probability distribution described by the inference network.

**Example 2**

```
BAG<
  TUPLE<
    time: Atomic<Time>,
    date: Atomic<Date>,
    keyframes:
      LIST<
        TUPLE<
          keyframe: Atomic<Image>,
          color:    CONTREP,
          texture:  CONTREP
        >
      >,
    audiotrack: Atomic<Audio>,
    transcript:
      TUPLE<
        transcript: Atomic<Text>,
        content:    CONTREP
      >
  >
>;
```

In combination with standard MOA structures like bag and tuple, we can now define and manipulate multimedia data collections and their metadata. For each feature space modelling the content of a multimedia object, we define a `CONTREP` structure. Since this structure is an orthogonal extension of MOA, we can also query the collection on the combination of content with conventional attributes. For example, we can easily restrict the query results of a content query to a ranking of only last week's news bulletins. Example 2 extends the type definition for the video archive example with its content representations. Of course, the content representations may be hidden from end users, such that they only see the definition of example 1.

# 5 Examples

## 5.1 Text retrieval

We first implemented a simplified version of the original inference network retrieval model, leaving out its proximity operators. Assume now that `docs` is a bag of content representations of text documents, `query` is a collection of query terms, and `stats` provides collection statistics such as *idf*. The MOA expression in example 3 computes $\Pr(\mathcal{I}|\mathrm{O}_i)$ as described in the previous section. A `map` on a bag performs an operation on all elements of the bag. In the specification of the operation to be performed, the bag's element is referred to as `THIS`. The `getBL` constructs a `DOCNET`, such that the inner `map` converts the bag of document representations in a bag of `DOCNET` structures. The outer map uses the 'sum' belief operator to compute the probability of relevance for each document.

**Example 3**

```
map[sum(THIS)](
  map[getBL( THIS,
    query, stats ) ]( docs ));
```

## 5.2 Compound documents

In the discussion of our retrieval model so far, the objects $\mathrm{O}_i$ have been assumed atomic. We will now rank compound documents on logical units like sections or chapters, rather than on their full content. In example 4, we model the content of a news document as a bag of items. The topology of the inference network specified by this particular query is taken from [3]. These experiments suggested that the best results are achieved when

a document is ranked by the contribution of its best section. Note the use of the `INFNET` structure to express the belief propagation through the extra layer of nodes in the query network.

**Example 4**

- *data definition for compound documents:*

```
BAG<
  TUPLE<
    Category : str,
    Content  : BAG< CONTREP >
  >
>;
```

- *ranking news documents by their best items:*

```
map[max( INFNET<THIS> ) ](
  map[ map[ sum(getBL( THIS,
         query, stats ))](
    THIS.Content ) ]( docs )));
```

### 5.3 Music retrieval

We conclude the paper with a small scale multimedia retrieval experiment using our experimentation platform. The results should not be given more status than just 'proof of concept'. Although the experimental evaluation has not been very thorough, the results are encouraging. Indeed, it seems possible to interactively retrieve groups of similar songs, in particular for well defined categories.

In multimedia retrieval, emotional and aesthetic values play an important role in the user's evaluation process [5]. Because subjective judgments seem especially important when we compare music fragments, we decided to try out the multimedia query processor on a content representation of music objects. Note that we assume the similarity between two fragments to be defined by the overall 'sound' of the music. The extraction of meta-data is based on [25]. We augmented the feature vectors with a simple rhythm indicator based on peaks in the autocorrelation function of the lowest parts of the frequency domain.

Data set **Symbol-1**, created in cooperation with the Dutch company 'Symbol Automatisering', consists of 287 songs. Domain experts of Symbol Automatisering have manually classified these songs into six main categories: rock, house, alternative, easy listening, dance, and classical. We sampled between one and two minutes of each song, that we segmented into fragments of 5 seconds each. The result is a data collection of 3363 fragments for which we computed the feature vectors. Feature clustering with Autoclass identified 53 different clusters; we assigned to each feature vector the concept node according to the cluster with the highest probability. We then modeled a song as a collection of these concepts. We treated this representation of songs as if they were text documents in which the concepts are the words. Thus, we simply used equation 1 to estimate $\Pr(f_j|O_i)$. In future experiments, we plan to evaluate the representation of songs in more detail, e.g. using the $\Pr(f_j|C_p)$ estimated by Autoclass, and using all concepts detected in the fragment.

We performed the following experiment with music retrieval from this collection. Simulating online relevance feedback, we constructed a query network of the concepts that occurred most frequently in half of the songs belonging to a category. We then tried to retrieve other songs of the same category. Of the top 20 songs for the query based on 'rock', 15 had also been classified manually as rock. Of the other 5 songs, only 2 clearly do not belong in the rock category. With the 'classical' and 'house' songs, we found hardly any misses. Results for the category 'alternative' were however hardly better than chance, but maybe this is partly because the category is not well defined.

## 6 Conclusion and future research

We developed a multimedia query processor that supports the end users of a multimedia database with query formulation. The architecture is extensible with new algorithms for meta-data extraction, and the query processor is designed to use the available representations transparently. The integration of the content-based query processing in MOA also allows the user to query both the logical and the content structure of multimedia objects. The main contribution of our work is the design for scalability.

Improving the basic functionality of the prototype is a topic high on our research agenda. From a technical viewpoint, we should implement clustering in our architecture. Also, we want to experiment with multiple representations in the database. The foundation of the model in the theory of probabilistic networks provides a strong theoretical framework [19, 11, 14]. Within this

framework, there is a lot of scope for experiments and we would like to investigate its use to model and learn dependencies between representations.

An important but open research issue is the development of an evaluation methodology for multimedia retrieval. The inherent subjectivity in multimedia searching makes it impossible to develop a test suite that is not related to a real user task. We believe the music domain provides a context well suited to evaluate how the query process adapts to subjectivity of the users. However, content modeling of music is not easy and the success criteria are vaguely defined. To evaluate the effect of multiple representations and their interdependencies in retrieval, retrieval from publishers' photo and video archives may provide a better context. However, the challenge in this domain is to construct a test suite with realistic user tasks and clearly defined success factors, without making the evaluation process too expensive (amount of data) and elaborate (user studies).

## Acknowledgements

Many thanks go to Annita Wilschut, who encouraged me to prototype my view on a multimedia DBMS using structural object-orientation. Dick Theissens of Symbol Automatisering generously provided data for the music retrieval experiment, and Harold Oortwijn implemented the feature extractor. I would also like to thank Henk Ernst Blok for proofreading draft versions of this paper and his useful feedback.

## References

[1] P. Boncz, A.N. Wilschut, and M.L. Kersten. Flattening an object algebra to provide performance. In *Fourteenth International Conference on Data Engineering*, pages 568–577, Orlando, Florida, February 1998.

[2] P.A. Boncz and M.L. Kersten. Monet: An impressionist sketch of an advanced database system. In *BIWIT'95: Basque international workshop on information technology*, July 1995.

[3] J.P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, July 1994.

[4] P. Cheeseman and J. Stutz. Bayesian classification (AutoClass): Theory and results. In *Advances in Knowledge Discovery and Data Mining*. AAAI Press, 1995.

[5] A.P. de Vries and H.M. Blanken. Database technology and the management of multimedia data in Mirror. In *Multimedia Storage and Archiving Systems III*, volume 3527 of *Proceedings of SPIE*, Boston MA, November 1998.

[6] A.P. de Vries and H.M. Blanken. The relationship between IR and multimedia databases. In *IRSG'98*, Autrans, France, March 1998.

[7] A.P. de Vries, B. Eberman, and D.E. Kovalcin. The design and implementation of an infrastructure for multimedia digital libraries. In *Proceedings of the 1998 International Database Engineering & Applications Symposium*, pages 103–110, Cardiff, UK, July 1998.

[8] A.P. de Vries, G.C. van der Veer, and H.M. Blanken. Let's talk about it: Dialogues with multimedia databases. Database support for human activity. *Displays*, 18(4):215–220, 1998.

[9] A.P. de Vries and A.N. Wilschut. On the integration of IR and databases. In *IFIP WG 2.6 Working Conference on Database Semantics - Semantic Issues in Multimedia (DS-8)*, Rotorua, New Zealand, January 1999. Accepted as short paper.

[10] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 3:231–262, 1994.

[11] R.M. Fung and B.A. Del Favero. Applying Bayesian networks to information retrieval. *Communications of the ACM*, 38(3):43–48, March 1995.

[12] W. Greiff, W.B. Croft, and H. Turtle. PIC matrices: A computationally tractable class of probabilistic query operators. Technical Report IR-132, The Center for Intelligent Information Retrieval, 1998. submitted to ACM TOIS.

[13] A.G. Hauptmann and M.J. Witbrock. *Intelligent multimedia information retrieval*,

chapter Informedia: news-on-demand multimedia information acquisition and retrieval, pages 215–239. AAAI Press/MIT Press, 1997.

[14] D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced technology division, March 1995. Revised edition November 1996.

[15] M. Markkula and E. Sormunen. Searching for photos - journalists' practices in pictorial IR. In *The challenge of image retrieval*, Newcastle upon Tyne, UK, 1998. University of Northumbria.

[16] T. Minka. An image database browser that learns from user interaction. Master's thesis, MIT, 1996. Also appeared as MIT Media Laboratory technical report 365.

[17] T.P. Minka and R.W. Picard. Interactive learning using a "society of models". Technical Report TR-349, MIT Media Laboratory Perceptual Computing Section, 1997. Submitted to Special Issue of Pattern Recognition on Image Databases: Classification and Retrieval.

[18] N. Nes and M. Kersten. The Acoi algebra: A query algebra for image retrieval systems. In *Advances in Databases. 16th British National Conference on Databases, BNCOD 16*, pages 77–88, Cardiff, Wales, UK, July 1998.

[19] J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of Plausible Inference*. Morgan Kaufmann, California, 1988.

[20] K. Popat and R.W. Picard. Cluster-based probability model and its application to image and texture processing. *IEEE Transactions on Image Processing*, 6(2):268–284, February 1997.

[21] B.A.N. Ribeiro and R. Muntz. A belief network model for IR. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR '96)*, pages 253–260, Zürich, Switzerland, August 1996.

[22] H.R. Turtle. *Inference networks for document retrieval*. PhD thesis, Univeristy of Massachusetts, 1991.

[23] H.R. Turtle and W.B. Croft. A comparison of text retrieval models. *The computer journal*, 35(3):279–290, 1992.

[24] C.J. van Rijsbergen. *Information retrieval*. Butterworths, London, 2nd edition, 1979.

[25] E. Wold, Th. Blum, D. Keisler, and J. Wheaton. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3(3), 1996.

[26] S.K.M. Wong and Y.Y. Yao. On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1):38–68, January 1995.

# An Overview of Information Extraction Technology and its Application to Information Retrieval

Douglas E. Appelt
Artificial Intelligence Center
SRI International
Appelt@ai.sri.com

## ABSTRACT

Information Extraction Technology refers to a collection of shallow natural-language processing techniques that are particularly well suited to extracting specific, targeted information from texts. Information Retrieval Technology is generally thought of as techniques for retrieving a set of documents maximally relevant to a query from a large set of documents. The two technologies, while related from a user's perspective, have generally been thought of as suitable for addressing completely different problems.

This article reviews information extraction technology, and considers how it can be applied to improving the precision of routing queries. An experiment in which the SRI FASTUS system is applied to a routing task in TREC-6 is discussed, as well as an Open Domain System that is being developed to make such undertakings easier and quicker in the future.

**Keywords:** Information Extraction, Information Retrieval, maximum entropy, routing

## 1 INTRODUCTION

Information Extraction is a general name for a variety of shallow natural-language processing techniques that are used to extract highly specific, targeted information from texts. The following characteristics are typical of problems for which current information extraction technology is an appropriate solution:

1) The relevant corpus of texts is moderately large – on the order of hundreds, or even thousands of texts.
2) The proportion of relevant text to irrelevant text is small. A typical document would have a handful of relevant sentences.
3) A small set of targeted entities is relevant to the task. Persons, companies, organizations and locations are entities relevant to extraction tasks in most domains.
4) A small set of targeted relations and events involving entities is sought.

There are several different information retrieval tasks. Probably the task which people are most familiar is the *ad-hoc query*, which is typically what one does with a web search engine. One formulates a query describing documents in which one is interested, and the retrieval system returns a list of maximally relevant documents, sorted in order of decreasing relevance, from a pre-existing database of documents.

Another information retrieval task of interest is the *routing* task. In this task, one has a standing information need for documents on a selected topic. However, instead of retrieving the documents from a pre-existing collection, new documents arrive, and they are then sorted in decreasing order of relevance to the topic.

There is an obvious connection between extraction and retrieval, in that it is possible to formulate a retrieval task as an extraction task. A document is deemed relevant for retrieval if it is possible to extract information relevant to the topic from it.

## 2 INFORMATION EXTRACTION

Information Extraction systems are characteristically shallow language processing systems. The shallowness of the processing is driven by the need to process large amounts of data in reason-

able periods of time, and to be robust against the many kinds of errors and disfluencies that one encounters in real-world texts. While information extraction systems differ in interesting details, at a high level of abstraction they can be viewed as a cascade of transducers. The lowest level transducer accepts the raw text as input, and each level imposes some additional structure on its input and passes the results along to the next phase. The phases themselves can be implemented as statistical annotators, or as in the case of the SRI FASTUS system [1] as finite-state transducers. While it is conceivable that more complex components, such as full parsers, could be used in information extraction systems, the necessity of processing large quantities of real-world data in reasonable time has strongly favored the simpler finite-state models.

The FASTUS system is a typical information extraction system that comprises six levels of transducers:

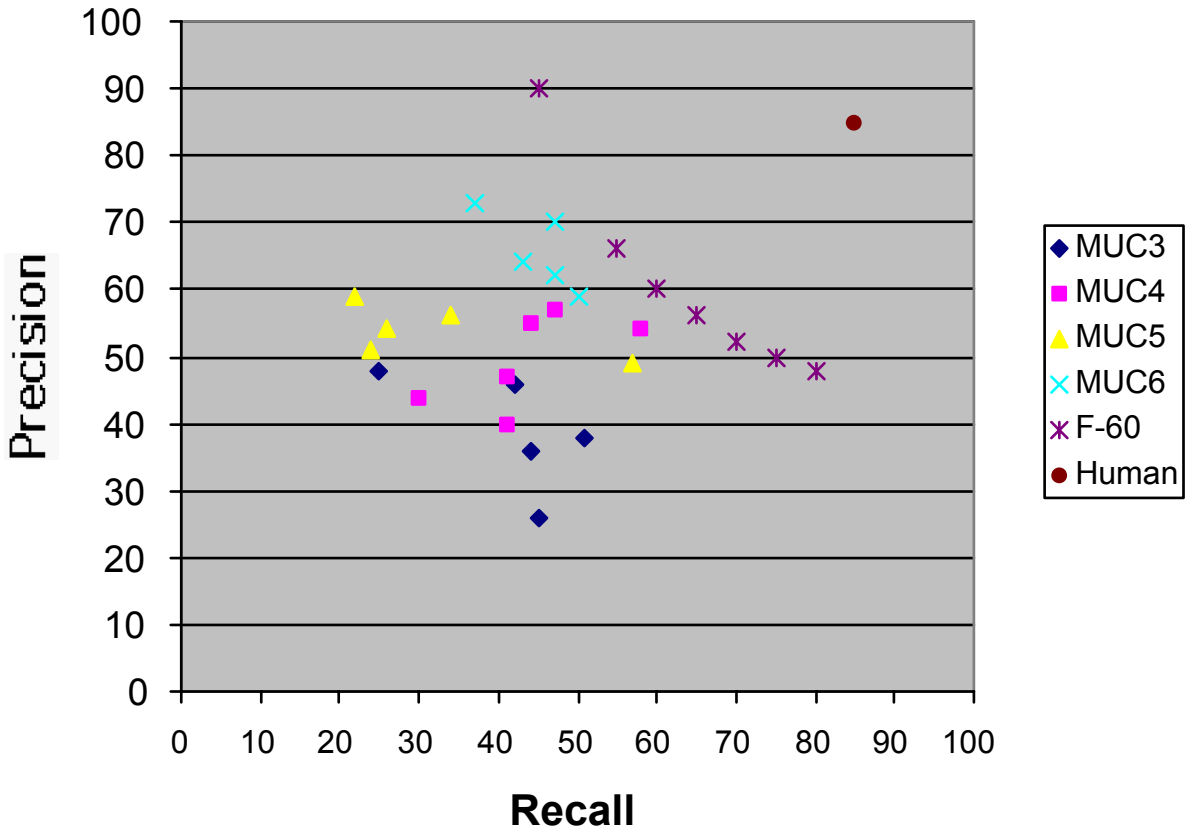
1) **Tokenization –** text is broken down into basic units (i.e. words, punctuation, SGML markup, etc) and words looked up in the lexicon, and annotated with possible features.
2) **Preprocessing** – certain highly regular and local constituents are identified and annotated. This includes date and time expressions, as well as spelled-out numbers.
3) **Named Entity Recognition** - this phase uses lexical information and internal structure, as well as capitalization, if available, to mark entities in the text with proper names. The basic FASTUS system includes name recognition rules for persons, locations, companies, and other organizations.
4) **Simple Phrase Recognition** – this phase produces a shallow finite-state parse of the input to annotate noun groups (including determiner, pre-nominal modifiers and head noun), verb groups (auxiliaries, adverbs, and main verb) and "particles" (prepositions, conjunctions, relative pronouns).
5) **Complex Phrase Recognition** – simple phrases are combined into complex phrases when appropriate. This handles *domain-relevant* prepositional attachment, conjunctions, and spatio-temporal locatives.
6) **Coreference Resolution** – Pronouns and definite descriptions referring to named-entities are resolved.
7) **Domain** – finds clause-level (i.e. subject-verb-object) patterns of relevant information in the text, and fills the slots in templates

Finite-state methods have become the "standard" approach for building information extraction systems simply because they have been shown to be the most effective in practice, particularly at the complex phrase and domain levels. A number of experiments have demonstrated that components of an IE system, particularly of the earlier stages that involve more local processing of the input, can be usefully constructed using statistical models.

Tasks such as sentence boundary detection in normal punctuated text can be performed by a statistically trained system (e.g. Reynar and Ratnaparkhi [11]), however, at least for English, it is possible with very little effort to hand-craft a system that has similar near-perfect performance.

Very capable statistical named entity recognition systems have been developed by Bikel et al. [3] and Borthwick et al. [4]. The performance of statistically trained modules still lags that of handcrafted models, however, the ease with which such systems can be trained may compensate for the relatively small loss in accuracy.

While statistical parsing algorithms have been much investigated in the literature, their application to information extraction systems has been limited. This is probably because such algorithms do not have a decisive advantage for producing local bracketing, and the ability to produce marginally better attachment decisions for more global analyses does not seem to have a significant effect on the ultimate correctness of information extraction system output.




## 3 THE LIMITS OF INFORMATION EXTRACTION TECHNOLOGY.

The Defense Advanced Research Projects Agency in the United States has been sponsoring regular evaluations of information extraction systems for about ten years. These MUC (Message Understanding Conferences) would require participants to actually build information extraction systems, whose output on a substantial blind test set (i.e. approximately 100 articles) would be scored for recall and precision against predetermined correct answers.

Figure 1 illustrates the results obtained by the highest-performing systems on different tasks of similar scope and complexity over several successive MUC evaluations. The chart shows that the clusters of scores have become tighter, which suggests that the adoption of similar finite-state technology by the participating sites has led, unsurprisingly, to systems with similar performance characteristics. Statistical significance testing (Chinchor [5]) has in fact shown that there were no significant differences between the top five systems in MUC-6 at a 98% level of confidence.

It is also evident that, while precision has shown a modest increase over the series of evaluations, recall has not shown a similar increase. This may lead one to suspect that the technology is nearing the limits of its effectiveness. Indeed it has been hypothesized that, as evaluated by F-measure (a geometric mean of precision and recall), that no information extraction system would be able to exceed an F measure of 60 on a typical MUC Scenario Template task (a full information extraction task of moderate complexity).

Recently both SRI International and New York University have conducted an experiment to discover the limits of the technology by continuing development of one of the MUC Scenario Template task domains (the MUC-6 Management Succession task) for a period of a year. The results of these experiments (unpublished) show that indeed performance in the low 60's seems to be near the

limits of what one can expect from current technology applied to similar tasks.

As part of MUC-7 some experiments were conducted to determine how well trained humans can perform extraction tasks. The results indicate that a low value for human performance is around F = 85, as indicated in the diagram on the previous page. This is significantly better than any extraction system has done to date on a moderately complex Scenario Template extraction task. If it is indeed the case that limitations to the technology will restrict us to employing extraction systems that are significantly lower than human performance, it is reasonable to ask what tasks might benefit from systems with performance characteristics such as we have available.

## 4 Applying Information Extraction to Information Retrieval

Closer inspection of the results from the MUC-6 scenario template task revealed that the extraction system had higher performance on some areas of the task than others. In particular, the system was quite good at recognizing the central events of relevance to the domain. The MUC score reports include a slot for "text-filtering" which scores whether a system is capable of identifying any relevant event in a text. The FASTUS system had a recall and precision of 87 on this task, which is considerably higher than the recall and precision on the task as a whole.

This experience suggests that an information extraction system that is built to extract certain kinds of information might be used as an effective document filter to improve the results of an information retrieval system. The question was whether the level of performance that can be realistically achieved is sufficient to have a significant effect on an already state-of-the-art retrieval system.

One of the queries in the TREC-5 evaluation was quite close to the MUC-6 scenario template task. The query was

> *A document will announce the appointment of a new CEO and/or the resignation of a CEO of a company.*

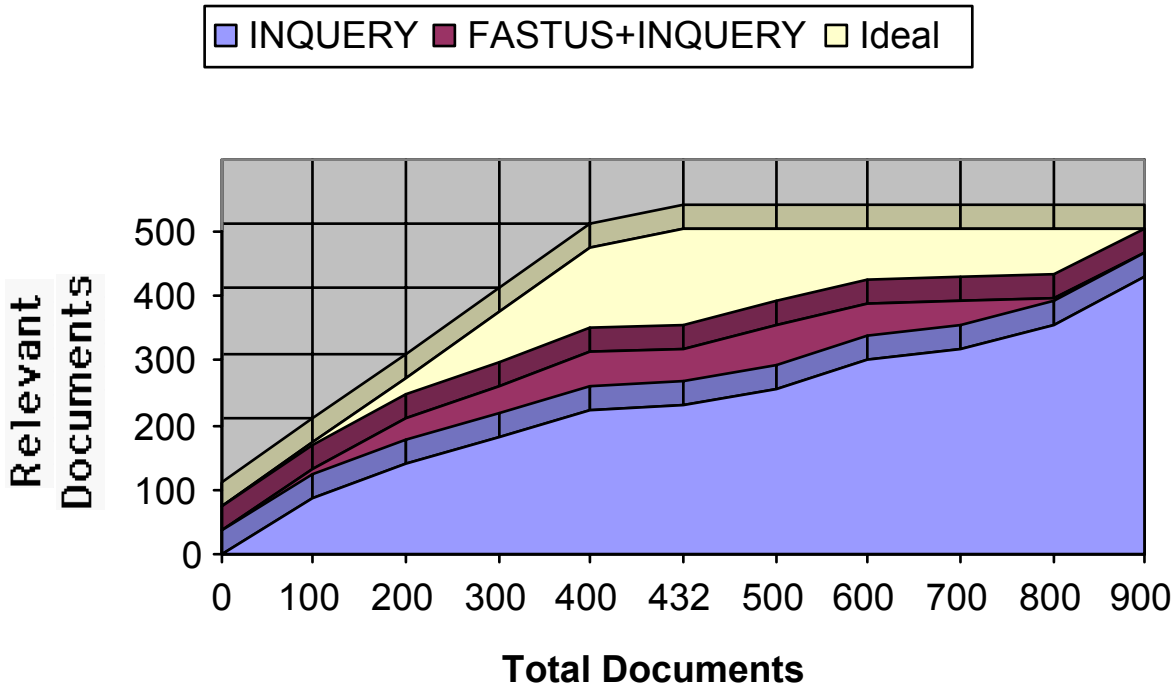
An analysis of the results of the University of Massachusetts INQUERY system [6] revealed certain systematic errors. One error was to return articles on CEO compensation packages in response to the above query. Another frequent error was to return an article in which some high-level manager at a company was appointed or resigned, and the opinions of the CEO of the company were solicited in the article. In each of these cases, the relative frequencies of relevant words were high, yet the requisite semantic relationships were missing.

Because these semantic relationships are precisely what an information extraction system is capable of producing, it seems natural to examine whether an information extraction system can be applied to information retrieval tasks.

The idea of using natural-language processing techniques in information retrieval is not new, and several ideas have been advanced (e.g., Sparck-Jones [12], Lewis [10]. The relatively high overhead imposed by natural-language processing systems has focused the application of natural-language processing techniques on the indexing and preprocessing side of the retrieval operation, rather than actual query processing. Although some small improvements have been reported, success in the application of natural-language processing techniques to information retrieval problems has so far been elusive.

There is a natural distinction to be made within document retrieval tasks between *ad hoc* and *routing* queries. An *ad hoc* query is one that would be asked only once, and therefore it is assumed that no significant analysis of the corpus of documents directed at processing a single query is performed. A routing query, on the other hand, is one that may be performed repeatedly on new sets of incoming documents, and is designed to fill a standing information need. For such queries, query specific tuning of an IR system is reasonable, and it is toward such queries that a combined IE/IR system could be directed.

In a routing task, as it is conceived in the TREC evaluations, a system is given a query and a set of training texts that are considered exemplars of relevant documents. The system's performance is optimized on the query using the training documents. Then a new corpus of test texts is given to the system, and the system pro-




duces a ranked order of the M most relevant texts from the new corpus. A figure of merit for routing systems is "precision at N", i.e. the percentage of relevant texts out of the first N retrieved.

The simplest possible application of information extraction to the routing task would be to build an extraction system to extract the information relevant to the query, and then run the extraction system over all the texts, and rank them in order of the quantity of relevant information it is possible to extract from each one. However, this does not seem to be a reasonable approach to the problem because, although extraction systems are intended to run quickly over large corpora, for very large corpora, the computational demands are still prohibitive. An approach that involves the integration of an extraction system with a retrieval system seems to be most promising. In a routing task designed to produce a ranked list of N documents, the retrieval system would produce a ranked list of considerably more documents, say 2N, and then the extraction system would attempt to extract information from the 2N documents. The results of the extraction would produce a reordering of the top 2N documents, from which the first N would be taken as the output of the hybrid system.

## 5 A Simple Experiment to Improve Routing Precision

The coincidental similarity between one of the TREC topics and the MUC-6 topic presented us with the opportunity to do a simple test to see how much one could improve the output of an IR system by using a fairly well developed extraction system. The MUC-6 scenario template task required about 12 person weeks to implement. This is almost certainly more effort than one would want to devote to optimizing a single routing query, but since the data was available, it was easy to conduct an experiment.

We used the output of the INQUERY system on the TREC topic regarding CEO succession. INQUERY produced a set of 1,000 text documents it deemed most likely to be relevant, ranking them in order of hypothesized relevance. The SRI FASTUS information extraction system was then used to assign a numerical score from 0.1 to 1,000 to the templates it produced as follows:

1) CEO + person name + company name – 1,000.
2) CEO + company name – 100
3) CEO + person name – 10
4) CEO + any transition event – 1
5) CEO + be-verb – 0.1

The score of a phrase was taken to be the sum of the scores of the templates created from that phrase, and the scores of each phrase were summed to give the score for an article.

The rationale behind this scoring scheme is easy to understand. If a template is produced from the FASTUS patterns describing a complete succession event, with company and person, the article is almost certainly relevant. The less information is extracted, the more likely it is that the template was overgenerated. We discovered that identification of a person was less reliable an indicator of a template's correctness than the identification of a company. The final rule was intended to capture the relatively rare cases in which all information about the transition event was stative, e.g., *John Smith is the new CEO."*

This single experiment produced quite positive results. The precision at 100 of the INQUERY system was 87, but the FASTUS reranking produced a precision at 100 of 96 — a considerable reduction in error rate. Overall, the error rate was reduced by a bit less than half, as indicated by the diagram on the previous page. Although extrapolation from such limited data can be risky, the results certainly suggested, that, at least on certain problems, the level of performance that can be realistically achieved by current information extraction systems is capable of significantly improving routing results.

## 6 Testing on a TREC Routing Task

The results on the initial experiment with INQUERY were positive enough to encourage us to attempt a more ambitious experiment. We entered the main routing track at TREC-6 with a hybrid system based on FASTUS coupled with a derivative of the SMART system developed by Tomek Strzlkowski of General Electric. Systems are evaluated on their performance on 47 different routing topics. Our strategy was to develop FASTUS information extraction grammars for as many of the routing topics as possible in the time available. Then, we would take the top 2000 articles as ranked by the GE-SMART system, evaluate them with FASTUS, and score them based on the number of domain patterns that were matched in the article. Ties were resolved by the original SMART ranking. The top 1000 articles would then be returned as the output of the combined routing system.

Because of time limitations, we were able to develop extraction grammars for 23 of the 47 topics, focusing the development on producing rules of high precision, rather than high recall. On topics for which we were unable to develop extraction systems, we simply returned the original GE-SMART ranking unchanged.

The overall result on the routing task was not impressive. We improved the average precision over all topics from 27% to 27.3%. However, it is more illuminating to look at the results from the combined system performance on the 23 topics for which we were able to write extraction grammars.

FASTUS improved the average precision compared to the raw output of the GE-SMART system on 17 of the 23 topics, and in seven of those cases, the improvement was substantial. One topic showed the highest average precision of any system. Unfortunately, on the other six topics, FASTUS lowered the average precision of the GE-SMART system, in two cases by a significant amount.

## 7 Better Integration of Extraction and Retrieval

The experiment with the TREC routing task convinced us that the key to making significant improvements to retrieval using the results of an information extraction system depended crucially on a close integration between the retrieval and the extraction systems for a particular topic.

Although it was impossible to evaluate the precise recall and precision of extraction systems on various topics, it is obvious to a developer of such systems that some topics are more amenable to identification of relevance through the simple predicate-argument structure that clause-level domain patterns in FASTUS are capable of identi-

fying. The management succession topic of the INQUERY system is obviously an example of a topic where information extraction can be usefully employed. However, for some topics, the central fact of the topic is difficult to capture as a subject-verb-object pattern occurring in a single sentence. There is no reason to believe that the same strategy for combining extraction results with retrieval results would work equally well in these different cases. It is also true that as an extraction system improves through continued development, that there is no reason to believe that any combination strategy that was appropriate for early stages of development would also be optimal for later stages of development.

For these reasons we attempted to find a trainable model that could be applied to each topic, for a given version of a retrieval and an extraction system that would yield an optimal combination strategy.

Kehler [9] developed a probabilistic model based on the maximum entropy approach (Berger et al. [2]). A set of contextual characteristics to conditionalize the model was identified that involved both the retrieval system and the FASTUS system. The particular set of features we chose were as follows:

1) A characteristic for each of the five types of contribution to the score (as described in Section 4)
2) A characteristic for whether there were at least two contributions to the score
3) A characteristic for whether there were at least three contributions to the score.
4) A characteristic for whether there were at least two contributions of 10 or above to the score.
5) A characteristic for each domain-phase rule (i.e. a single family of SVO patterns and their syntactic variants)
6) Characteristics based on the level of the retrieval system's ranking of the document.

A binary feature function *f(x, y)* was created that paired each characteristic of a document x with the result of the document being relevant (y). For example, a feature $f_k(x,y)$ pairing a document's score being at least 1000 with its being relevant would be defined as follows:

$$f_k(x,y) = \begin{cases} 1 & Score(x) \geq 1000,\ y = 1 \\ 0 & otherwise \end{cases}$$

These features then define potential constraints on the model that is learned, in which we constrain the expected value of the feature with respect to the probabilistic model ($p_m$) to be that for the feature with respect to the training data ($p_d$). The constraint is defined as follows:

$$\sum_{x,y} p_d(x,y) f(x,y) = \sum_{x,y} p_d(x) p_m(y \mid x) f(x,y)$$

Initialized to the uniform distribution, the algorithm approximates the gain in the predictiveness of the model that would result from imposing each of the possible constraints, and selects the one with the highest anticipated payoff. A constraint corresponding to a feature $f_k(x,y)$ is incorporated by training a multiplier $\lambda_k$ in a model of the form

$$p_\lambda(y \mid x) = \frac{1}{Z_\lambda(x)} e^{\sum_i \lambda_i f_i(x,y)}$$

in which

$$Z_\lambda(x) = \sum_y e^{\sum_i \lambda_i f_i(x,y)}$$

The feature selection process is iterated until the approximate gain for all the remaining unactivated constraints is negligible. The result of the algorithm is the model that has the maximum entropy out of the set of models in which the selected constraints hold.

Using the same data as in the earlier INQUERY experiment, we trained the model with 875 randomly selected texts, and evaluated it on the remaining 125 texts. Positive lambda values were learned for the first four of the five FASTUS score types (1000, 100, 10, and 1). Positive values were also learned for the first four rank features (i.e. INQUERY ranked the text in the top 100). Negative lambda values were learned for two of the verb-class features (the verb BE, as in *Smith was the CEO* and the verb ASSUME, as in (*Smith*

*assumed the post of CEO).* All other features were inconsequential.

For the test, INQUERY alone achieved a precision of 79, INQUERY with manually tuned FASTUS had precision of 88, and the maximum entropy model had a slightly better performance at a precision of 89.

Although the difference in this case was not large, the maximum entropy approach gives us a means of separating those cases in which FASTUS makes a positive contribution to the reranking of the retrieval system output and those cases in which its contribution is negative or indifferent. Training and evaluating this model on each of the 23 topics of the TREC experiment described in the previous section is work that is currently in progress.

## 8 Rapid Application Development Tools for Information Extraction

It is clear that any application of information extraction to information retrieval in the manner suggested in this paper depends crucially on the ability to develop reasonably performing information extraction applications in a new domain with minimal effort, preferably by users who are relatively inexperienced with issues in computational linguistics. This problem is addressed by the *Open Domain System* currently under development by SRI International.

Information extraction systems derive much of their strength from the fact that they are narrowly tailored to a particular domain. One of the weaknesses of earlier efforts in information extraction, typified by SRI's TACITUS system [8], was that, by basing a system on a fully general model of English, one exposed oneself to a variety of problems involving search through large, underconstrained spaces. While such systems might have been capable in principle of providing general solutions to problems in linguistic analysis, in practice unconstrained search meant overall sluggish performance, and efforts to limit the search often resulted in incorrect results as well.

The critical insight that enabled information extraction technology was that it was possible to combine a relatively domain-independent description of simple constituent structure (i.e. noun groups and verb groups) with a relatively small number of highly domain-specific rules for recognizing clausal patterns. Such rules might make many mistakes on sentences outside the system's narrow domain (including some that might contribute to overgeneration of responses), but this was more than offset by fast, robust, generally correct processing of domain-relevant sentences.

Of course, after working within the information extraction paradigm, one is quickly reminded of why generality is a good thing. While it became possible to construct information extraction systems with higher performance on tasks than had previously been possible, it was evident that building such systems was a time consuming and labor-intensive task that required specialized expertise of the implementers. This barrier is one of the primary obstacles to widespread adoption of information extraction technology. If information extraction technology is going to be applied to information retrieval, then the overhead of its application must be considerably reduced, since even if the advantages of error reduction can be significant in many cases, it may be the case that the effort required to take advantage of the error reduction is not worth the cost.

To address this problem, we are currently developing what we call an *Open Domain* extraction system.

In contrast to highly domain-specific extraction systems, an Open Domain system is targeted toward extracting a large variety of simple events of relevance to a very broad domain. In our case, we chose general business and economic news as the targeted Open Domain. This Open Domain includes many of the more specific topic areas for which application systems have been built, including investment, manufacturing, product development, marketing, and management. Another possible Open Domain would be directed toward processing military messages, including information about units, equipment, military facilities, and the kind of events they engage in, including surveillance, deployment, combat, supply, evacuation, etc.

An Open Domain system is not intended to be the final information extraction system, but rather it is a framework and toolkit to enable a user to easily construct an extraction system targeted toward the specific needs of an application within

the general Open Domain. In particular, an Open Domain system provides the following capabilities for the user:

1) Built-in handling of locative and temporal adverbials, avoiding the duplication of these capabilities with each new domain.
2) Basic syntactic metarules that allow a user to specify a subject-verb-object pattern from which rules recognizing syntactic variants such as nominalizations, passives, relative clauses, etc. can be generated automatically.
3) An enriched lexicon that gives specific information about predicate-argument mapping for a sizeable collection of open-domain-relevant verbs.
4) A basic shallow ontology for the Open Domain that identifies the most critical concepts relevant to the domain, and which can be specialized to the requirements of a particular application.
5) A complex ontology for the Open domain that identifies domain-relevant events and their causal relationships.
6) Lexical semantics connecting the basic ontology to words in the lexicon.
7) Subject-Verb-Object patterns for events in the complex ontology.
8) Support for rule learning from examples.

In the business-economics domain, the basic ontology covers such concepts as persons, companies and their subsidiaries, and legislative bodies. The ontology contains a large number of categories in the general domain of "goods and services" although this is quite open-ended and not yet complete. Also included is a category of assets, including real estate, plants, currency, and financial instruments. Each of these ontological categories is arranged in a hierarchy, and paired with a relatively complete correspondence between terms in the lexicon and ontological categories. It is assumed that this basic ontology can provide a good starting point for customizing the system to a domain. For example, the category of real-estate, which includes various kinds of buildings, structures, and real property, can be specialized to a more specific category like "multi-family dwellings" by the user building the domain-specific application. Graphical tools are under development to assist the user in defining these categories and making them complete.

Because the ontology is focused on a broad domain, it is often more useful than an ontology like that of WordNet [7], which was not developed with a particular domain in mind. For example, in WordNet, "business" and "company" are not synonyms, although in most of the articles that one would process in business and economic news, the two are interchangeable.

The core patterns cover the 150 most common verbs and their associated nominalizations from the Wall Street Journal. These patterns can be used with meta-rules to expand each rule into about 15 syntactic variants, that handle locative and temporal adverbials as well.

The Open Domain System is also capable of supporting rule-learning form examples. When the user highlights a sentence, and specifies the relevant information to be extracted, a pattern can be generated that matches precisely the instance in the text. Then parts of the pattern can be generalized along dimensions suggested by the Open Domain Ontology. Without such an ontology, it would be difficult to produce useful generalizations from a small number of training examples.

We used an early version of the Open Domain System to assist in the development of the small extraction domains (called "FASTLETS"). The system at that time did not include the full ontology or business-domain subject-verb-object patterns, but it did include the basic capability of instantiating multiple rules from a single collection of features.

Although a systematic evaluation of the recall and precision of the FASTLETS was impossible due to the absence of scoreable keys, we found that systems with sufficient capability to have a positive impact on the refinement of the routing query output could be constructed with only a half-day's effort by a computer-literate undergraduate with no special training in computational linguistics.

## 9 CONCLUSION

Because tailoring a natural-language processing system to a specific domain has always been such a labor-intensive process, and the relatively sluggish performance of general NLP systems, it has long been assumed that natural-language

processing techniques are of limited use in information retrieval applications.

We feel that this assessment of the contribution of NLP to information retrieval should be re-evaluated in the light of recent advances in information extraction technology. These advances have led to systems that are fast enough to process large volumes of text in relatively short periods of time, and through advances such as the Open Domain technology, it has become feasible to build domain-specific systems with relatively little effort by people without highly specialized knowledge. For routing applications where a standing information need is satisfied by the appropriate classification of new incoming documents, improvements in retrieval accuracy can well be worth the relatively modest investment in system-development time.

## REFERENCES


[1] D. Appelt et al., A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text, in Roche and Schabes (eds.) Finite State Devices for Natural Language Processing, MIT Press, 1996.

[2] A. Berger et al., A Maximum Entropy Approach to Natural-Language Processing, Computational Linguistics 22 No. 1, 1996.

[3] M. Bikel et al, Nymble: A High Performance Learning Name Finder, Proceedings of the Fifth Conference on Applied Natural-Language Processing, 1997, pp. 194-201.

[4] A. Borthwick et al., Exploiting Diverse Knowledge Sources via Maximum Entropy in Name Recognition, Proceedings of the the Sixth Workshop on Very Large Corpora, Montreal, Canada, August 1998.

[5] N. Chinchor, Statistical Significance of MUC-6 Results, Proceedings of the Sixth Message Understanding Conference, 1995, pp. 39-44.

[6] B. Croft et al. Recent Experiments with INQUERY, Proceedings of the 4th Text Retrieval Evaluation Conference (TREC-4), 1996.

[7] C. Fellbaum (ed.) WordNet: An Electronic Lexical Database, MIT Press, 1998.

[8] J. Hobbs, Interpretation as Abduction, Artificial Intelligence 63, No. 1, 1993, pp. 69-142.

[9] A. Kehler et al. Using Information Extraction to Improve Document Retrieval, unpublished manuscript.

[10] D. Lewis, Text Representation for Intelligent Text Retrieval: A Classification-Oriented View, in P. Jacobs (ed.) Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval, Lawrence Earlbaum and Assoc., 1992.

[11] J. Reynar and A. Ratnaparkhi, A Maximum Entropy Approach to Identifying Sentence Boundaries, Proceedings of the Fifth conference on Applied Natural Language Processing, 1997, pp. 16-19.

[12] K. Sparck-Jones, Assumptions and Issues in Text-Based Retrieval, in P. Jacobs (ed.) Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval, Lawrence Earlbaum and Assoc., 1992.

# Combining linguistic and knowledge-based engineering for information retrieval and information extraction

Paul E. van der Vet and Bas van Bakel
Vossius Laboratory for Content Engineering
CTIT, University of Twente
P.O. Box 217, 7500 AE Enschede, the Netherlands
Phone +31 53 489 3694, fax +31 53 489 3503
Email {vet, bakel}@cs.utwente.nl

## Abstract

Controlled-term indexing (the method of choice for multimedia collections and still very popular for purely textual material), appears an expensive solution because it takes huge resources and manual indexing. It is not possible, however, to perform a well-founded asessment of various approaches to information retrieval. We discuss ways to improve controlled-term indexing and illustrate these by looking at the Condorcet project carried out at Twente by us and co-workers. We round off with a discussion that, we hope, will raise more questions than it answers.

**Keywords:** Knowledge-Based Systems, Language Technology, Information Retrieval, Multimedia, Information Extraction

## 1 Introduction

Among the papers of the French mathematician and social philosopher Condorcet (1743–1794) found after his death, there is a proposal to assign each and every piece of knowledge a unique code. The code would serve two purposes: it could be used to organise libraries, and by comparing the codes with what was known, lacunae in knowledge could be discovered. Organisation of libraries was becoming a pressing problem in Condorcet's time because the production of printed material was turning into a flood. (Still, the antique library of Alexandria with its 700,000 manuscripts must have been a nightmare for newcomers,too.) Condorcet's proposal basically defines what we now call *controlled terms*: tokens taken from a pre-defined list with a meaning that is fixed with respect to that of the other tokens in the list. Then, as now, these terms were assigned by hand. One of the advantages of controlled terms, already noted by Condorcet, is their independence of medium in which the information is expressed. Controlled-term indexing therefore is the method of choice for multimedia collections.

We will first contrast controlled-term and uncontrolled-term approaches to find the relative pros and cons of each. The pressing question is: which method is preferred, relative to the situation? As we will show, this question cannot be answered for lack of data. We then turn to improvements of current controlled-term indexing practices. We discuss one technique in more detail by way of an overview of an information retrieval project carried out at our group, named after Condorcet. We round off with a discussion rather than conclusions.

## 2 The mystery of the continued use of controlled terms

About a decade ago, the Dutch sociologist Laeyendecker wrote a book under the title "Does progress bring us any further?" (our translation) [19]. For information retrieval (IR), many IR experts say 'yes' but not every end-user believes them. Probabilistic approaches are claimed to be the cheap and satisfactory solution to *ad-hoc* IR problems but many end-users stick to controlled terms. A probabilistic approach produces a document representation that consists of uncontrolled terms (basically, stemmed words or regularised phrases from the text itself minus the so-called stopwords). Controlled terms, by contrast, are taken from pre-defined resources such as thesauri and classification systems and need not occur in the document to which they are assigned. Given the impressive investments needed to make

controlled-term systems work, one wonders why such systems are still around and whether they perform as well as their users expect. These questions seem simple but turn out to be very difficult to answer.

The two approaches can be contrasted as if they were rivals, which they are not. (Any sensible system designer will offer users both possibilities.)

Controlled terms have two advantages. First, indexers and users alike at least share a common resource: the store of controlled terms. This reduces uncertainty. Second, because controlled terms are assigned by hand, they are completely media-independent. Texts in different languages about tigers, photographs and videos of tigers, and audio files with tiger sounds all receive the same controlled term, say, `tiger`. Against this, manual work is error-prone. When the *Chemical Abstracts* thesaurus was converted from hardcopy into an electronic version, many errors were introduced [23]. In our own investigations we have inspected *Engineered Materials Abstracts* and *Excerpta Medica* and found indexing errors, although there were fewer errors than we (biased computer scientists) initially thought. Further, it is a nuisance that proper names are seldom, if ever, declared to be controlled terms. In a number of retrieval situations we will want to search for proper names. Finally, since the development and maintenance of the store of controlled terms and indexing the documents have to be done by hand, a huge investment is needed. For comparison, Chemical Abstracts Services has close to 1,000 employees.

The big advantage of uncontrolled terms is low costs of indexing. Preparation of document representations can be fully automated, and costly investments like those for term resources are wholly avoided. The major disadvantages are: inability to abstract from the media used, and ambiguity of tokens: natural-language words or pictorial elements.

As regards effectiveness, there is ample material on uncontrolled-term systems. For controlled terms, there is no empirical material of a comparable quality and breadth. We simply do not know how well controlled-term systems perform, so a comparison between the two kinds of system on effectiveness is impossible.

In a famous experiment [3, 2], Blair and Maron measured the effectiveness of a STAIRS system that used uncontrolled-term indexing. They found the measured effectiveness disappointing and certainly below the requirements imposed by the situation. Searching was hindered by the very many ways the same subject can be characterised in natural language and even by cross-document anaphora (like "the subject of your last letter"). Blair and Maron concluded that it is simply infeasible for users to predict what words, word combinations or phrases would occur in the documents they sought and would not occur in the documents they did not seek. They advocated the use of controlled terms to enhance effectiveness, although they have not conducted a follow-up investigation to substantiate the claim. We find their argument plausible, at least for the situation and corpus investigated. For us, it is among the reasons to pursue a controlled-term approach in our own research.

Later experiments at TREC [34, 28] show that the situation has improved with respect to the figures found by Blair and Maron, but not spectacularly so. At the last TRECs, results seem to have reached a plateau [28]. From this, one cannot conclude that uncontrolled-term systems perform in an unsatisfactory way. After all, not every IR situation is as demanding as that investigated by Blair and Maron. We estimate that uncontrolled-term systems are a good choice for quite a number of applications.

The main problem in comparing approaches is that there are no estimates, and *a fortiori* no reliable estimates, of the total cost-benefit balance of an IR session. Benefits include expenses avoided and direct gains. Costs include:

1. Costs of setting up the system (including indexing), depreciated over sessions.
2. Costs of use, which can be broken down into:
   (a) Hardware costs (processing time, memory usage, network usage).
   (b) Costs of searching (handling some requests may take hours or even days of query construction and interactive refinement).
   (c) Costs of sifting the set of retrieved documents.
   (d) Costs incurred by missing relevant documents.

Item 2(a) can be safely neglected relative to the other cost items. The familiar measures of precision and recall bear on items 2(c) and 2(d) only and both the measurement and the subsequent interpretaion of these quantities is fraught with difficulties (see, for instance, [16]). Results in

terms of precision and recall at best represent an incomplete picture.

We simply do not know key items such as the costs of searching and costs of missing relevant documents. Just to illustrate how reliable cost-benefit figures, if they were available, would affect our judgments, consider a fictitious comparison between a controlled-term system and an uncontrolled-term system with identical recall and precision. The huge investments needed to get the controlled-term system into the air are earned back if the average session lasts significantly shorter than the average session on the uncontrolled-term system.

Cooper [7] has proposed to measure retrieval effectiveness in terms of the amount of money a person is willing to pay for having a system process an information request. Obviously, such a person has little to go on.

## 3 New approaches to controlled-term indexing

In our own research, we further explore controlled-term indexing. Computers can be employed in this approach, too, to obtain a more effective and efficient way of working. We investigate two improvements: better term resources and lowering of costs.

Term resources can be improved because current thesauri and classification systems are not very expressive. This state of affairs is due to the fact that, until recently, these resources had to be distributed and consulted in printed form. Computer manipulation opens new possibilities. A tangled hierarchy spanned by a number of different relations, for instance, becomes unreadable in printed form but is easy to understand and use with the help of computer programs. Modern jargon calls the computer-age successors of thesauri and classification systems *ontologies* [12, 22]. See [13, 32] for examples of ontologies that are too complex to be handled by other than automated means.

Ontologies allow indexers to assign *co-ordinated index terms* to documents to enable more precise searching. For example, suppose `aspirin` and `headache` are both controlled terms. With the help of co-ordination, we can specify the nature of the relation between these two terms in cases where they are both assigned to a document: for instance, `cures(aspirin, headache)` or `causes(aspirin, headache)`. Searchers can thus limit their search by specifying the relation. The query engine we developed for these terms [33] also handles generalisations, *e.g.*, `cures(any(medicine), headache)` will retrieve all documents about medicines against headache.

Costs can be lowered by partially automating the process of assigning controlled terms. (We say 'partially' because fully automatic assignment will not be both technically and economically feasible for a long time to come.) Text understanding and figure understanding are the fields that will have to spawn the necessary techniques. Figure understanding is a long way off, but text understanding is within reach. Indexing texts using text-understanding techniques is the subject of the next section. Documents in other forms will still have to be indexed by hand. The advantage of complete media-independence of controlled terms is not abandoned.

What remains are the costs of maintaining the resources. Resources are nowadays often simply lacking, so on the short term there are additional costs for making those resources in the first place. The resources include ontologies, grammars of natural languages, lexica that map natural-language words and phrases onto conceptual equivalents, knowledge bases with domain knowledge, and programs. We estimate that an ontology alone is more expensive than a thesaurus or classification system, so on the face of it this route only augments already substantial costs.

There are grounds, however, to think that many of the required resources will come into existence anyway. Unlike thesauri and classification systems, the resources required for semi-automatic indexing are also valuable for other applications. It is not difficult to foresee a future in which manipulation of information on the level of its content is commonplace. Workers in medicine have realised this earlier than their colleagues in other disciplines. The *Unified Medical Language System* (UMLS) [21] will grow into a body of resources that covers most of the needs of a semi-automated indexing system. Other disciplines will undoubtedly follow.

In our Condorcet project, we use UMLS as a resource. Basically, UMLS is a collection of thesauri, a lexicon that maps nouns and some phrases onto thesaurus terms, and a semantic network. The semantic network defines what are called *types* in a taxonomic hierarchy. Every term in every thesaurus is assigned a type to disambiguate meanings, for example, `cold` as indication of temperature *versus* `cold` as a disease.

From the thesauri, we use the MeSH Main Headings and the MeSH NM file (terms for chemicals). The combination of MeSH term and UMLS type is a concept in the sense of an ontology. The semantic network further defines about fifty relations that may hold between terms, depending on their types. We use these relations as co-ordinators to construct co-ordinated index terms such as `affects(zonisamide, epilepsy)` to index a document that discusses the use of zonisamide as anti-epileptic.

## 4 Condorcet

### 4.1 Overview

Condorcet (funded by the Dutch Technology Foundation (STW) through the *Werkgemeenschap Informatiewetenschap*, the Dutch Society for Information Science) focuses on semi-automatic indexing using controlled terms. We present an overview here; readers are referred to the Condorcet web site at

`http://www.cs.utwente.nl/condorcet/`

for more information and publications, including the three *Annual Reports* that have appeared so far.

Condorcet aims to build a prototype indexing system for large volumes of documents covering two scientific domains: *mechanical properties of engineering ceramics* as a field of materials science, and *epilepsy* as a subfield of medicine. Two domains rather than only one were chosen to avoid bias in the design of the indexing system. Ideally, when switching to another domain only the domain resource has to be changed. The documents in the development corpus are taken from machine-readable one year volumes of two bibliographic journals: the 1988 volume of *Excerpta Medica* from Elsevier Science Publishers, and the 1990 volume of *Engineered Materials Abstracts* from Materials Information. The prototype will be tested on 400 documents. Figures 1 and 2 present examples of document descriptions taken from the two sources. In the course of designing the system, we continuously incorporate techniques that enable the system to process much larger volumes, up to several hundred thousand documents.

Basically, indexing by Condorcet consists of mapping title plus abstract onto terms and co-ordinators by making intensive use of three kinds

AN: 88100203
TI: Effects of zonisamide in children with epilepsy
AB: The effects of zonisamide (1,2-benzisoxazole-3-methanesulfonamide: AD-810) were studied in 50 children with epilepsy, ranging in age from 3 months to 20 years (mean, 10.5 years). The types of epilepsy were primary generalized in one case, secondary generalized in 32, and partial in 17. The initial dose was 1-6 mg/kg/day and the dose was increased to 1.5-15 mg/kg/day. Four cases (8%) showed a complete disappearance of seizures and thirteen patients (26%) had a disappearance rate of 50% or more of seizures. Disappearance or improvement of seizures was obtained in 31% of the cases of generalized epilepsy and in 41% of the cases of partial epilepsy. Zonisamide was effective in 39% of cases of Lennox-Gastaut syndrome. Seizures completely disappeared in three of the four new cases. Spike discharges disappeared or significantly decreased in 22% of the cases that had undergone electroencephalograms. The blood levels of zonisamide were 10.8-18.8 mu g/ml in the three new cases when the seizures were controlled. Side effects such as drowsiness, ataxia, and salivation were observed in 42% of the children, more particularly in children receiving polypharmacy.

Figure 1: Part of a document description from the epilepsy domain, © Elsevier Science. 'AN' identifies the primary key, 'TI' and 'AB' the title and abstract parts. The present text reproduces the ASCII text as it is found in the file, hence, for instance, the string 'mu g/ml' for the SI unit $\mu$g/ml.

of knowledge. Knowledge of language and knowledge of the domain are combined to generate conceptual representations for the sentences in the document description, and indexing knowledge is used to generate index concepts from these conceptual representations. This indexing strategy is based on the idea of efficient use of the different kinds of knowledge. It is fully tuned to the objective of controlled-term indexing rather than focused on either linguistic or knowledge-based engineering, as is done in quite a number of other research projects [1, 11]. We return to this point below, in the discussion.

The problems involved in mapping document descriptions onto index terms and co-ordinators are linguistic problems and problems that involve inferences using domain knowledge. Therefore, combining linguistic and knowledge-based engineering appears a logical (but far from trivial) answer. Apart from how to make the combination conceptually, a more practical problem Condorcet tackles is how to design and develop a prototype indexing system that meets the design and

01 9001C1-C-0019
02 Influence of Ambient Temperature Sliding Velocity Under Unlubricated Sliding Conditions on Friction and Wear of Si sub 3 N sub 4 Up to 1000 deg C.
03 The tribological behaviour of Si sub 3 N sub 4 /Si sub 3 N sub 4 sliding pairs in pin-on-disk configuration for sliding velocities between 0.03-3 m/s, constant load of 10 N and environment-temperatures between 22-1000 deg C is dependent on the overlap ratio, the temperature and the sliding velocity. An influence of the phase composition was not observed for the three tested commercial Si sub 3 N sub 4 materials. The results are: (1) Coefficient of friction lies for solid state friction under steady state conditions between 0.5-1. (2) Wear rate increases with rising ambient temperature–especially at sliding speeds < 1 m/s. (3) The tribological behaviour for temperatures => 400 deg C is characterized by a high wear/low wear transition with increasing velocities. (4) The influence of overlap ratio on wear increases with increasing ambient temperature. A small overlap ratio is tribological disadvantageous for Si sub 3 N sub 4 sliding pairs. Si sub 3 N sub 4 /Si sub 3 N sub 4 sliding pairs do not meet for the described sliding claims without lubrication.

Figure 2: Part of a document description from the materials science domain, © Materials Information. As in the epilepsy example, the present text reproduces the ASCII text of the source. The string 'Si sub 3 N sub 4' stands for the chemical formula $Si_3N_4$, and '=>' for the symbol '$\geq$'.

development criteria set out at the beginning of the project [31]. In this respect, Condorcet has clearly been a two-faced research project from the start: in order to build a working application (the main objective of the project), the entire indexing process had to be conceptualized first. The outcome of the latter may be regarded as Condorcet's contribution to IR, and in the long run it may prove instrumental for the more difficult and ambitious task of information extraction as well. Although we were lucky to be able to draw on substantial experience from an earlier project with this approach [30, 26], we still had to tune the results of this earlier work to Condorcet's task, and build a working prototype in accord with the design criteria.

Condorcet's approach to document indexing by employing linguistic engineering can hardly be considered new. There are many examples of IR systems in which linguistic engineering plays a prominent role – e.g., ADRENAL [20], FERRET [24], MEDLEE [10], and AIMS [17]. Not everyone is prepared to regard these contributions as being 'significant': for instance, Harman [15] asserts that linguistic engineering still has to make its first significant contribution to improving document retrieval systems. Smeaton [27] offers a reason for this perceived inadequacy: according to him, IR and linguistic engineering are inherently different processes. IR is inexact whereas linguistic engineering is not, and only a change of approaches in both IR and linguistic engineering will lead to progress, as the current approaches only cause "the 'butting of heads', which we see at present with IR attempting to cherry-pick any appropriate techniques from NLP" ([27], p. 136).

In contrast to Harman, we think the cited works do contribute to better indexing systems. We also disagree with Smeaton's opposition between IR and linguistic engineering. In our view, an index term is abstract rather than vague; see the discussion at the end of this paper. In Condorcet, then, the linguistic engineering module is tuned to the specific needs that apply to controlled-term document indexing, causing it to differ from general-purpose linguistic engineering systems. Linguistic engineering within Condorcet is highly application-oriented; the knowledge-based approach guarantees that the linguistic engineering system is based on linguistic principles, and that therefore no *ad hoc* solutions will be applied.

## 4.2 System design

The design criteria underlying Condorcet are mainly concerned with costs of setting up and maintaining the system, and anticipating reuse - at least of parts of the system - for tasks similar to IR, like information extraction and text summarization. This has led to a sequential modular system, in which different kinds of knowledge are used by separate parts of the indexing system, which is depicted in figure 3. To anticipate indexing of reality-level volumes of documents, indexing and retrieving documents should be fast and robust to reduce costs of using the system to an acceptable level. Reuse of existing domain knowledge resources like UMLS is another cost-saving measure. Maintainability and extendibility are served by following the familiar principle of knowledge-based engineering to separate knowledge from the programs that use it.

The system design makes it easy to determine which knowledge contributes in what way to the overall task of indexing, and therefore the system can be optimised for the task of indexing. This will be done after evaluation of the entire sys-

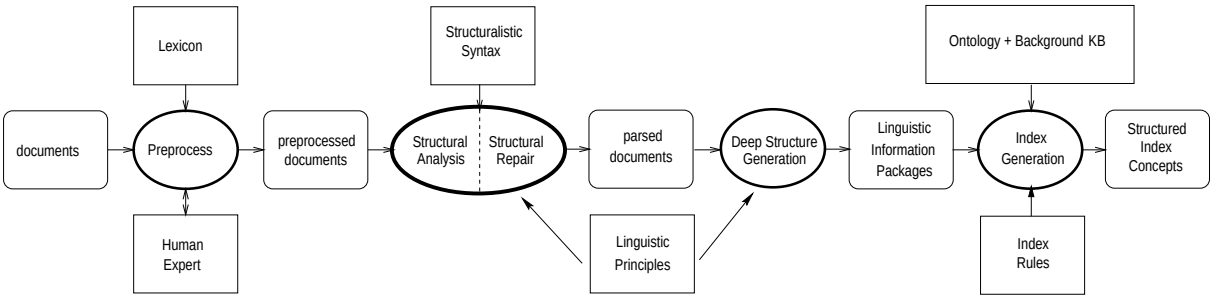



Figure 3: Condorcet's indexing process.

tem. It should be noted that this is only possible because we started out with a conceptualisation of the indexing task rather than an implementation, and because we took a rigid approach to the design of a sequential modular system, by postulating a separate module for each different type of information used. This is why the indexing process consists of several subprocesses, discussed below.

## 4.3 Assiging index terms

Index terms, possibly co-ordinated, are assigned to documents in a four-step process, see figure 3. The first step is pre-processing. The texts are converted from the format in which they are found into a canonical format in which SGML tags are used to delimit and identify the various parts. Further, the text is tokenised, which means that lexical units are recognised and tagged with the appropriate part-of-speech information (like plural noun, determiner, passive participle of verb, and the like). In case of lexical ambiguity, simply all possibilities are given. Tagging is based on the CELEX lexicon [4], transformed to reflect the parts of speech we want to distinguish, and on additional information for lexical items not found in CELEX. Tokenisation is interactive: when a lexical item cannot be recognised, the user is asked to supply the missing information. The pre-process is stable.

The texts are now ready for semi-automatic assignment of co-ordinated and/or unco-ordinated index terms. Obviously, the assignment will have to be based on an analysis of the natural-language text. The major problems in mapping descriptions to concepts and relations are linguistic in nature. Therefore we need knowledge on how concepts and relations can be expressed in natural language. It appears that there are many possible ways, by using different syntactic constructions. Consider the following sentences:

- Effects of zonisamide in children with epilepsy.
- Zonisamide affects epilepsy.
- Epilepsy was affected by using zonisamide.
- Zonisamide was effective in 39% of cases of epilepsy

Given the coarse granularity of the ontology used, these sentences all express the same co-ordinated index term `affects(zonisamide, epilepsy)`, only in a different syntactic form. In order to produce the proper structured conceptual representations, we not only need to determine the syntactic surface structures of these sentences but also their underlying deep structure. The deep structures contain the necessary information for mapping natural-language utterances onto terms and co-ordinators.

To obtain deep structures we use syntactic principles of Chomsky's Government & Binding (GB) theory [5]. This theory is chosen for theoretical and practical reasons. First and foremost, Chomsky's *Principles & Parameters* framework can explain a wide variety of language phenomena using just a few assumptions (see also [9]). Using GB therefore makes it possible to develop a relatively small and elegant, principle-based linguistic engineering system. As it is of secondary interest how these principles should be formalized in GB [6], we can freely formalize and implement them, and separate the linguistic knowledge resources from the processes as required.

### Structural analysis

Deep structures are generated from surface structures. The Structural Analyzer produces a sur-

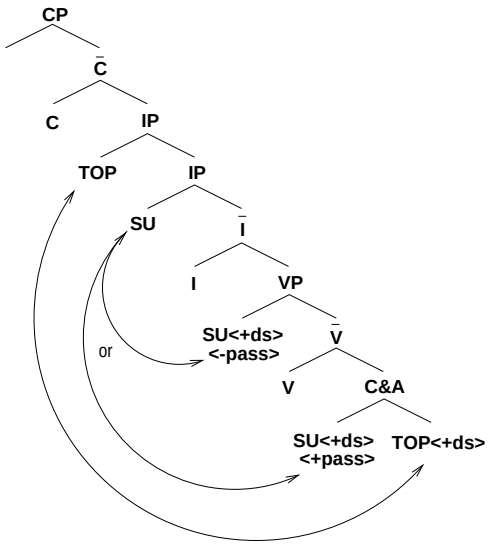



Figure 4: Enriched canonical tree structure. SU-node contains the syntactic subject and TOP collects all topicalised elements (CP's, PP's, NP's, adverbs). V contains the main verb, and C&A all verbal complements and adverbials. In enriched structures, TOP and SU are linked to their deep structure positions. SU is linked to the deep structure subject position of the main verb in case of active, and to the leftmost position under C&A in case of passive sentences.

face structure in a canonical format for every sentence in the document description, according to X-bar theory [18]. In this process, N,V, A, and P are regarded as *lexical heads*, and C(omplementizer) and I(nflection) as *non-lexical heads*. At structural level, the major categories are analysed in accord with the $\bar{X}$ Conventions. The *maximal projections* for the lexical heads are represented as NP, VP, AP, PP, IP and CP, respectively.

The Structural Analyzer, developed by Condorcet team member Erik Oltmans [25], is robust. It handles erroneous input like misspelled words and ungrammatical sentences by means of reanalysis. It has been everyday linguistic engineering practice in the last decade or so to tackle erroneous input with some robustness device, but Condorcet's Structural analyzer is unique in that it performs reanalysis in a purely principle-based fashion. It contains a number of reanalysis rules based on linguistic principles that transform partial parses (containing *chunks*) into complete parses, in accord with the canonical X-bar format. Three strategies are used in this respect: *chunking*, *sloppy agreement* and *catch-all rules* [25]. The catch-all rules ensure that the system displays behaviour known as graceful degradation. A parse is found that contains as much syntactic information as possible.

### Deep structures

The next indexing step involves the generation of deep structures from the surface structures. Actually, *Enriched Surface Structures* (ESSes) rather than deep structures are generated. ESSes are constituent structures in which constituents are linked to their deep structure positions, without changing the word order of the sentence. Deep structure generation is performed by a transformational process, based on *Move* $\alpha$ rules and *Control Structure* rules, reflecting the principles and parameters of GB theory. A crucial condition for all *Move* $\alpha$ rules is that they obey the *Subjacency Condition*, thus adhering to the principle of strict cyclicity [5]. Linking constituents to deep structure positions is making use of *Case Theory* and *Theta theory*. Generating the ESS of a sentence consists of linking constituents to their deep structure positions. The result of this process is illustrated in figure 4. Deep structure positions are the deep structure subject position for the external theta role, and positions under C&A for internal theta roles.

ESS generation is complicated for a number of syntactic constructions. Consider infinitival clauses lacking overt syntactic subjects, like in the sentence *"The purpose was to inquire into the determinants of psychopathology"*. It is the task of deep structure generation to make the semantic subject explicit.

### Linguistic Information Packages

After ESS's have been generated, *Linguistic Information Packages* (LIP's) can be generated for all XPs in the sentence, in a simple fashion. A LIP consists of the head of the (lexical) XP, and the heads of the lexical XPs that are in theta role positions (i.e., subject position and object position) of the matrix XP (see figure 5). LIPs contain the essential linguistic information from which structured concepts can be derived by using domain and background knowledge only. In other words, once LIPs have been generated for all XPs in a sentence, no further inferencing using linguistic knowledge has to be performed.

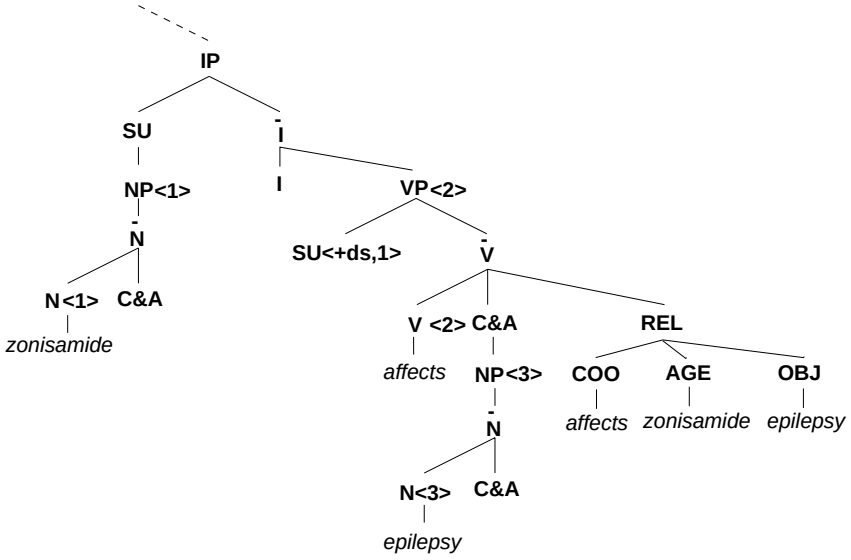



Figure 5: Syntactic tree containing the Linguistic Information Package for the VP of *zonisamide affects epilepsy*. The LIP consists of a candidate *relation* (REL), consisting of a candidate *co-ordinator* (COO), an *agent* (AGE) and an *object* (OBJ). Constituents are linked to deep structure positions by indexes (e.g. `<1>`).

**Index generation**

LIP's are passed on to the last module of Condorcet, called 4MOD. It consists of a knowledge base and an index generator. The knowledge base consists of four parts: (1) a lexicon that maps natural-language words (and occasionally phrases) onto thesaurus terms; (2) a list of terms; (3) a semantic network; and (4) a rule base. With the exception of the rule base, all components are taken over from the UMLS knowledge sources. For the materials science domain, there is no knowledge resource comparable to UMLS and Condorcet's knowledge base for this domain is in an embryonic stage. The rule base contains mainly disambiguation knowledge to select, for instance, the correct interpretation among the many potential interpretations of 'of': compare "effects of zonisamide" (co-ordinator `affects` with `zonisamide` as first argument and no apparent second argument) and "the inferior colliculus of rats" (co-ordinator `part-of` with `inferior colliculus` as first argument and `rat` as second argument).

The index generator takes a LIP as input. It first makes explicit all possibilities implicitly coded in the LIP. Then, if there are several possibilities, a selection process is started to rule out candidates. The knowledge is supplied by the knowledge base. For example, suppose we have received a LIP with a head that, according to the concept lexicon, might give rise to the co-ordinator `affects`. The index generator will now search the semantic subject and semantic object positions in the LIP for strings that give rise to index terms. If found, those index terms may be arguments to the co-ordinator just found, `affects`. However, the terms that are allowed as first and second argument of this co-ordinator have to be of specific types. Using the type assignments, this is checked. If the candidate co-ordinated term meets both linguistic and type-compatibility criteria, the interpretation is judged correct. Else, it is discarded and a search starts for a new interpretation.

When all LIPs have been processed, the result is polished. Duplicates are removed, as are index terms that are superterms of terms also on the list.

## 4.4 Provisional results

We are now in the last year of the project. Provisional testing has yielded promising results. The indexing system except the last module is able to process the larger part of the development corpus. Only the last module of Condorcet is still under development. We still need to conduct evaluation

experiments.

We find that combining linguistic and knowledge-based engineering strategies in document indexing is a viable strategy. Especially the use of substantial linguistic engineering has paid off, even though not all possible linguistic structures (adposition, extraposition, to name a few) are covered by the system. We think that coverage of these structure types is not needed for indexing purposes. We cannot substantiate this, however, because we are unable to compare the current approach with one in which these structure types are covered. We expect that for the more challenging task of information extraction, we will have to cover these structure types.

## 5 Discussion

Above, we have observed that one of the salient advantages of controlled terms is their independence of the language in which the document happens to be written. But the semi-automatic indexing system that assigns such terms cannot be language-independent, at least not entirely. The ideal is a system with clearly separated language-dependent and language-independent modules. Switching from English to Japanese texts would then require replacement of the modules for English by their counterparts for Japanese while the rest of the system remains unaffected.

Here the questions start. One is: is this modular design possible? Doubts are raised by observing that certain tasks need both linguistic and domain-related knowledge. In an earlier study [29], we demonstrated that anaphora resolution improves by having the program take recourse to domain knowledge in addition to linguistic knowledge. In Condorcet, disambiguation of PP-attachment is performed using UMLS constraints on relations, surely domain knowledge. Thus, it is simple to keep linguistic and domain-related resources apart but it is an open question whether the programs that use both kinds of resources can be ported to other languages without difficulty.

The combination of linguistic engineering and knowledge-based engineering is fascinating in its own right. Earlier publications (like [11, 1]) have approached the subject from the linguistic point of view. Like other work we have done in this direction, Condorcet approaches the subject from the point of view of the application to be built. The issue then becomes one of selecting resources. Sticking to the Condorcet example, for any text the search space is formed by all controlled terms (co-ordinated or not) defined by the ontology. The analysis steps are there to make constraints explicit. The constraints narrow down the search space, eventually leaving only those terms that can be assigned legitimately. (See [14] for a similar approach.) This view treats linguistic and domain knowledge as being completely on a par, without any pre-defined sequence or priority. This way of viewing the problem raises a host of interesting research questions. One of our favourites is: would it help (be more effective, be more efficient) to perform a tentative mapping on a knowledge representation first and perform linguistic analysis only later to narrow down the remaining possibilities?

Another direction in which this work can be extended is that of information extraction. From the Condorcet point of view, the difference between assigning controlled terms and transforming a text into a knowledge representation is gradual. Controlled terms, particularly co-ordinated terms, are viewed as knowledge representations that abstract from what the text actually asserts about the subject. To illustrate, `cures(aspirin, headache)` is a controlled term while ¬`cures(aspirin, headache)` ("aspirin does not help against headache") or `cures(aspirin, headache, 85, human)` ("aspirin cures headache in 85% of human patients") are ways to express what the text asserts. The Condorcet system can be enhanced to deliver the latter kind of output, turning it into an information extraction system.

Information extraction is the inevitable successor of information retrieval that, in the way we have discussed it, is better called document retrieval. A document is a combination of content and wrapper. Now information is exchanged over networks, the wrapper stands in the way of reuse of information by desktop applications of the user. We do not claim that each and every message can be couched in a knowledge representation language while preserving the nuances and modalities. But in particular in the natural sciences and engineering there is a growing need for information that is exchanged in more formal languages. The forerunners here are molecular biologists, who exchange genetic information in a form that facilitates reuse by the receiver. There is even a system that supports peer review of electronically exchanged genetics findings [8]. At the moment, however, the majority of researchers stick to articles as their mode of communication. Information extraction can be used retrospectively and concurrently to make the in-

formation available in a form fit for computer manipulation that augments the article itself. To make this possible, we need a more thorough understanding of the delicate interplay of linguistic and domain-related knowledge.

## Acknowledgements

The ideas discussed in this paper have been formed in discussions with Condorcet team members Reinier Boon, Nicolaas Mars, and Erik Oltmans and former team member Jeroen Nijhuis. Other people who have contributed in one way or another, sometimes unwittingly and perhaps unwillingly, are Harold Boley, Peter Bosch, Theo Huibers, Franciska de Jong, John Mackenzie Owen, Gerrit van der Veer, and Arjen de Vries.

## References

[1] James F. Allen. Natural language, knowledge representation, and logical form. In Madeleine Bates and Ralph M. Weischedel, editors, *Challenges in natural language processing*, pages 146–175. Cambridge University Press, Cambridge, 1993.

[2] David C. Blair. STAIRS redux: thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science*, 47:4–22, 1996.

[3] David C. Blair and M.E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28:289–299, 1985.

[4] Gavin Burnage. *CELEX - A guide for users*. Centre for Lexical Information, Nijmegen, The Netherlands, 1990.

[5] Noam Chomsky. *Lectures on government and binding*. Foris Publications, Dordrecht, The Netherlands, 1981.

[6] Noam Chomsky. On formalization and formal linguistics. *Natural Language and Linguistic Theory*, 8:143–147, 1990.

[7] William S. Cooper. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24:87–100, 413–424, 1973.

[8] Jérôme Euzenat. Building consensual knowledge bases: context and architecture. In Nicolaas J.I. Mars, editor, *Towards very large knowledge bases. Knowledge Building and Knowledge Sharing 1995*, pages 143–155. IOS Press, Amsterdam, 1995.

[9] Sandiway Fong. *Computational properties of principle-based grammatical theories*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Mass, 1991.

[10] C. Friedman, G. Hripsak, W. DuMouchel, S.B. Johnson, and P.D. Clayton. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1:83–108, 1995.

[11] Peter Gerstl. Linking linguistic and non-linguistic information. *Data and Knowledge Engineering*, 8:205–222, 1992.

[12] Thomas R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43:907–928, 1995.

[13] Thomas R. Gruber and Gregory R. Olsen. An ontology for engineering mathematics. In Jon Doyle, Erik Sandewall, and Pietro Torasso, editors, *Principles of knowledge representation and reasoning: proceedings of the fourth international conference (KR'94)*, pages 258–269, San Francisco CA, 1994. Morgan Kaufmann.

[14] Udo Hahn and Klemens Schnattinger. Ontology engineering via text understanding. In José Cuena, editor, *IT KNOWS (Information technology and knowledge systems). Proceedings of the XV. IFIP World Computer Congress, 31 August – 4 September 1998, Vienna/Austria and Budapest/Hungary*, pages 429–442, Vienna, 1998. Österreichische Computer Gesellschaft.

[15] Donna Harman, Peter Schäubele, and Alan Smeaton. Document processing. In Giovanni Battista Varile and Antonio Zampoll, editors, *Survey of the state of the art in human language technology (http://www.cse.ogi.edu/CSLU/HLTsurvey/)*. Centre for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, 1996.

[16] William R. Hersh. *Information retrieval: a health care perspective*. Springer, New York, 1996.

[17] Julia Hodges, Shiyun Yie, Ray Reighart, and Lois Boggess. An automated system that assists in the generation of document indexes. *Natural Language Engineering*, 2:137–160, 1996.

[18] Ray Jackendoff. *$\bar{X}$ syntax: a study of phrase structure*. MIT Press, Cambridge, Mass, 1977.

[19] Leonardus Laeyendecker. *Brengt de vooruitgang ons verder?* Ten Have, Baarn, 1986.

[20] David D. Lewis, W. Bruce Croft, and Nehru Bhandaru. Language-oriented information retrieval. *International Journal of Intelligent Systems*, 4:285–318, 1989.

[21] D.A.B. Lindberg, B.L. Humphreys, and A.T. McCray. The unified medical language system. *Methods of Information in Medicine*, 32:281–291, 1993.

[22] Nicolaas J.I. Mars. What is an ontology? In Alex Goodall, editor, *The impact of ontologies on reuse, interoperability and distributed processing*, pages 9–19. Unicom, Uxbridge, Middlesex, UK, 1995.

[23] Sabine Martin and Günter Bergerhoff. Chemical abstracts online: a study of the quality of controlled terms. *Journal of Chemical Information and Computer Sciences*, 31:147–152, 1991.

[24] Michael J. Mauldin. *Conceptual information retrieval. A case study in adaptive partial parsing*. Kluwer Academic, Boston, 1991.

[25] Erik Oltmans. A two-stage model for robust parsing. In Chadia Moghrabi, editor, *Proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP+IA'98)*, pages 233–239, Moncton, New Brunswick, Canada, 1998. GRÉTAL, Université de Moncton.

[26] Geert J. Postma, B. van Bakel, and G. Kateman. Automatic extraction of analytical chemical information. system description, inventory of tasks and problems, and preliminary results. *Journal of Chemical Information and Computer Science*, 36:770–785, 1995.

[27] Alan F. Smeaton. Information retrieval: still butting heads with natural language processing? In M.T. Pazienza, editor, *Information Extraction - A multidisciplinary approach to an emerging information technology*, pages 115–138. Springer, Berlin, 1997.

[28] Karen Sparck Jones. Summary performance comparisons TREC-2, TREC-3, TREC-4, TREC-5, TREC-6. In Ellen M. Voorhees and Donna K. Harman, editors, *The sixth text retrieval conference (TREC-6)*, pages B–1 – B–8, Gaithersburg MD, 1998. U.S. Department of Commerce, National Instititute of Standards and Technology.

[29] Laudy E.H.M. ter Haar, Ivana Korbayová, Paul E. van der Vet, and Toine Andernach. Use of domain knowledge in resolving pronominal anaphora. *Belgian Journal of Linguistics*, 10:12–35, 1996.

[30] Bas van Bakel. *A linguistic approach to automatic information extraction*. Ph.D. thesis, University of Nijmegen, The Netherlands, 1996.

[31] Bas van Bakel, Reinier T. Boon, Nicolaas J.I. Mars, Jeroen Nijhuis, Erik Oltmans, and Paul E. van der Vet. Condorcet annual report. Technical report UT-KBS-96-12, University of Twente, Enschede, The Netherlands, September 1996.

[32] Paul E. van der Vet and Nicolaas J.I. Mars. Bottom-up construction of ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 10(4):513–526, 1998.

[33] Paul E. van der Vet and Nicolaas J.I. Mars. CQE: a query engine for coordinated index terms. *Journal of the American Society for Information Science*, forthcoming, 1999.

[34] Ellen M. Voorhees and Donna K. Harman. Overview of the sixth text retrieval conference (TREC-6). In Donna K. Harman, editor, *The sixth text retrieval conference (TREC-6)*, pages 1–24, Gaithersburg MD, 1998. U.S. Department of Commerce, National Instititute of Standards and Technology.

# Information retrieval: how far will *really* simple methods take you?

Karen Sparck Jones
Computer Laboratory, University of Cambridge
New Museums Site, Corn Exchange STree
Cambridge CB2 3QG, England
*sparckjones@cl.cam.ac.uk*

## Abstract

This paper considers document (text) retrieval techniques and their relations with natural language processing (NLP). It summarises retrieval requirements, illustrates current approaches to retrieval using a probabilistic model, and shows why these strategies are effective though they do not invoke conventional NLP. The paper then argues that the statistical methods used for retrieval have potential applications in conjunction with NLP for other important tasks, taking automatic summarising as an example.

**Keywords:** Document / text retrieval, statistical methods, probabilistic model, natural language processing, summarising

## 1 Introduction: the task

Document, or text, retrieval is an important form of information retrieval. There are many variations in the task including, for example, recovering an imperfectly remembered but previously seen item; or seeking items for which some relatively unambiguous, non-content specification like an author's surname can be given. But the more important and more challenging case is where the user is seeking documents about some *topic* without prior knowledge of what there is in the file, or reliable pointers to it.

*Adhoc* searches of this kind may alternatively start from a given known document, seeking others like it. They may also subsume *iterative* searches in which both user need and search query are modified on the basis of what is retrieved each time. Iterative searching on some given occasion has much in common with topic-based information *filtering* and routing over a period of time: but I shall concentrate initially on adhoc searching and return to filtering later.

## 2 Retrieval essentials

Adhoc searching has two key features. First, however much a user already knows, they are by definition 'ignorant' or they would not be seeking more information. Second, whether or not the user's information *need* is expressed as a question, e.g.

`Is prefabricated furniture manufactured in developing countries?`

it should not be assumed that, even if they system was intelligent enough to answer the question, this would necessarily suffice. The actuality in many cases, and thus the default presumption, is that the user wants to find out about something and therefore wishes to be provided with materials on it for their own review and analysis. This is required to allow them to properly relate any new information to what they already know. The user's linguistic input is taken as a *request* for documents giving information on, i.e. about, the user's topic, for the user to read: e.g. documents on

`prefabricated furniture manufacture in developing countries`.

The frequently-made assumption is that since the user's request is a linguistic object - phrase, sentence, known document, i.e. a text or discourse of some sort, and each document in the file is also a linguistic object, natural language processing (NLP) is required for effective retrieval. Thus given that, as illustrated above, the input topic is a complex relational concept and, further, that this can be expressed in different ways, syn-

tactic and semantic text analysis is needed both to determine the nature of the topic and to allow one expression of it to be related to another. We suppose, for instance, that the meaning of the example topic is found to be

'the making of household chairs etc that are in kit form when done in nations that are not

economically advanced',

and that this allows the original expression of the topic to be seen as equivalent to, e.g.

`manufacturing prefabricated furniture in countries that are developing`

or

`developing country manufacture of furniture which is prefabricated`.

Adopted wholeheartedly, this view has an important consequence. It implies the ability to identify not just the presence of a topic in a document, but that a document is *about* a topic, if not exclusively at least to a substantial extent. Thus when the retrieval file contains full text documents as opposed to titles or abstracts, the implication is that document *index descriptions*, indications of what a document is about, are summaries. But making summarising a condition of indexing is a demanding strategy.

The fact that modern retrieval systems apply wholly different approaches to indexing and searching is not, however, because we cannot deliver good automatic summaries. These methods deliberately eschew syntactic and semantic processing of the conventional kind, and rely heavily on word-based statistics. They have been developed and tested, over a long period and on a large scale, in comparative experiments which have shown both that they are effective in their own right and that they perform as well as methods exploiting NLP (TREC, 1993-1998). From the NLP point of view these methods are astonishingly simple. But though it seems surprising that NLP-based indexing does not do better, the established statistical techniques are in fact well suited to the retrieval task. They are also more sophisticated than they at first appear. So their successful application in retrieval may have lessons for other language-based information processing tasks.

## 3 Statistically-based retrieval methods

Statistically-based retrieval strategies are underpinned by formal models, notably the vector space, probabilistic, and inference net models (Salton and McGill, 1983; Sparck Jones et al., 1998; Turtle and Croft, 1991). At the foundational level there are real differences between these. But as they all have to work from the same basic data and, for the good reasons detailed later, exploit these in much the same way, practical implementations, especially of the vector space and probabilistic approaches, look much the same. For present purposes it is not necessary to elaborate on the underlying theories. I shall treat one approach, the probabilistic one, as an example and, taking the theoretical account presented in Sparck Jones et al. (1998) for granted, use the following summary account of the form in which the model is practically applied (see e.g. Robertson and Sparck Jones, 1994/7; Okapi, 1997) to illustrate the key features of modern retrieval.

### 3.1 An illustration

We start with simple index terms, say single words. Requests and documents are represented, for indexing, just as sets of terms. In particular, a search *query* is a set of terms. These terms are *weighted* using three types of information, to determine the *matching score* between the query and a document. The first type of information is the frequency of a query term in a document, `TF`. In general, the more frequently the term occurs in a document, the more likely the document is to be about the query topic and hence *relevant* to the user's need. The second type of information is the document length, `DL`. Given two documents with the same length, if the query term is more frequent in the first than the second it is reasonable to conclude the first is more likely to be relevant than the second. But if the first document is longer than the second, the greater term frequency may be attributed to greater verbosity and not be especially useful. We therefor modulate term frequency by document length.

The third type of information is collection frequency, CF, the number of documents in which the query term occurs. Since retrieval is about selecting the (typically few) relevant documents from a mass of non-relevant ones, terms with high collection frequency are poor selectors, Further, their occurrence in a particular document may follow simply from this general high frequency. Collection frequency should thus be exploited inversely, and a query term's weight for a particular document reflect the extent to which TF and CF are interestingly related: a good query term, i.e. one to which a high weight is assigned, is one with high TF (subject to DL) and low CF, i.e. with high inverse CF as naturally represented on the logarithmic scale.

Taking these three types of information together thus produces a formula for the *combined weight*, CW, of a specific query term i occurring in a specific document j as follows:

```
CW (i,j)  =  [ CFW (i) * TF (i,j) * (K1+1) ] /

                    [ K1 * ( (1-b) + (b * (NDL (j)) ) ) + TF (i,j) ]
```

In this formula CFW defines a CF-based weight. Given N, the total number of documents in the file and (for later convenience) relabelling CF as n:

```
CFW (i)  =  log N - log n
```

NDL is a normalised document weight defined as:

```
NDL (j)  =  (DL (j)) / (Average DL for all documents)
```

K1 and b are tuning constants.

Query-document matching scores are therefore not based just on whether query terms are present in a document, but are the sum of the weights of the present terms. As discussed further below, there is no requirement to match on all the given query terms. But the use of weights allows more refinement than simply counting the number of matching terms, since a match on fewer, but good terms, can give a higher score than a match on more, but less good ones. The output of a search is thus a list of matching documents, *ranked* by score and, hopefully, with highly ranked documents relevant to the user's need.

The basic scheme just described can be modified in important ways. First, by using a fourth type of information: where requests are long (as they might be with example documents), the weighting formula can be adjusted to take account of the fact that terms may have different within-query frequency, since more frequent terms are presumably more valuable. This is done simply by multiplying the CW formula by QTF. The more significant modification is where there is some information about known relevant documents, as in iterative searching. Given this fifth type of information, individual terms can be characterised not just by their simple collection frequency, but by their respective relevance and non-relevance frequencies, and by the former, RF, in particular. Thus we first define a relevance weight RW for each query term i:

```
RW (i)  =  log [ ( (r+0.5)(N-n-R+r+0.5) ) / ( (n-r+0.5)(R-r+0.5) ) ]
```

where N and n are as above, R is the number of known relevant documents and r (i.e. RF) is the number of known relevant documents for the term, and 0.5 is included to allow for predictive uncertainty.

Then for iterative searching we can replace CW by an iterative weight CIW, defined as:

```
CIW (i,j)  =  [ RW (i) * TF (i,j) * (K1+1) ] /

                    [ K1 * ( (1-b) + (b * (NDL (j)) ) ) + TF (i,j) ]
```

Finally, it is reasonable to assume that other terms for which relevance information is available may be valuable additions to the query, given the variable ways that topics can be expressed. Query *expansion* increases the chance of matching. Thus if all the terms occurring in the known relevant documents are pooled, they can be ranked by offer weight, OW, i.e. by

```
OW (i) = r * RW (i)
```

and the top, e.g. 10 or 20, terms added to the request. The query terms are then weighted using CIW. This process is using a sixth type of information, which we can label co-relevance frequency, CRF. (It is of course also possible to combine QTF and CIW, in the obvious way.)

The effect on practice for the example topic might be that a term with low CF, typically high TF, and high RF, say prefabricated, will score very highly, perhaps ensuring that a document containing only prefabricated and furniture scores more highly than one containing furniture, manufacture, developing and countries.

It must be emphasised that the theory given in Sparck Jones et al. (1998) motivates this specific set of formulae: they are not just pulled out of the air as more or less plausible. Equally, it must be emphasised that applying them delivers a good standard of performance, compared both with other intuitively obvious techniques and with ones using NLP. Further, the other statistically-based models give very comparable performance, showing that it is right to use these types of information for retrieval purposes.

## 4 Topic and content

From the point of view of topic representation and (document) content capture, the essential feature of the modern approach to retrieval is most clearly seen in the simple unweighted case, and where there are more than one or two terms in the initial query, e.g.

cheap prefabricated furniture manufacture.

The key to modern retrieval is the *best match principle*. While, ideally, one would like to match all of the query terms, this is unlikely to be achieved in practice. Moreover, while any document matching on all terms where many are given is highly likely to be relevant, there will normally be other documents that are relevant although they do not match on all terms. Failure to get complete matches follows naturally from the uncertainties that characterise the whole situation:

1. partial expression of need (users don't know so much);

2. indirect access to content (words only say so much);

3. variable presentation of content (people don't coincide too much).

These uncertainties cannot be overcome by greater control aimed at fully explicit, normalising topic/content representation. This is most obvious in the first and second cases, but also applies to the third. Thus being about a topic is an essentially loose notion: there is no one correct way of being about something. Content representation and matching have therefore to be on a 'do as best you can' basis.

But while this tolerance, or hospitality, may seem attractive from one point of view, the obvious points against it are that matching on sets of terms does not guarantee that the correct term senses will be selected or, more significantly, that the correct structural relations between terms will be captured. We are all familiar with the claim that the strategy described is inadequate because it cannot distinguish a blind venetian from a venetian blind.

The reality is different. Just matching on several terms at once normally succeeds only for appropriate senses and the required structural relations, or at least for broad approximations to these. Conjoint word matching de facto reflects suitable sense selection conditions and semantic relationships. For example,

cheap prefabricated furniture manufacture,

though it may match documents that are about cheap furniture or cheap manufacture, is unlikely to encounter and hence match documents that are about the vulgar [i.e. cheap] contents [i.e.furniture] of someone's mind. This is clearly less likely when either some small subset of a large initial set of terms matches, e.g. cheap and manufacture, or when there are only a few terms to start with, e.g. developing and countries, which may match on quite independent occurrences of countries and developing. There are nevertheless successful matches even in these conditions, specifically ones which capture the intended concepts (and, we assume, therefore also retrieve relevant documents). The simple set model is indirectly capturing real complex concepts and finessing the need to specify these concepts explicitly, in full detail. Further, while matches may sometimes be inappropriate, retrieval is usually directed at obtaining sets of documents (say to cover a subject thoroughly), which means that while some matches may be duds, the user benefits from other successful ones. However the fact that in modern situations, e.g. World Wide Web searching, requests typically start with, maybe, only two terms suggests that, whatever other problems these short

requests present, they offer little scope for NLP.

The more important point about partial matching is the real flexibility it offers: this is to permit retrieval even when the language or perspective of request and document are slightly different. Specifically, the set-based approach, especially where there are many query terms, as with expanded queries, assists with both precision and recall: the former by offering alternative forms of multi-term matching, the latter by achieving at least some match.

### 4.1 Key concept emphasis

The points just made apply to the basic unweighted case. But many tests have demonstrated that very large performance gains can be obtained with weighting, particularly when searching full text. Weighting clearly allows partial matching, just as the simple case does, but it does far more as well. The specific purpose of weighting is to meet the need for documents that are primarily, or substantially, about the query topic. Weighting signals that a term is *important*, because frequent, in a document. We do not know what it means, but this does not matter. Term weighting thus picks out the main concepts in the text and, where a high matching score is achieved, incidentally marks the fact that the query topic is important for the text. It does this somewhat crudely, just as set matching is a rather crude way of capturing structure. But it is sufficient for the purpose, even if it only generally, but not universally, holds. It also works in conjunction with partial matching.

### 4.2 Wider application

The successful approach to adhoc document retrieval just described can be extended to nearby related tasks. With filtering, for instance, as TREC has shown (TREC, 1993-1998), it is possible to apply the model very effectively when a large amount of relevance information is available. The same generic approach can be applied to passage retrieval, or short text extraction, and to the construction of new, hypertext documents (Salton et al., 1997). It has also been applied successully to other languages including character rather than word-based ones (e.g. Chinese in TREC), and to spoken documents, which may have distinctive discourse styles (as in broadcast news) as well as recognition errors.

The more interesting question is whether these strategies can be adapted to automatic document transformation and reduction, as in summarising, as opposed to document selection. I shall return to this after considering the role of NLP in retrieval in more detail.

## 5 NLP in retrieval

In spite of the points made in the last section, it is widely believed there is a manifest need to apply NLP for term and topic specification in retrieval, and attempts have been made to do this at least since the sixties (Sparck Jones, in press *a*).

The most basic, but also a very appropriate, use of NLP has been for word stemming, since there is good (though not universal) evidence that conflating e.g. singular and plural noun forms, or verb variants, is helpful in promoting recall without loss of precision. Stemming has been very effectively implemented with a procedure like Porter's (Porter, 1980), which uses a suffix list but no stem dictionary. It has been most widely exploited for English, but has also been applied to other languages, e.g. Spanish in TREC. In the original topic example, for instance, it could deliver the term `develop-` which would match `developing`, `developed`, `develops`, etc. Compounding languages require more effort for word decomposition and a reference lexicon, as well as more complex index management; but there is no reason in principle why the statistical methods described earlier should not be applied to the constituents of compounds.

This single term application of linguistic processing is nevertheless only minor, as well as uncontroversial. The belief in the value of NLP refers to a more ambitious application, to derive complex index descriptions. However the evidence from conventional manual indexing, where whole documents may be summarily indexed by complex proposition-like descriptions has shown that these are too demanding for effective retrieval, regardless of whether available NLP techniques can deliver them: see e.g. Keen (1973). Thus the potential use of NLP for retrieval has focussed on the identification and exploitation of compound terms, as represented by e.g. nominal groups. The argument here is that the components of topics may themselves be compound concepts, though it is sensible, in the interests of flexibility, to allow them to be used within the same set-based method of dealing with whole topics as is the norm for simple terms.

There are thus two issues: one is whether *com-*

*pound terms*, often called *phrases*, are really effective; and the other is whether, if they are used, they have to be defined linguistically as opposed to statistically. The argument for the linguistic approach is illustrated by Strzalkowski (1994): parsing identifies head-modifier units, which may then be conflated, for the same reason as with word variants, and also stemmed, so that e.g. the original `furniture manufacture` will also match `manufacture of furniture`, `furniture manufacturing` and `furniture that is manufactured`. Unfortunately, while there is some support for the claim that phrasal terms contribute usefully, if not spectacularly, to retrieval performance, there is no evidence that there is any gain from using NLP as opposed to statistical techniques (Mitra et al., 1997). In fact, just as with the set combination of single terms, a pair of words that recurs in a file is likely to represent a genuine compound, which can be identified by this recurrence without the need for parsing. This strategy could, for instance, pick up the adjacent pair `prefabricated furniture` along with, after stemming, `prefabricating furniture`, and also `furniture prefabrication`. (In line with the general need for flexibility, the elements of phrasal terms are used in requests as well as the phrases: the latter tend to have higher weights, so are more valuable, but the former contribute to matching too).

## 6 Lessons for NLP

The main lessons to be learnt from experience with the document retrieval task are first, the fact that at least some language-using tasks can be effectively done in a very crude way. This is not only because they are actually rather undemanding tasks. The second and more important lesson is that these tasks can be effectively done in a crude way because language, i.e. discourse, is *redundant*. The whole approach depends on repetition and reinforcement, with the same concepts (words) figuring in multiple slightly different ways in a text. It is no accident that frequency signals importance. NLP workers have always recognised that corpus data is valuable e.g. for grammar refinement or lexicon tailoring, precisely because it captures significant differences, marked by relative frequency of coccurrence and cooccurrence, in usage. However the point that emerges from the experience of statistically-based retrieval is that is that it may be worth trying to extend the application of statistical methods further in building tasks systems. In particular, retrieval suggests that statistical data should be exploited not just to improve the knowledge resources that are applied, but in a stronger sense, motivated by some underlying model such as the probabilistic one, to drive task processes.

Automatic summarising is a critical test for this. Attempts in the past to derive summaries by extracting and concatenating source text sentences, chosen using word frequencies, have not been convincing (Paice, 1990). But as summarising comes in many flavours for many purposes (Sparck Jones, in press *b*), there may be a role, for instance, for 'semi-summaries' consisting of statistically-grounded phrases. This would not just be a re-run of retrieval methods, first because phrases would be chosen to be representative of the individual document alone, not to discriminate one document from others. More importantly, it would be necessary to combine NLP of the conventional kind with the use of frequency data: first because when confined to a single document it is necessary to analyse it more carefully than for retrieval, since it is not possible to rely so heavily on file information to identify important phrases; second to capture genuine phrasal variants as fully as possible; and third because for individual documents it is desirable to find long phrases, since these are more informative for the reader and there is no need to consider the matching constraints that make shorter phrases more helpful in the retrieval case.

The kinds of possibilities that are beginning to be explored are illustrated in e.g. Mani and Maybury (1997, in press). Some Cambridge work also shows how the general suggestion that NLP and statistical methods might be helpfully combined can be applied at different levels of analysis: more simply at a very shallow level, or at a somewhat deeper level, with less reliance on the form of source expressions.

In a modest study along the first lines, Dersy (1996) used robust parsing to identify e.g. noun groups, conflation of variants, phrase selection based on occurrence frequency but constrained by rules designed to avoid phrases with heavy lexical overlap, and generation of an output list of phrases in preferred surface form and in text order. Tucker's recent work (see Sparck Jones, in press *b*), on the other hand, seeks to escape surface linguistic expressions and make more explicit use of the discourse relationships between source text concepts. Thus in his approach, sentences are parsed into logical forms from which primitive

predications are extracted. These are connected e.g. by predicate or argument sharing and thus form a cohesion graph embodying the overall attentional structure of the source text. Operations on the graph are designed to select subsets of nodes taken to represent the main source content, applying criteria e.g. for salience, representativeness and coverage of the original. Predications in the selected subset are merged where possible, by using e.g. common arguments, and and an output list of surface phrases is then generated to provide, as with Dersy, an indicative summary. In contrast to Dersy, however, the merging can lead to complex and thus more informative, phrases. Evaluation work, though difficult and hence limited, suggests this type of approach to summarising can be of considerable value for browsing and preview purposes, though such summaries cannot take the place of 'proper' condensation and reformulation, based on source text understanding and delivering a coherent output text.

Similar methods would also appear to be applicable to information extraction, as general strategs designed to identify key source text elements without relying on prior specification of what is required (e.g. terrorist incidents, company merger information).

Current NLP techniques for robust parsing are sufficiently effective to act as one of the necessary supports for identifying important types of text unit. My claim here is that they may have a natural partner in current retrieval techniques. These have become well-established in their own area: the nature of their success there suggests they may be more widely relevant to other language-processing tasks than hitherto supposed, so further research on combining them with NLP is well warranted.

## References


Dersy, J. (1996) Unpublished MPhil Dissertation, University of Cambridge.

Harman, D.K. (1996) Evaluation techniques and measures. *The Fourth Text REtrieval Conference (TREC-4)*, (Ed. D.K. Harman), Special Publication 500-236, National Institute of Standards and Technology, Gaithersburg, MD, 1996, A6-A14.

Keen, E.M. (1973) The Aberystwyth index languages test. *Journal of Documentation*, 29 (1), 1-35.

Mani, I. and Maybury, M. (Eds. (1997) *Intelligent scalable text summarisation*, Proceedings of a Workshop, Somerset NJ: Association for Computational Linguistics.

Mani, I. and Maybury, M. (Eds.) *Advances in automatic text summarisation*, Cambridge MA: MIT Press, in press.

Mitra, M., Buckley, C., Singhal, A. and Cardie, C. (1997) An analysis of statistical and syntactic phrases. *Proceedings of RIAO-97, Computer-Assisted Information Searching on Internet*, Centre de Hautes Etudes Internationales d'Informatique Documentaires, Paris.

Okapi (1997) Papers on the Okapi system, Special Issue of *Journal of Documentation*, 33, 3-87.

Paice, C.J. (1990) Constructing literature abstracts by computer: techniques and prospects, *Information Processing and Management*, 26 (2), 171-186.

Porter, M.F. (1980) An algorithm for suffix stripping. *Program*, 14, 130-137.

Robertson, S.E. and Sparck Jones, K. (1994/7) Simple, proven approaches to text retrieval. Technical Report 356, Computer Laboratory, University of Cambridge. (http://www.cl.cam.ac.uk/ftp/papers/reports/TR356-ksj-approaches-to-text-retrieval.ps.gz)

Salton, G. and McGill, M.J. (1983) *Introduction to modern information retrieval*, New York: McGraw-Hill.

Salton, G., Singhal, A., Mitra, M. and Buckley, C.. (1997) Automatic text structuring and summarisation. *Information Processing and Management* 33 (2), 193-207.

Sparck Jones, K. (in press *a*) What is the role of NLP in text retrieval? In *Natural language information retrieval* (Ed. T. Strzalkowski), Dordrecht: Kluwer.

Sparck Jones, K. (in press *b*) Automatic summarising: factors and directions. In *Advances in automatic text summarisation* (Ed. I. Mani and M. Maybury), Cambridge MA: MIT Press. (http://xxx.lanl.gov/cmp-lg/9805011)

Sparck Jones, K. Walker, S. and Robertson, S.E. (1998) *A probabilistic model of information*

*retrieval: development and status*. Technical Report 446, Computer Laboratory, University of Cambridge. (http://www.cl.cam.ac.uk/ftp/papers/reports/TR446-ksj-probabilistic-information-retrieval.ps.gz)

Strzalkowski, T. (1994) Robust text processing in automated information retrieval. *Proceedings of the 4th Conference on Applied Natural Language Processing*, Somerset NJ: Association for Computational Lingustics, 168-173.

TREC (1993-1998): D.K. Harman (Ed.) *The First Text REtrieval Conference (TREC-1)*, Special Publication 500-207, National Institute of Standards and Technology, Gaithersburg, MD, 1993; *... Second ... (TREC-2)*. SP 500-215, NIST, 1994; *... Third ...(TREC-3)*, SP 500-225, 1995; *... Fourth ... (TREC-4)*, SP 500-236, 1996; Voorhees, E.M. and Harman, D.K. (Eds.) *... Fifth ... (TREC-5)*, SP 500-238, 1997; *... Sixth ... (TREC-6)* (1997), SP 500-240, 1998.

Turtle, H. and Croft, W.B. (1991) Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9 (3), 187-222.

# Cross-Language Information Retrieval: Some Methods and Tools

Raymond Flournoy
Stanford University
Department of Computer Science
flournoy@csli.stanford.edu

Hiroshi Masuichi
Fuji Xerox Co., Ltd.
Corporate Research Labs, Kanagawa
masuichi@csli.stanford.edu

Stanley Peters
Stanford University
Department of Linguistics
peters@csli.stanford.edu

## Abstract

We describe two related methods of cross-language information retrieval which share a new approach that does not involve machine translation of queries. We compare their strengths with methods that require translating queries.

**Keywords:** Language Technology, Cross-Language Information Retrieval

## 1 Introduction

One approach [1] to crosslingual information retrieval (CLIR) is to convert a query Q posed in language L into a query Q′ in another language L′ by means of machine translation (MT), and then use monolingual information retrieval (IR) on Q′ to retrieve documents written in L′. In this paper, we describe two methods for CLIR which differ from this standard approach. In these new methods, we transfer across languages the specification of relevant documents which IR computes from a query, rather than translate the query itself.

Recall as background that IR is traditionally regarded as comprising two stages: (1) indexing a collection of documents by some method, and (2) responding to any query Q by (i) converting it to a search specification S(Q), then (ii) searching for documents whose index meets the specification—or rank ordering documents according to how close they come to meeting S(Q).

In CLIR, where one must retrieve documents written in a different language than the query, there are always different languages L and L′ involved, and typically also collections of documents in each of these languages. We propose to perform CLIR by (i) converting Q to its search specification S(Q) for documents in L, (ii) mapping this to a corresponding search specification S′(Q) for documents in L′, and (iii) searching for documents in the L′ collection whose index meets this specification S′(Q) as nearly as possible. The first of these steps is identical to the first half of search in monolingual IR in L, and the last is identical to the second half of search in monolingual IR in L′. We need to show how the intermediate step of mapping a search specification S(Q) for Q in L to a corresponding specification S′(Q) for search in L′ can be performed successfully. This paper discusses two different methods we are investigating.

An important advantage they share over the customary approach to CLIR is that removing MT from the search process avoids introducing a considerable set of errors MT brings with it. We provide some evidence that converting search specifications across languages loses less information—introduces less 'noise'—and is accordingly more accurate than translating queries by MT.

## 2 Converting Queries in Concept Space

Many information retrieval systems are based on keyword search, in which the target texts are searched for exact matches with the keyword strings. This leads to the familiar failures of retrieving too many documents, because the keyword or keywords are used too widely, or retrieving too few documents, because the desired documents do not contain the exact literal keywords

used in the query.

Keyword-based search suffers from excessive dependence on the specific words used to query the system. Instead, we want the system to see a query not as a collection of words, but as the ideas or concepts represented by the words. Furthermore, we would like the system to know how two different words or concepts are related, and to recognize when concepts are highly related or suggestive of one another. If a system could do this, it could take a query about *earthquakes* and would know that words such as *tremor* and *fault* are highly related to earthquakes, and therefore articles containing those words would be marked as possibly relevant to the query, even if the word *earthquake* never appeared in those articles.

The concept of relatedness which we would like to exploit is an extremely complex one, which goes beyond simple synonymy. Because of this complexity, having humans specify the relatedness of all concepts by hand would be at best impractical, and most likely impossible. Rather we would like to gather information automatically, to allow concepts to organize themselves according to the training data.

Our method for doing this is based on an approach first proposed by Hinrich Schütze [2]. In this approach, which we refer to as **information mapping**, we compile a list of, say, 1000 'content-bearing words' and for every word in the vocabulary we scan the training corpus and count the total cooccurrences between the vocabulary word and each content-bearing word. In this way, we produce for each word a list of 1000 numbers which represents the distributional behavior of the word. This list of numbers can be viewed as a 1000-dimensional vector within what we call a **concept space**, and our belief is that words with similar or related meanings will have similar distributional profiles, and thus will have vectors which lie close together within the 1000-dimensional concept space. We conduct the training, and automatically the words fan out into the concept space, clustering with other words which are related in meaning, and because of the extremely high dimensionality of the space, the concepts are able to organize themselves into complex webs of relatedness, all without human direction or intervention.

Once the concept space is built, we can then map articles and queries into the space. Executing information retrieval becomes the simple task of determining which article vectors are closest to the query vector S(Q) of the query Q within the concept space.

Since the concept space is meant to represent abstract concepts and not literal words, it is a natural step to extend this idea to multiple languages. The same region of the space would represent the same concept across languages, even if the concept is expressed in different words in different languages. For example, if a number of words related to politics, such as *election* and *president*, have clustered in one area, we would like to be able to place Japanese words such as *senkyo* and *seifu* into that same region.

We can extend the formalism to multilingual information retrieval by requiring that the 1000 content-bearing words which form the framework of the concept space represent the same concepts across languages. In other words, if two of the dimensions are represented in English by the words *economics* and *vaccine*, they would be translated into German as *ökonomie* and *impfstoff*. Training would proceed as before, but by counting cooccurrences with the proper forms of the content-bearing words. By doing so, we expect that articles about, for example, immunization among the poor would fall into the same area along these two axes of the concept space, regardless of whether the articles were written in English or German. This entails that the search specification $S'(Q)$ in $L'$ for a query Q in L can be identical to the search specification S(Q) in the query's 'source' language. It is important to note that we do not need to translate the raft of stored documents or even the query words into other languages, but only the small set of content-bearing words. We establish the same set of basic ideas to form the framework of the multilingual concept space, and training automatically populates the space with our full vocabularies, regardless of the language.

Information mapping's strength is its ability to characterize subtle relationships between words through the high dimensionality of the concept space. And since the space is built automatically through training on text corpora, this advantage does not come with prohibitive labor costs. We hope to leverage this strength further by abstracting to the conceptual level as much of the multilingual information retrieval task as possible, and allowing simple transfer of queries and articles across different language's versions of the concept space. Attaining this would achieve true 'information' retrieval, and not the simple keyword searching generally employed now.

## 3 Converting Queries in Mutual Information Space

A different method for converting search specifications from one language to another utilizes IR based on mutual information between documents to be retrieved and words in a subset of texts that constitutes the search specification. This method uses a parallel corpus to replace the specification S(Q) of a user's query Q in terms of mutual information in his or her language L with a set of texts S(Q′) in another language L′. We can then retrieve texts in the target language which are most relevant to the query, i.e., those with highest mutual information to words in S(Q′).

Below, we first explain the monolingual IR system and then show how to realize a multilingual text retrieval system. We close with a specific example that shows how this method eliminates problems that would result from attempting to translating the query itself.

### 3.1 Monolingual Text Retrieval by Mutual Information

This monolingual text retrieval system requires as input a subset of the whole collection of texts in the database. The system ranks all texts in so that the more related a text is to the input, the more highly the text is ranked. The following is the algorithm for the system.

1. Make a list of all the word types $W_1, W_2, \ldots, W_n$ which appear in input texts.

2. Calculate the following measure of 'mutual information' for each word $W_i$ $(1 \leq i \leq n)$.

$$\mathrm{mi}(W_i) = \log \frac{P(W_i, S)}{P(W_i)P(S)} \qquad (1)$$

where
$P(W_i, S) = \frac{a_i}{M}$
$P(W_i) = \frac{b_i}{M}$
$P(S) = \frac{s}{M}$
$M$ = the number of texts in the whole database
$a_i$ = the number of input texts that contain $W_i$
$b_i$ = the total number of texts that contain $W_i$
$s$ = the number of input texts

3. Calculate the following summation of 'mutual information' for each text $T_j$ $(1 \leq j \leq M)$.

$$MI(T_j) = \sum_{i=1}^{n} F(i, j) \qquad (2)$$

where
$F(i,j) = \mathrm{mi}(W_i)$ if $T_j$ includes $W_i$
$= 0$ if $T_j$ does not include $W_i$

4. Sort the texts in the database according to their MI scores.

### 3.2 A Multilingual Text Retrieval System

With access to a parallel corpus, we can adapt this monolingual text retrieval algorithm to multilingual text retrieval. We apply this algorithm twice for this purpose. The first application replaces a source language query by a set of target language texts; the second obtains from them the desired set of result texts. Assume that the parallel corpus consists of English sentences with Japanese translations, and the input to the system is a Japanese sentence. Then the system ranks all the English sentences in the corpus according to relatedness to the input using the following algorithm:

1. Make a list of all the input sentence's Japanese words (other than stop words) $JW_1, JW_2, \ldots, JW_L$.

2. Extract from the parallel corpus all Japanese sentences that contain any word in the list.

3. With the Japanese sentences just extracted as input texts, use the monolingual text retrieval system described in the preceding section to rank Japanese sentences. The system's database texts should be all Japanese sentences in the parallel corpus.

4. Extract the $N$ most highly ranked Japanese sentences $J_1, J_2, \ldots, J_N$ and the parallel English sentences $E_1, E_2, \ldots, E_N$.

5. Rank all the English sentences using the monolingual text retrieval system again. The database texts of the monolingual system should be all the English sentences in the parallel corpus and the input texts to the monolingual system should be the English sentences just obtained in Step 4.

Note that in Step 5, the system can rank English sentences without Japanese translations. Thus

the database at this stage can be a larger collection of English texts than merely sentences appearing in the parallel corpus.

We carried out experiments with this algorithm using a corpus of 300,000 English sentences and Japanese translations.

## 4 Comparison

We spend the rest of this paper comparing our approach with some conventional multilingual text retrieval approaches. Consider, for example, the Japanese sentence

(a) *Shidai ni hoso-ku na-ru*

as a query. In this case, the English sentences most relevant to the input among the 300,000 English sentences are

(b) *It tapers down to a point.*

and

(c) *It tapers into a sharp point.*

### (I) Pattern Matching in Source Language

One could extract the Japanese words *shidai* and *hoso-i* from input sentence (a) using a Japanese morphological analysis system (omitting some stop words), as we do. Neither sentence (b) nor (c) can, however, be retrieved with the query (*shidai* and *hoso-i*), because the Japanese translations of (b) and (c) do not include *shidai* or *hoso-i*. This approach succeeds only when a corpus contains Japanese translations that are almost the same as the input sentence. Because one concept can be expressed in various ways and it is rare that nearly identical translations for an input sentence exist in a corpus, this approach fails in many cases.

### (II) Pattern Matching in Target Language with Word-Based Translation

¿From a Japanese-English dictionary we can obtain for *shidai* the English translations *gradually*, *by degree*, and *little by little*, and for *hoso-i* the translations *thin*, *narrow*, *fine*, *slim*, and *slender*. However, the query

(*gradually* or *by degree* or *little by little*) and (*thin* or *narrow* or *fine* or *slim* or *slender*)

will not retrieve (b) or (c), because neither (b) nor (c) includes any of these English expressions. In general, the meaning of a word differs considerably depending on its context. *Hoso-i* has the flavor of *taper*, as in (b) and (c), only when it cooccurs with *shidai* or other specific words; so word-based translation of a query is not enough to support accurate multilingual text retrieval.

### (III) Statistics in Source Language

The underlying assumption of statistical text retrieval approaches such as vector space models and probabilistic models is that related words tend to cooccur in texts. For example, a text that includes *weather* is expected to include the words like *rain*, *snow*, *dry*, etc. This assumption does not always hold when a text is very short, as in this experiment in where texts comprise a single sentence. In fact, the Japanese translations of (b) and (c) are not highly ranked by our monolingual text retrieval system, which is a probabilistic model, in Steps 3 and 4 of this algorithm.

Though the result produced in Step 4 does not include the translation of either (b) or (c), it includes sentences translating the following English sentences containing the word *taper*:

*American aid tapered off.*
*The purchase of California wine tapered off.*
*Her muscular legs tapered to slender ankles.*
*When they realized what he was saying, the applause tapered off uncertainly.*

This is because the Japanese translations of these sentences include either *shidai* or *hoso-i*. As a result, *taper* has a high 'mutual information' score and we find both sentences (b) and (c) in the top 10 sentences out of 300,000 by our multilingual text retrieval system.

## 5 Conclusion

In this paper we have introduced two methods for CLIR which eliminate the need for translation at run-time while reducing the amount of noise introduced by the cross-lingual component. In both methods, instead of translating the list of query words, we take a representation of the entire query and transfer that representation over to another language setting, where the actual IR search occurs. In one method, we transfer a query vector in concept space into an overlaid concept space in another language. In the other method, we use mutual information to find a representative set of sentences from a parallel corpus, and

this becomes our transfered query in the target language. Both methods avoid the problem of noise introduced through single-word translations and also avoid the need for translating the entire corpus to be searched. Our new approach enables us to adapt established, well-developed monolingual IR methods to CLIR by applying the first half of retrieval in them in the first language for CLIR and applying their second half in the second language as the last step of CLIR. As a result, the need for translation is simplified to finding corresponding content-bearing words in L and L′ or obtaining a high-quality corpus of parallel L and L′ sentences.

Naturally, the documents in L′ that CLIR returns can be translated to L if desired, either by MT or by a human translator; this subsequent step is equally possible for the standard approach and the new approach we have described.

## References


[1] Hull, David A. and Gregory Grefenstette. Querying across Languages: A dictionary–based approach to multilingual information retrieval. In *Proceedings of SIGIR*. pp. 49-57. 1996.

[2] Schütze, Hinrich. *Ambiguity in Language Learning: Computational and Cognitive Models*. CSLI Lecture Notes, No. 71. Stanford, California. 1997.

# Talking Pictures: Indexing and Representing Video with Collateral Texts

Andrew Salway and Khurshid Ahmad
Department of Computing, University of Surrey
Guildford, GU2 5XH United Kingdom
a.salway@surrey.ac.uk, k.ahmad@surrey.ac.uk

## ABSTRACT

The relevance of collateral texts for building knowledge-based visual information systems is discussed, with reference to moving images of dance. The knowledge acquisition technique Protocol Analysis is applied to elicit verbal reports from experts watching moving images with complex contents and interwoven meanings. These reports are analysed at lexical, clausal and discourse levels, using text analysis methods. The results show a potential for using these reports to index and represent moving images, through the creation of lexical resources and knowledge-bases. The KAB system integrates text analysis modules with a video object database to process collateral texts.

**Keywords:** Knowledge-based Visual Information Systems, Video Indexing, Collateral Text, Protocol Analysis

## 1 INTRODUCTION

There is a need for technologies that can assist in the retrieval and presentation of visual information. These technologies must provide ways of attaching useful indices to both still and moving images so that they can be matched against user queries. Furthermore, representations of image content in knowledge-based visual information systems should address the fact that an image can mean many things to many people.

The inclusion of human knowledge may be crucial in domains where the perception and understanding of images is an expert task. The knowledge of experts is realised in the language of their written and spoken discourse. The texts produced by experts can be used as high-level expressions of image contents - from which indices and representations for computer systems could be derived.

Typically, image and video retrieval research has sought to compute indices from raw image data - attaching colour, texture and shape features to still images, and segmenting and selecting key-frames for moving images. In the field of computer vision, researchers have developed picture grammars which build up image descriptions in terms of primitives such as edges, corners and surfaces. These approaches work well in many cases when matching on *perceptual similarity* is required.

However, much visual information is complex, comprising interwoven strands of meaning that may be confounded in the image. In these cases, indices and representations of images need to be high-level and symbolic - to refer beyond the physical image contents.

Certain visual information may be best understood, and hence explicated, by the experts of a particular domain: consider, a surgeon examining an X-ray image; a meteorologist making predictions from moving images of weather systems; a scene-of-crime officer recording photographed evidence; and art critics and dance scholars who can elucidate meanings in complex images which would not be apparent to a lay person.

In these cases the words spoken and written by the experts about the visual artefacts will be high-level expressions of their contents. In order to exploit this fact for building knowledge-based visual information systems it is important to understand how experts articulate their knowledge, and how their articulations relate to visual information.

Researchers have already exploited textual information that co-occurs with everyday visual information. For example, Srihari reported how newspaper photograph captions were processed to constrain subsequent image analysis algorithms that detected the faces of people described in the caption [1]. She used the phrase ‘collateral text’ to denote textual information which related in some way to visual information. Recent research has used the textual component of video email and of news video for indexing purposes, see [2] for examples.

This paper considers how the written and spoken words of experts can be utilised as collateral texts to build knowledge-based visual information systems for specialised images. This question is explored in the exemplar domain of dance - chosen for its stylized moving images, diverse multimedia information and consolidated expert knowledge.

The work of dance scholars is first discussed to highlight the ways in which a moving image can be analysed, and to suggest how texts written by dance experts could be collateral to dance images (Section 2). An investigation is then reported in which spoken texts were elicited from dance experts as they watched dance videos; following the knowledge acquisition technique of Protocol Analysis. Their verbal reports were analysed for their lexical, clausal and discourse-specific behaviour. This analysis sought to explicate the ways in which the experts articulate their knowledge about moving images (Section 3). The findings of these investigations have been used to refine the specification of the KAB system. KAB (Knowledge-rich Annotation and Browsing) integrates digital video and collateral texts to give knowledge-rich representations of moving images (Section 4). The paper closes with some remarks about how collateral texts might be used in knowledge-based visual information systems with integrated video analysis functionality (Section 5).

## 2 Dance Experts' Discourse on Moving Images

Dance comprises stylized movements which are rhythmic and usually set to music. Recorded dance images provide a source of diverse, but interrelated, multimedia information about human movements, music, costume and the stage set. The dancing body may be described in terms of its parts, or itself may be described as a part of a greater whole – when groups of dancers form patterns. Dance can convey emotions, tell stories and make social comments and cultural statements. In dance, whether recorded images or live performance, there are interwoven strands of meaning.

### 2.1 Dance Analysis

There are specialists who study dance as an academic, and as an applied subject. They deal with the perceptual and cognitive aspects of dance and discuss dance in historical, cultural and political contexts.

Dance scholars have evolved notation systems for recording muscular movement (with similarities to musical notation). The emphasis here is on recording the muscular-skeletal movements of the dancers and their positions in space. Movement notation systems are expected to 'provide the key to relatively unambiguous communication through the creation of an agreed symbol system' [3]. An example of a prominent system, *Labanotation*, is shown in Fig. 1. The notation is read from bottom to top, along a vertical temporal axis delimited by bars akin to those of a musical score. Symbols to the left of the centre refer to movements made by the left hand side of the body: the foot, leg, torso, arm, hand and fingers - in that order. The symbols' points, shadings and size capture the movement dynamics of direction, level of extension and duration.

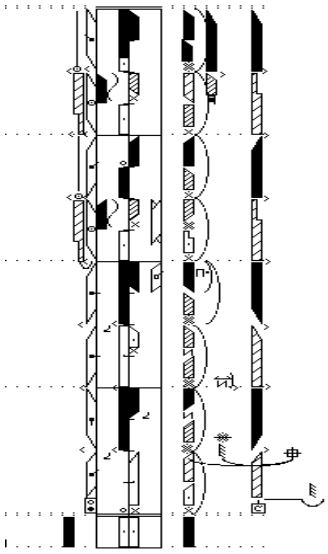


Fig. 1: An example of *Labanotation*

Movements are also described with the terminology of established dance genres, like classical ballet's *plié*, *relevé, attitude* and so on. The dance analyst then discerns how individual actions and gestures are combined in spatial and temporal forms, e.g. to show the interaction between dancers, or the recurrence of a *motif*.

It is problematic to separate the *objective* description of dance (movement durations, directions, accelerations) from its *subjective* facet (meanings and emotions). For, whilst a *slow movement* can be assessed in biomechanical terms, it might also be, depending upon contextual factors, a *tender movement* or indicative of *wandering*.

These movement qualities guide the dance scholar in explicating the narrative and intention of a dance. Individual dances and choreographers are considered and a view of genres and styles of dance is developed on a case-by-case basis. Thus, Judith Mackrell has

written that classical ballet tries to create 'the illusion of flight', and that some classical Indian dances 'are grounded on earth' [4]. The motivations for some modern American dance are given in comments on Martha Graham - that her dancing 'was based on the pull of gravity' [ibid.]; and on Merce Cunningham – that he wanted dance to 'reflect the dense information overload that we're used to processing every day in the modern world' [ibid.]. When making such interpretations, it has been argued, it is important to recognise that 'the conventions and traditions of the context, genres and styles presuppose and, therefore, prescribe specific ranges of *subject matter and the manner of treatment*' (emphasis original) [3].

The dance theorist draws on theoretical perspectives in order to place the dance in an historical, cultural and theoretical framework. Dance genres and styles may be used to classify a dance, and relate it to its forebears. Other political and cultural theories may be adapted, for example, Marxism, feminism and psychoanalysis. In the course of evaluation, reference is made to other dances as well as sources including books, films, historical events and cultural phenomena.

Janet Adshead-Lansdale has outlined a four stage method for analysing dance [3]. The method involves describing movements, discerning spatial and temporal forms, interpreting meanings and evaluating the aesthetic merits of a dance. This framework was used in the research reported here for focussing experts on particular aspects of moving images during knowledge acquisition. Similar frameworks have been used in other research concerning visual information: Erwin Panofsky's three levels of meaning in fine art [5] have been adapted by information scientists for picture classification [6]; and, Christian Metz's five levels of cinematic analysis [7] motivated a semantic data model of video [8].

## 2.2 DANCE TEXTS AND DANCE KNOWLEDGE FOR IMAGE RETRIEVAL

The knowledge used by dance experts to analyse and understand dance at perceptual, cognitive and aesthetic levels is relevant to the ways in which images of dance can be stored and retrieved. The content of a dance, from a retrieval perspective, may include isolated muscular movements, patterns formed by dancers, metaphoric references, e.g. to 'pull of gravity' or 'information overload', as well as historical, cultural and possibly political allusions.

For us, a verbal expression about a dance can be collateral to still and moving images of that dance. This collaterality can be found in the program of a dance performance, in a dance critic's newspaper review, in popular books explicating the works of choreographers and in learned journal articles.

The question is how to use these different types of texts for the task of labelling and retrieving dance sequences. A dance program could be used to label long stretches of movement, like acts and scenes (circa 15 - 20 minutes). A dance critic's output might highlight noteworthy aspects of a dance to be labelled, e.g. certain sequences of movement, the performance of certain dancers or the use of costume and sets. More learned writing includes the dissection of short movement sequences (*motifs*) as well as the principled grouping (*classification*) of dances or choreographers into genres and styles.

The information retrieval literature would suggest the building of a thesaurus, comprising terms related to dance movements, key historical events and even philosophical and political trends: these would then be used to label all or parts of a dance video - much as is practised in the keywording of journal articles and textbooks.

The different types of dance text discussed above are produced for information dissemination to a well-defined readership. The intention of the authors is not to index a dance for retrieval purposes. These texts can merely serve as mnemonics, placing the burden of interpretation of collateral words, phrases, clauses and whole texts, on the reader. Nevertheless, it is these texts that help their readers to see beyond the physical image: beyond the surfaces, edges and contrasts of light and shade. These texts have the potential for helping end-users of an image retrieval system by expanding and refining queries. For instance, these texts might be used to produce a specialist lexicon of dance, or of a particular genre of dance with contextual information about most of its lemma entries. Key-word-in-context analysis may help to build a lexical semantic network linking lemmas with arcs labelled by semantic relations like synonymy and hyponymy.

Though the extant texts of the dance domain are collateral to moving images of dance, the fact that they are written means they do not share a straightforward temporal relationship with their subject matter. Authors bring together examples of movement from different parts of a dance, or from different dances. The collaterality is further obscured by the changing focus of the text - in one paragraph the discussion might be of muscular movement, in the next it might turn to the historical building in which the dance was staged. Thus, whilst these extant texts may be collateral to dance images, they would pose problems if used in the initial development of a knowledge-based visual information system.

In contrast, a recorded running commentary on a dance sequence will preserve collaterality, whilst still being a linguistic artefact – amenable to subsequent analysis. To ensure that the commentary is knowledge-rich, i.e. it contains so-called domain objects and heuristics, it can be elicited from a domain expert. Finally, the expert should be instructed to focus on a particular aspect of the dance in a single recording. Protocol Analysis is a technique for prompting experts to articulate aspects of their knowledge in a *verbal report*.

# 3 VERBAL REPORTS OF DANCE SEQUENCES

Researchers in various disciplines have been concerned with how subjects talk and write about images. Firschein and Fischler elicited descriptions of aerial city photographs from subjects with a variety of tasks; their investigation concerned the descriptive representation of pictorial data for computer vision systems [9]. In information science, researchers have studied both how indexers attach keywords to images [10], and how searchers of images phrase their queries [11]. Linguists have recorded narratives from subjects about a film they watched; the transcripts became the basis for a cross-cultural study of discourse and the relationship between conscious experience and the spoken word [12].

The artificial intelligence and cognitive psychology literatures discuss knowledge acquisition techniques which can be used to elicit, analyse and represent aspects of human knowledge for use in computer systems. Knowledge engineers have access to two sources of knowledge: human experts and the texts the experts have produced. Knowledge can be elicited from experts through brainstorming, interviews, questionnaires and the like. Techniques to extract knowledge from text may be applied to the transcripts of such knowledge acquisition sessions, and also to texts extant in the domain. Text analysis can assist in knowledge engineering, with the elaboration of domain terminology leading to the modelling of concepts and then propositions and rules [13].

Protocol Analysis is a knowledge acquisition technique, in which an expert is asked to 'think aloud' as they perform a task: the resultant verbalization is taken to reflect their cognitive processes - and hence their expertise [14]. Protocol Analysis is used to access the steps that an expert takes in performing a task, i.e. to understand 'how' the expert does it. The expert's verbalization is recorded and becomes the object of investigation - hence the claim that verbal reports provide us with data about the subject's cognitive processes. There is a similarity with the research method of Content Analysis, which 'procedures to make valid inferences from text' [15].

In the current research, Protocol Analysis was applied to access an expert's perception and understanding of complex visual information in real time. The goal was to elicit verbal reports that (i) would help to understand the expertise used in analysing moving images; and, (ii) would serve as collateral texts for indexing and representing them.

Experts were prompted to talk about a moving image as they watched it: the instructions they were given before talking were used to focus their verbalization on particular aspects of the image contents.

## 3.1 METHOD

Five dance experts[1] were twice recorded speaking as they watched a video compilation of dance excerpts lasting 20 minutes. Four of the excerpts were duets, a fifth featured 12 dancers. The types of dance included neo-classical ballet and modern dance.

Before the first recording the expert was prompted with an instruction to 'Describe'[2] the dances, speaking as they watched. For the second recording the instruction was to 'Interpret'[3]. These instructions were worded by the co-ordinating expert, Prof. Janet Adshead-Lansdale, to reflect tasks familiar to the other experts.

The experts' verbalizations were recorded onto one of the sound tracks of the video cassette that they were watching; this maintained the temporal relationship between word and image. They were transcribed by one of the authors. The flow of spoken language was broken into speech fragments on the basis of the speaker's pauses and the perceived completeness of each speech fragment. A start-time (m:s) was manually inserted before each transcribed speech fragment. For the current research there was no need to transcribe intonation information and lengths of pauses, but this information might be valuable in future studies.

[1] A university lecturer and four post-graduate students from the same Dance Studies MA course at Surrey.

[2] The written instruction advised *'by describe we mean, focus particularly on the detail of the movement, its use of space and its dynamic emphasis'*.

[3] This time the elaboration was *'by interpret we mean, outline one or many kinds of significance you might attribute to the interaction in this section'*.

The verbal reports for the 'Describe' task, henceforth D-texts (for the sake of clarity in this paper), totaled 11,300 words, and 1600 speech fragments; from 100 minutes of speech. The most fluent speaker averaged a rate of about 150 words per minute, the least about 75: there was no significant variation in the rate of speech between different dance excerpts. Half the transcripts from the 'Interpret' task (I-texts) have been analysed, these total 6289 words.

## 3.2 Analysis of Verbal Reports

Initial analysis of the verbal reports manually examined the extent (i) to which the linear order of their contents matched those of the video; and, (ii) to which subjects agreed on what to speak about and how. Observation of aligned verbal reports suggested a reasonable correlation between their contents and the videos' contents. Table 1 shows samples from two D-texts (from different experts) and one I-text, corresponding to 40 seconds of an excerpt from Matthew Bourne's *Swan Lake*.

It is clear that in the D-texts the experts have chosen to focus on similar aspects of the video, however they describe them at different levels of detail, and using different words. For example, in describing the man walking, one expert noted *where* - 'across the stage area' and another noted *how* - 'at a very slow pace'; and the arms of the swans are 'extending .. outwards and behind their backs' (D-text 2) or 'locked behind their backs' (D-text 2).

The I-text in Table 1 says less about the individual actions, and more about the mood of the characters - 'a sense of loneliness' and 'looking longingly'; and, about their relationship - the man is 'separated from the rest of the characters'.

Generally the experts' expressions were observed to range from single words - naming individual actions and gestures almost as soon as they recognised them, to longer speech fragments – detailing the time, location and quality of the movements; and, in the case of the I-texts, making cases for their interpretations. The experts generally kept up with the moving image – their words tended to lag no more than a few seconds behind the subject matter. However, there were examples of experts referring back to earlier sequences, e.g. 'back to the pas de deux type lifting movement'; and of experts referring to longer sequences in their interpretations, e.g. 'their relationship to each other alters during the course of the piece.'

Table 1: Excerpts from two corresponding D-texts and one I-text

| **D-text 1** | **D-text 2** | **I-text** |
|---|---|---|
| [0:05] a single man walks across the stage area<br>[0:10] his back is to the audience<br>[0:11] he hugs himself<br>[0:13] he is surrounded by a group of dancers who are bent over from the waist<br>[0:18] extending their arms outwards and behind their backs into a cross shape<br>[0:25] they are looking upwards<br>[0:27] the central character who is a male who has walked across the stage wanders towards the audience looking around<br>[0:36] meanwhile the male group of dancers of about twelve are continuing to spread their arms and they are running around, arms undulating | [0:00] we see a big scene full of blue figures<br>[0:05] a man is walking at a very slow pace<br>[0:08] see a lot of back of people<br>[0:10] their arms are locked behind their backs<br>[0:16] they are actually higher than their backs<br>[0:17] and they gradually move their torsos up and they are standing<br>[0:22] they are all men with the ~left ~foot, left leg bent<br>[0:26] we see the man who was walking turning round and his face is looking upwards in a sort of romantic pose<br>[0:36] we see the arms of the men who wear the white pants | [0:07] the male character seems to be out looking at the moon, searching for something, he has a sense of loneliness and isolation about him<br>[0:14] the swan characters in the chorus are very earthy<br>[0:17] seem very still and calm in comparison to the man<br>[0:19] he is looking longingly at the moon<br>[0:25] he is unhappy, distressed, a bit soulful about something<br>[0:32] his mood is accentuated by the swans – we see them at one with their environment<br>[0:35] whereas he somehow seems dislocated<br>[0:38] being separated from the rest of the characters |

The verbal reports were analysed with regard to the distribution of their lexical items; the information content of their clauses; and cohesion phenomena. The analysis was performed with Surrey's text analysis package *System Quirk ©*. Results show how experts' knowledge for analysing moving dance images can be articulated to: name movements and their qualities; elaborate on gestures, actions and poses; highlight important sequences; and make interpretations.

### 3.2.1 Terminology for describing dance

A statistic which divides the relative frequency of a word in a collection of specialist language texts (SL) by its relative frequency in a general language sample (GL) gives a word list in which words peculiar to the texts rise to the top. Table 2 shows some words which appeared in the D-texts 100+ times relatively more

often than they did in a general language sample (10 million words of text from the Longman Corpus). The 25 words are from a specialist dance terminology; they constitute just under 20% of the D-text words with SL / GL > 100. Apart from the generally used terms *bodyshape, choreography, duet(s), footwork* and *motif*, Table 2 contains terms from a balletic vocabulary – perhaps reflecting the background of the experts, and the kinds of dance they spoke about.

Table 2: Dance Terminology in the D-texts

| |
|---|
| adage, adagio, arabesque(s), balletic, battement, batterie, bodyshape, choreography, développé(s), duet(s), footwork, jeté, motif, passé, penché, pirouette(s), planche, plié(ing), promenaded |

A second subset of words with SL / GL > 100, comprising a further 43 % of the total, is listed in Table 3. These are words which would be familiar to non-experts, but that have been appropriated, sometimes with a shift in meaning, by the dance experts; e.g. *pedestrian* which refers to everyday movement in this context. They are split into two groups: (i) nouns and verbs which denote movement and actions; and (ii) adjectives and adverbs which denote the quality of the movement and actions. The descriptions of quality at times cross the boundary into interpretation, e.g. *animalistic* and *robotic*.

Table 3: Preponderant general language lexical items in the D-texts

| | |
|---|---|
| **Movement & Action** | arching, balances, balancing, caresses, clasping, flicks, hops, interlocking, jumping, jumps, kneeling, leans, leaps, lunge, lunges, manipulates, manipulating, pivoting, pivots, pushes, quivers, rotates, slicing, spins, spiralling, spirals, splaying, stroking, sways, tilts, tottering, totters, twisting, twitching, undulations, weaved, wiggles, wraps |
| **Quality** | animalistic, dynamic, flexed, gestural, jerky, lyrical, pedestrian, rhythmical, robotic, stuttered, swirly, synchronised, undulating, unison, unisonal, unisonally, virtuostic, wiggly |

The mid-ranges of the SL / GL list were filled with more generally familiar words which describe movement and actions, e.g. *bend, come, hold, roll, kneel, walk*. Words that locate movements and actions in space and time also appear, e.g. *diagonal, forward, left, right, across* and *continuing, occasional, while, sporadic*.

The absolute frequency of words denoting body parts perhaps says something about how the dance analyst attends to movement - it may also say something about the genres of dance being described. The D-texts gave the following result, in descending order of absolute frequency: *arm(s), leg(s), hand(s), head(s), foot / feet, body/ies, back(s), shoulder(s), chest, neck, torso, waist, face(s), elbow(s), hair, knees, palms, spine*.

Compound terms were identified using a method that extracts strings occurring between lexical items given in boundary lists, giving, e.g. *corps de ballet, rond de jambe* and *pas de deux*.

Distinct collocation patterns involving the words *position, gesture* and *action* were noted - in each case a particular type of preceding word was seen to collocate about 50% of the time, Table 4. These examples might be considered to be 'semi-fossilized phrases' [16] - in which one word predicts a limited number of collocating words.

Table 4: Semi-fossilized phrases in the D-texts

| Nucleate | Total Freq. | Typical preceding word | Example Phrases |
|---|---|---|---|
| **position** | 68 | first – fifth (44%) | first position ... fifth position |
| **gesture** | 19 | 'BODY_PART' (53%) | leg gesture<br>head gesture |
| **action** | 18 | 'METAPHOR' (56%) | pendulum action<br>sawing action |

### 3.2.2 Clauses bearing information about *gesture, action* and *pose*

Classification of the speech fragments was made in terms of the information content of their clauses. Observation of the linguistic data suggested three categories of clause for this purpose: these were validated by the co-ordinating dance expert. *Gesture* clauses describe a spatial reconfiguration of body parts in relation to one another; *Action* clauses describe a spatial relocation of the whole body along spatial pathways, this includes locomotion; *Pose* clauses describe dancers' locations on stage, held positions and gazes.

A manual analysis of one D-text (80 speech fragments describing 333 seconds of dance) gave the distribution of clauses shown in Table 5 (a speech fragment may contain more than one clause). Twelve speech fragments did not fit the scheme; they referred to costumes or camera actions.

Table 5: Distribution of clauses by information content in one D-text

| Information Content | Freq. | Examples |
|---|---|---|
| ***Gesture*** | 27 | .. he hugs himself ..<br>.. his arms are undulating .. |
| ***Action*** | 26 | .. a single man walks across the stage area ..<br>.. they circle around each other .. |
| ***Pose*** | 27 | .. there is a group of four of them in the background ..<br>.. they are looking upwards ..<br>.. goes into an arabesque.. |

The even spread of clauses suggests that this is a useful classification for further analysis to be based on. Each clause can refer to 1, 2 or 3+ dancers - dancing in unison or taking different roles in an interaction. The contents of a clause can be modified by adjectives and adverbs to describe quality, and by prepositional phrases to situate them in space and time.

### 3.2.3 Lexical and syntactic cohesion

Halliday has described linguistic phenomena relating to cohesion and the creation of 'texture' [17]. Two textural characteristics of the D-texts are noted here. Some passages are marked by repetition of semantically-related words, exemplifying lexical cohesion, Table 6.

Table 6: Cohesion through lexical repetition in D-texts

...
[2:44] there are a lot of **leaps**
[2:45] **hops**
[2:46] **jumps**
[2:48] with legs usually extended
[2:51] there are different qualities of **aerial steps**
...
[5:11] they are now face-to-face in an **embrace**
[5:15] the swan character **clasps** the male character in an almost foetal position
[5:20] he is **clinging** his arms around the back of the swan character's neck
...

Other passages focus on a particular dancer, or group of dancers. Here cohesion is maintained by reference, Table 7.

Table 7: Cohesion through reference in D-texts

[1:00] **the** central character is kneeling
---
[1:05] there is now **another** character entered
[1:07] **who** turns, spirals and goes into an arabesque position
[1:14] _ curving
[1:15] _ twisting
[1:16] **his** arms are undulating
[1:18] sometimes **he** is stretching upwards
[1:20] sometimes **he** is curving inwards
---
[1:22] **the** central character remains kneeling, glancing upwards

In the above passage, a mention of a previously unseen dancer is cued by an indefinite article, *another* in this case. Subsequent mentions are referential - *who, his, he* - or elliptical. The return of a previous dancer is cued by the definite article *the*.

### 3.2.4 From description to interpretation

The fuzzy boundary between description and interpretation marks the passage from the literal to the metaphorical. Table 8 lists words with SL / GL > 100 in the I-texts (cf. Tables 1 & 2 for D-texts). The contrast between Tables 1 & 2 and Table 8 suggests that the experts successfully differentiated the 'Describe' and 'Interpret' tasks. However, words underlined in Table 8 also appear in Tables 1 & 2. Some of these are such general words they might be expected in any kind of dance text, e.g. *duet* and *footwork*; others are indicative of the problem in keeping movement descriptions objective, e.g. *animalistic*. What properly distinguishes the I-texts is the unusual occurrence in Table 8 of abstract nouns like *ethereality, mentalities* and *recalcitrance*.

Table 8: Words with SL / GL > 100 in the I-texts

allures, animalistic, balletic, constriction, counterbalances, duet, dynamic, ecstasy, endlessness, ethereality, figment, floorspace, flurries, footwork, hinging, homoeroticism, impacting, instigating, interlinked, layerings, magnets, manipulative, mentalities, palpability, quirkiness, recalcitrance, repelling, seducing, skyscrapers, soulful, swan, swans, thematically, togetherness, torsos, transversed, twittering, unisonal

Some interpretative statements take a metaphoric form so that a phrase with relatively objective content is linked with one that is more imaginative: in this way the signification of muscular movements is expounded. The phrases are linked by one of six phrases, Table 9.

Table 9: Linking of phrases to form interpretations in the I-texts

| Linking phrase | Freq. | Example |
|---|---|---|
| **seem*** | 32 | although her involvement in this seems perhaps more vital |
| **sense (of)** | 19 | holding the wrists, a sense of being bound |
| **suggest*** | 17 | aerial steps, which could also suggest flight |
| **as if** | 16 | it is in blue, sort of dark, as if he is dreaming |
| **like** | 16 | the stretching of the neck, like a swan |
| **appear to be** | 4 | the constriction also appears to be a support |

## 4 Processing Verbal Reports as Collateral Texts

The textual component of video has been exploited for indexing purposes by using combinations of established language technologies like speech recognition, information retrieval and information extraction, for examples see [2]. The analysis of the verbal reports suggests that they could be analysed by IR and IE techniques to index moving dance images: they are rich in specialist terminology and other keywords, linked in time to the moving image; the semi-formal characterisation of clausal information content might help to extract representations of image contents; and, the observed cohesion phenomena might support video segmentation.

As the first stage in building a knowledge-base, the elicitation of verbal reports on moving images gives the knowledge engineer a rich source of domain concepts and a set of cases. The cases can form the basis of further sessions in which the experts might elaborate the reasoning behind their descriptions and interpretations.

The collaterality between verbal reports suggests the possibility for extracting terminologies rich in lexical relations, following an approach like that used to process aligned texts in multilingual systems. Furthermore, the link to the moving image makes ostensive definitions available.

The KAB system (Knowledge-rich Annotation and Browsing) uses collateral texts to index and represent digital video. An early version facilitated the browsing of extant domain texts alongside video [18]. This was implemented in Macromedia® Inc.'s *Director* - a commercial multimedia authoring system. Following the study of verbal reports, the specification of KAB has been extended to implement a *video object* database, alongside text analysis modules. Collateral text is processed in order to semi-automate video annotation by generating video objects with reference to a knowledge-base, Fig. 2.

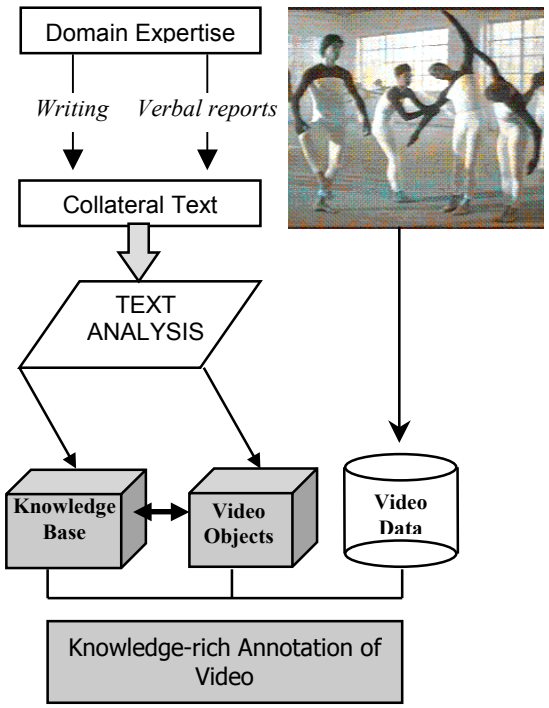



Fig. 2: The KAB Video Annotation Overview

Following Oomoto and Tanaka, a video object is a data structure which refers to a (possibly discontinuous) sequence of video frames, marked by start and end times, with an expression which represents the image contents [19]. This model is chosen for KAB because it can be easily adapted to handle different forms of representation - including: keyterms, attribute-value pairs, predicated expressions and links to other multimedia documents. With current video coding standards it is sufficient for video objects to refer only to sequences of video frames: this situation may change with the advent of object-based video coding - the nascent MPEG-4 standard[4] will give the objects portrayed in moving images their own identities.

The KAB system is being implemented in the object-oriented Java programming language, extended by the Java Media Framework[5] (JMF). This combination offers high-level operations for both text analysis and video presentation. The JMF abstracts from the physical layer of video so that programs can

[4]http://drogo.cselt.stet.it/mpeg/standards/mpeg-4/mpeg-4.htm
[5]http://www.javasoft.com/products/java-media/jmf/index.html

be written as platform and *codec* independent. The JMF API (Application Programming Interface) provides high-level commands for controlling video, e.g. 'stop', 'start' and 'go to time X'. The KAB system incorporates the text analysis software used previously in the analysis of the verbal reports.

The KAB prototype, Fig. 3, lets the user build collections of linked videos and collateral texts. Annotations can be attached to the video in the form of video objects through a series of dynamic menus which show a selection of available representations (updated through the 'Add Lexical Knowledge' option). Searching is achieved by making a selection from similar menus, which returns a set of matching video objects. Current work is implementing the 'Process Texts' function so that collateral texts are analysed to automatically suggest video objects – grounded in lexical resources and knowledge-bases. As well as being used to match queries for retrieval purposes, the expressions attached to video objects can also be used to explain the video contents to the viewer when browsing, e.g. by showing an expert's commentary on a sequence or offering a link to related media. For further information about the development of KAB see [20].

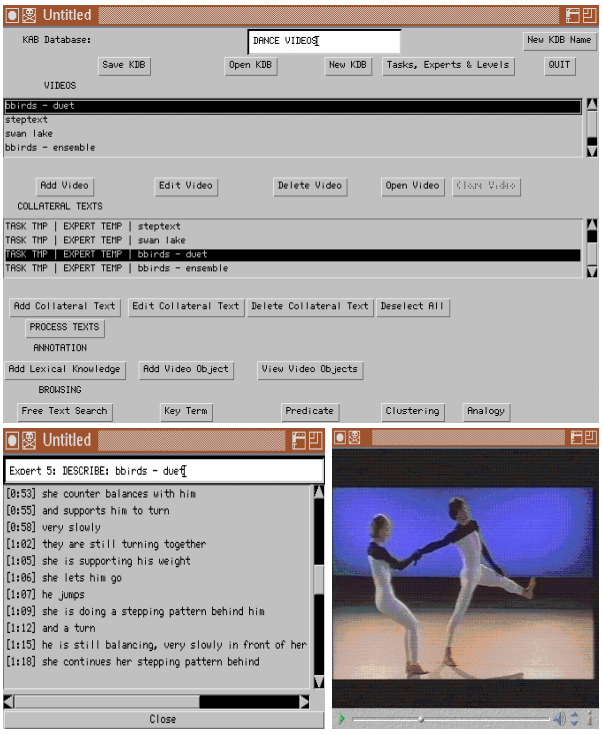



Fig. 3: The KAB Prototype main menu and example video with collateral text

# 5 CLOSING REMARKS

The use of keywords and other linguistic expressions as pointers to image contents is contentious philosophically in that it posits a primacy for language over other modes of communication. However, the words of experts elaborating upon the visual artefacts of their domain can help in understanding the artefact, especially for pedagogic purposes, and by extension for information retrieval tasks. The collaterality of texts with moving images was discussed and a suggestion made for how spoken collateral texts can be elicited from experts. It was shown how experts analyse an image sequence, both literally and metaphorically, and as it were take us beyond the image. Verbal reports generated by experts serve as content-rich running commentaries of moving images, as in the KAB system above.

The use of verbal reports for indexing images appears to be labour intensive and this fact should perhaps discount the approach for routine or large-scale image indexing tasks. Dance images, though, may be a special case in that they comprise themed sequences, realised by idiosyncratic motion patterns. Thus, particular genres, dancers, choreographers and narratives, for example, may be linked with 'signature sequences'. The verbal reports help to identify where examples of these idiosyncratic motion sequences occur. A system like KAB can then elaborate indexes and annotations of relevant image sequences for a domain.

Research in video analysis, particularly of human movement [21] [22], will perhaps lead to systems for recognising these idiosyncratic sequences in previously unseen video data. Such a system could then interact with a system like KAB to attach linguistically-based expressions to the new sequence, based on previous examples already aligned with collateral texts. This scenario would exploit automatic image analysis techniques for low-cost indexing, and would give knowledge-rich representations from the results of an initial, 'one-off', knowledge acquisition stage.

# REFERENCES

[1] R. Srihari. Use of Captions and Other Collateral Text in Understanding Photographs. In *Artificial Intelligence Review* 8 (5-6), pages 409-430, 1995.

[2] M. T. Maybury, editor. *Intelligent Multimedia Information Retrieval.* Menlo Park CA: AAAI Press / MIT Press, 1997.

[3] J. Adshead, editor. *Dance Analysis: Theory and Practice*. London: Dance Books, 1988.

[4] J. Mackrell. *Reading Dance*. London: Michael Joseph, 1997.

[5] E. Panofsky. *Meaning in the Visual Arts*. Harmondsworth: Penguin, 1970.

[6] S. Shatford. Analyzing the Subject of a Picture: a Theoretical Approach. In *Cataloging and Classification Quarterly* 6 (3), pages 39-62, 1986.

[7] C. Metz. *Film Language: a semiotics of the cinema*. New York: Oxford University Press, 1974.

[8] C. Lindley and U. Srinivasan. Query Semantics for Content-Based Retrieval of Video Data: an Empirical Investigation. To appear in *Procs. Storage and Retrieval Issues in Image and Multimedia Databases, DEXA '98*, 1998.

[9] O. Firschein and M. A. Fischler. A Study in Descriptive Representation of Pictorial Data. In *Pattern Recognition* 4, pages 361-377, 1972.

[10] P. G. B. Enser. Pictorial Information Retrieval. In *J. of Documentation* 51 (2), pages 126-170, 1995.

[11] L. H. Armitage and P. G. B. Enser. Analysis of user need in image archives. In *J. of Information Science* 23 (4), pages 287-299, 1997.

[12] W. L. Chafe, editor. *The Pear Stories: cognitive, cultural and linguistic aspects of narrative production*. Norwood NJ: Ablex Pub. Corp., 1980.

[13] J. Boose. Knowledge Acquisition. In S. C. Shapiro, editor, *The Encyclopedia of Artificial Intelligence, Vol. I*, 1992.

[14] K. A. Ericsson and H. A. Simon. *Protocol Analysis: Verbal Reports as Data*. 2nd Edition, Cambridge MA and London: The MIT Press, 1993.

[15] R. P. Weber. *Basic Content Analysis*. 2nd Edition, London: Sage Pubns., 1990.

[16] G. Kjellmer. A Mint of Phrases. In K. Aijmer and B. Altenberg, editors, *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, pages 111-127, 1991.

[17] M. A. K. Halliday. *An Introduction to Functional Grammar*. 2nd Edition, London: Edward Arnold, 1994.

[18] K. Ahmad, A. Salway and J. Adshead-Lansdale. (An)notating Dance: Multimedia Storage and Retrieval. In H. Selvaraj and B. Verma, editors, *Procs. ICCIMA '98*, pages 788-793, 1998.

[19] E. Oomoto and K. Tanaka. Video Database Systems - Recent Trends in Research and Development Activities. In W. I. Grosky, R. Jain and R. Mehrotra, *The Handbook of Multimedia Information Management*, pages 405-448, 1997.

[20] A. Salway. *Forthcoming Ph.D. dissertation*, Dept. of Computing, University of Surrey.

[21] C.-C. Lien and C.-L. Huang. Model-based articulated hand motion tracking for gesture recognition. In *Image and Vision Computing* 16, pages 121-134, 1998.

[22] T. Ahmad et al.. Tracking and recognising hand gestures, using statistical shape models. *Image and Vision Computing* 15, pages 345-352, 1997.

# POP-EYE
# Using Language Technology in Video Retrieval

Wim van Bruxvoort[1]

VDA informatiebeheersing bv
Seinstraat 32, p.o. box 2294, 1200 CG
Hilversum, The Netherlands
E-mail: wbruxvoort@vda.nl

## ABSTRACT

This document describes the Pop-Eye project.

**Keywords:** Language Technology, Multimedia Information Retrieval

## 1 INTRODUCTION

To reuse video and filmmaterial, detailed and comprehensive documentation and profiling of the archived material is very important.Advanced methods information retrieval can support processes in the domain of textual digital libraries, but for information that is contained in visual material, as in film and video, there as so far no effective methods for automatic profiling and indexing. The advances in the field of automatic image-recognition are significant, but not in that order that they can provide a sufficiently robust basis for profiling large amount of the afore mentioned,visual data in an effective way.

The European project Pop-Eye attempts to address this problem.The automatic recognition of images, and video as an extension of that, will for a long number of years, be insufficient to provide the users of this kind of data with automatic disclosure and retrieval functionality. Human language will be the only possible basis of automatic indexing and retrieval, in profiling and in searching.In the project Popeye the subtitles are the key textual material that is related to the video. Indexes are built on the subtitles with a timecode stamp on them. Searching the subtitles give direct access to the videomaterial associated to the subtitle. The user does not start at the beginning of the video, does not need to "wind" forward to the time code as if it where on a tape, but can directly watch the video.With the use of automatic translation technology the Popeye systems provides cross-lingual access possibilities to the users.

In an extension of the Popeye project, the Olive project, also the spoken word is important. In this project speech recognition will be used, as well as linking these texts (subtitles and speech recognized text) to other text documents related to the video such as storyboard, scripts and other.

In the rest of this document we present the Popeye system in the situation at the user sites, some functionality and a brief description of the technical design and implementation of the system.

## 2 ARCHIVING AND REUSE OF VIDEO PRODUCTIONS

The primary users of the Pop-Eye projects are major European Television Stations, comprising VRT (formerly BRTN, Brussels, Belgium), SWR

[1] This article is based on a early article on Popeye by Klaus Netter (DFKI) who was the first to promote and present the Popeye project, and furthermore I like to thank Joop van Gent and Wessel Kraaij (TNO-TPD), Franciska de Jong (TNO/University of Twente) and Godfrey Smart and Erik Aarts (VDA), their committment to the development of the system is enduring.

(formerly SWF, Baden-Baden, Germany), and TROS (Hilversum, Netherlands. For all of these broadcasters archiving of video productions plays an important role. There are several purposes to reuse video material. The first, and maybe most important for reducing costs, is the reuse of material in new productions. The second purpose is for general research goals by programm producers. Other uses are reselling or re-broadcasting of existing video material.Reuse and research purposes make that the users have the opportunity to access detailed information about the content of the video material, without the need of viewing the material first.

Describing the content of video material needs to be done by experts in the field of broadcasting and archiving. Besides that, it is also very time consuming. To describe one hour of video material takes over ten hours of time of a trained documentalist. For large broadcasters, with a large, daily growing, video archive, it is therefore very expensive. The archive of video consists up to this about some keywords per video, to together with a short summary of the content. Notable exceptions are among others the extensive documentation provided through the FESAD (IBM) database in the German ARD federation, in use at SWR, or the content disclosure of the BRTN video archive in their Basys-system. In the Fesad system even the raw data, the unused material, is documented. One can imagine that all together this data is of great importance and value to the user organization.

In an ideal situation the user at a broadcast organization, most likely a producer, is able to access the video archive through an internet or intranet. As a result of his query to the database he can decide to download or directly (on-line) view the video material in a format and quality that he prefers. This format can be only an image (per subtitle), a low quality (one to sixteen frames per second), or a format of near-broadcast quality (MPEG2).In case the user decides to use the material he knows the physical location of the original analog tape or can use the digital material directly.

## 3 AUTOMATIC INDEXING AND RETRIEVAL

To answer such problems and demands as just described, Pop-Eye attempts to provide online access to video material on the basis of linguistic material associated with the content of the material. The main tasks performed by the system are capturing, indexing and retrieval.

Capture. Video material is digitized and aligned with the subtitle text files by inserting the time code of the video into the subtitle textfile.

Indexing. The texts are processed on the basis of state-of the art language technology and different indexes are constructed from the text. Where possible the texts or the indices are translated.

Retrieval. In response to a search query by a user, the system retrieves textstrings that match his query, and allows for downloading and viewing the corresponding video sequence using the time code.

At all user sites the actual broadcast quality video material is still broadcasted in analog format, and archived in analog format. This makes it necessary to integrate the digitizing part of the work into the capture software. This can be done using commercially available software, like Adobe Premiere.

In digitizing the videomaterial choices have to be made regarding the requested quality. Once a video is digitized in a low quality, for instance at a rate of only 8 frames per second, it can not be upgraded to a higher quality withour redoing the digitizing process. Currently, the users digitize video in a MPEG1-format and the digital file is stored on a CD-Rom. Then it is transcribed from MPEG1-format to the RealVideo format. The latter allows viewing the file over an intranet with acceptable performance. RealVideo also allows to start a videofile at a certain timecode, which is not possible with the MPEG1-format while using a standard available viewer. Still the digitizing is not directly done in the RealVideo format because in the near future a new standard may be used, even netwerk performance may increase so that viewing the MPEG-file directly is possible.

As an alternative to view the whole video, a storyboard can provide the user with sufficient information about the content of a video. A

storyboard contains captures images ("shots") of the video, mostly on scene-changes. A sequence of images is then shown on the screen. Applications to automaticcally generate this kind of storyboard are available. An exaple of this is the storyboard developed in the Euromedia project (http://www.foyer.de/euromedia/). Besides the feact that a good storyboard gives an overview of a video on one screen in one time, downloading a storyboard takes less time then downloading an video, since it is much smaller in file-size.

The text associated with a video is in Popeye the subtitle. This can be a translated text for non-native speaking viewers, or an abbreviated text representation of what is spoken as a service to the bad of hearing. Of course the spoken text itself can also be a source for data to associate with the video, this requires automatic speech recognition technology. The latter is not part of the Popeye project, but is one of the goals of its' succesor, the Olive project.

After capturing the video and the textual data and mapping the two together, indexes need to build from the textual data. The linguistic approaches used to analyse the text and build the indexes in Popeye, will not be presented in detail in this paper. [more can be added: Among the approaches are fuzzy matching, phrase extraction and proper name identification. To have multi-lingual retrieval functionality off-line translation is applied. The full text can be translated (using the Logos-system) or the index terms are translated thus supplies a partial translation. Both allow the user to enter a query in his or her own language and retrieve documents in another language, but view these (partially) translated. The partial translated text is imperfect, but sufficient to enable the user to judge the usability of the retrieved documents.

One special aspect in retrieving video through subtitle texts is explained below, to give an impression of the special character of retrieving video through text. It is possible that a long sentence is divided amongst two or more subtitle lines, each of them with their own timecode. While there is one scene, possibly without any scene changes, the subtitle text can contain phrases that are relevant to the scene, but are divided into several lines of text. Retrieval of text from idexes build on the seperate subtitle lines is less precise than it would be when the subtitle lines related to the (one) scene are viewed as one sentence. (Counterwise one subtitle with one timecode can be displayed during several scenes. For instance a subtitled videoclip of a pop artist or group on MTV.) toThese may seem rather trivial aspects, but they are not from a users' perspective.

The subtitle apparrently does not describe the image or sequence of images that is on the associated video. Non-automatic profiling of video material by expert documentalists is also imperfect, but visual aspects, such as the launching of a rocket or the raising of a flag, can be described by an expert, but are not contained in the subtitles. So the association of the subtitle with the image is a major area of investigation during the user evaluation of the Popeye project. So far the first preliminary results have not been disappointing. As stated in the introduction of this article, it is very important to keep in mind that using the text is the only currently available technology to automatically disclose broadcasted video material on a large scale.

One of the extensions of the Popeye system at one of the user sites that is currently investigated is a connection of the system directly to the broadcasted TV-signal, including the broadcasted subtitle text. With this application a broadcasting company can completely automatic extend their content management system with a retrieval function to their own broadcasted material. It is a clear that here the afore mentioned imperfectness of the functionality, subtitles not fully describing the visual content, is less relevant for the user organisation, since it takes no effort (of human experts) to build the archive. The advantage of automating the disclosure is clear.

## 4 TECHNICAL IMPLEMENTATION

The system is implemented through the co-operation of several technology providers, research institutions and universities. These include TNO-TPD Delft, which built the core indexing and retrieval functionality, VDA informatiebeheersing bv Hilversum, which is developing commercial software for the Broadcasting and Publishing industry and which built the video capturing software and is responsible for system integration, the University of Twente and the LT Lab of DFKI GmbH

Saarbrücken, which are responsible among others for the language technology, the University of Tübingen, carrying out the evaluation in Pop-Eye.

The technical design and implementation of the Popeye systems allows the user to access the system through an Internet or Intranet with the use of a standard browser. The user retrieval interface is completely build in Java. The data is stored in an Oracle database, but other databases can be used as well, using JDBC (Java DataBase Connection), the digitized video is stored on a separate database-server. This allows every user organization to connect to a video storage system that is preferred.

The capture software is build in Visual Basic and allows the user to import (subtitle) text files in their own format or in a standardized EBU (European Broadcasting Union) format.The user management software is again build in Java and accessible through common internet technology.All indexing software is available through API's on the server, with the Microsoft NT operating system.

The future of the Popeye system is that it will be integrated in the content management system DAS, a VDA product. Furthermore the system will be linked to the existing archiving systems at the user sites: the afore mentioned FESAD system at SWR, the Basys system at VRT and the POP-system at TROS. Popeye will be reused and extended in the Olive project, and VDA intends to make it available as an API-like functionality for other content or asset management systems


**Figure 1 Part of the search screen ('milieu'='environment')**

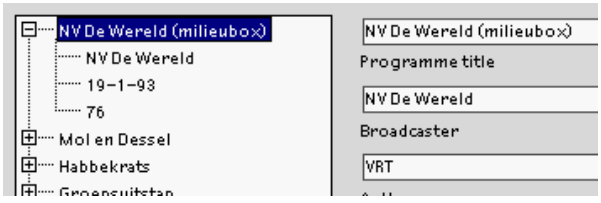



**Figure 2 Part of the results**

| score | timecode | subtitle |
|---|---|---|
| 76 | 00:05:20.24 | De milieu-box wordt gemaakt bij |
| 76 | 00:10:16.16 | Zouden ze hier ook een milieu-box |
| 76 | 00:02:02.19 | Hebt u al een milieu-box? -Wat? |
| 76 | 00:00:00.05 | Mexicaanse illegalen Dag mevrouw. |
| 76 | 00:02:07.20 | Uw milieu-box. -Ik denk het wel. |

**Figure 3 Text hits in the subtitles**


**Figure 4 Part of the story board (images)**


**Figure 5 The video**

# Going digital at SWR TV-archives

# New dimensions of information management for professional and public demands

Istar Buscher
Suedwestrundfunk (SWR)
Documentation and Archives, Research and Develeopment section
Hans-Bredow-Strasse, D-76530 Baden-Baden
istar.buscher@swr-online.de

## ABSTRACT

This document describes SWR's (German public broadcaster, member of ARD-network) involvement in European multimedia projects for digital video archive solutions. Main focus is on the results of the Esprit project Euromedia and it's first implementation for productive usage at one production department. Also the paper describes the technical and administrative difficulties for TV-archives while shifting to digital archiving.

**Keywords:** Multimedia Information Retrieval, Pattern Recognition, TV archive, Media Asset Management, Euromedia, Digital Video Archive

## 1 INTRODUCTION

To a large extent still unnoticed by the TV consumer, at present a technical change is taking place in the entire spectrum of television production. From the studio technology to film editing and video archiving the digital revolution enters the TV companies.

This digital revolution in Germany is correlating with the merger of two of the eleven members of the public ARD-network. Since September, former Suedwestfunk (SWF) and Sueddeutscher Rundfunk (SDR) form together the new company Suedwestrundfunk (SWR). Both partners within SWR have already a large common experience in handling the new techniques. Since five years now the TV archives of SWF and SDR are together involved in different national and European research projects. The most successful project so far, the Esprit-project Euromedia will be concluded still in this year.

Reason enough for SWR to try to balance the rapid development on the sector of digital archive solutions and to resume the problems that one is confronted with while organizing the digitisation of TV-archive material.

Before I present the results of Euromedia in detail, I think it is necessary to describe the specific workflows of TV archiving for a better understanding of the background. I understand my paper as a view from the practise on the developments offered by the IT market.

## 2 A TV-ARCHIVE

Generally speaking, a TV-archive keeps TV-programmes for the purpose of possible re-use of stock footage in new productions, as well as for re-broadcasting and sale. Compared to other archive sections (like a press archive or library), the main task of a TV-archive is to find again picture content to illustrate new programmes, not the background information about something. It's strategic meaning grows with the expansion of broadcasting and cross-media production within the digital age.

Since the beginning of production at SWR's predecessors SWF and SDR in the early 1950s, there is at least more than 250.000 items of stock footage, mostly broadcast material, stored at our TV Archives. Because of the historically grown structure of SWR, it is distributed over 3 main sites in 3 towns. According to the technical development, there's a large variety of cassette formats stored: 35 and 16 mm film, 1" VTR (earlier stored 2" VTR were completely copied to 1" VTR to save storage capacity), D5, Betacam SP, Digital Beta, VHS (preview copies) etc.

To facilitate the retrieval process, additional text-based information (i.e. content analysis and meta-data) is recorded after the programme is finished and/or has

been broadcast. At SWR, there is a STAIRS-based database called FESAD established since 1985. Content analysis (sometimes called disclosure) means first the description of the picture content (as detailed as possible), second the description of the subject. Meta-data are facts about the staff involved in the production, the broadcast date and time and the condition of the archive material. It also includes IPR-rights information.

Journalists are used to choose between online access to the TV-archive database from their own desktop or to ask one of the documentalists[1] to do a research. However, the retrieval may still end up becoming a time-consuming and tedious task, since the video material is stored separately on video cassette tapes. Also, the relationship between a programme and its content disclosure, is still limited, e.g. because of information lacks.

The number of lend tapes, cassettes or film reels increases. The growth rate of external loans increased from 1995 to 1996 13,9 %, the in house loans risen 28 %, as a result of the grown number of channels, that have to be delivered with content (already broadcast productions) or footage.

To find solutions for a more convincing kind of storage by using the chances of digitisation, it is highly relevant to examine the demands being placed on the TV-archive - first by it's own management, second by it's environment (journalist staff and broadcast management). In this field, we have to distinguish between traditional and new needs for the production workflow.

## 2.1 DEMANDS BY THE ARCHIVE MANAGEMENT

The increasing number of lend stock footage maybe a positive hint for the benefit and functionality of a TV-archive, but brings also a lot of problems. Providing the archive materials, i.e. making them available and accessible in time, is getting harder and harder. Among those problems arising with an expanding footage expedition are the following main issues, that were originally formulated by my colleague Albrecht Haefner for sound archives (but are also valid for any kind of media archive)[2]:

"An extremely delicate area is the preservation of the stocks, especially the quality control and the safeguarding. Due to the fact that the public broadcasting companies could afford the best material and equipment as well as optimal storage conditions in the past, regular quality control and safeguarding hardly existed".

Archivists who have not performed quality control measurements regularly in the past are not able to give any evidence about their collection's condition for now.

As Albrecht Haefner further says, it is the video tape itself as a physical subject, that causes troubles in many ways:

- tapes required for transmission, issued on loan or simply lost, are not available;
- tapes are worn out, sometimes even damaged;
- tapes need proper housings, appropriate storage conditions and storage space;
- supplying tapes requires costly staff;
- copying is possible only in real time (and sometimes even not, if recorders and viewing places are occupied by colleagues);
- parallel or simultaneous use of tapes is not possible.

## 2.2 JOURNALISTIC DEMANDS

The primary service offered by a TV archive is assisting journalists (i.e. reporters and directors) in finding pieces of video from the archive[3]. In general, there are two different types of video information present in a TV archive: the final version of a television production as it is broadcast, and the un-edited rushes or raw footage that is the outcome of video recording. At SWR we used to store only the broadcast version of a production because of saving storage capacity.

In most cases, archive users are looking for pieces of video that can be used to visualise a specific (mostly abstract) problem or issue; they are also searching for video with a certain image or audio content. In other cases, reporters are doing more thorough programme research where the target may be old documentaries or news related to a specific topic. Thus, video archive search tools should be capable of performing queries on video document topics as well as image/audio contents and should allow the user to determine the type of video to retrieve.

The context into which a piece of video belongs, has a strong influence on how this piece of video is interpreted. Thus, when a piece of video is retrieved from the archive, it is often necessary to view parts of it's context to interpret it.

### 2.2.1 Criteria for re-use of footage

Analysing user questions to documentalists in the past, we can distinguish the need for six different categories of description for stock video footage[3], of which the content of the shot is only one. Each of these ways of describing a video sequence might be germane to the

question of whether a video can be used for a certain purpose in a new production.

a.) The content of the scene (information about the shot location, the activities that are occurring in the video, the types of people and the roles they are playing in the those activities, and the salient visible objects).

b.) The points illustrated by the video (video often needs to communicate abstract ideas and relationships, and stock videos are often sought to carry that communicative burden).

c.) The composition and camerawork of the video (e.g. camera angle, motion - like zoom or pan - and focus).

d.) The likely functions of the video in a larger narrative (transitions, interludes, prologues, background or hooks).

e.) Information about the source of the video (meta-data: the physical location of the original video cassette, the clip's start and stop time-codes, and licensing or copyright restrictions that may apply).

f.) The relationships to other videos in the TV-archive (some videos will depict scenes of the same place on the same day, or even at the same time. Some will include the same actors or the same objects from different vantage points. Others may have been recorded with the intention of continuity, that is, one shot may be a reaction shot to something in another video. In addition, both typicality and rareness in the way of presenting a content can make materials valuable for reuse).

### 2.2.2 Problems of current disclosure work

The documentalists classify, describe, and annotate the contents of video documents and recordings to make searching and browsing possible in an efficient manner. Often, information related to the video material which is known during the production phase, is not very well taken care of and may never find its way into the database. In many cases, the documentalists have to do research for this information from the production phase and register it in the database. Therefore, cataloguing is also time-consuming.

Because of the lack of detailed meta-data information about the content and the requirements to work efficiently, these descriptions are usually coarse grained and describe long pieces of video - e.g., complete news items. As a result, the user can only retrieve a general description about these long pieces of video even when one is looking for smaller pieces - and has to review the tape.

### 2.2.3 Problems of the current retrieval possibilities

There are a lot of more problems related to today's way of archiving and documentation. The most obvious ones are:

*Availability*. Physical distance determines how easy access a user can have to the archive. Users working nearby the archives location have easy access while users working in offices situated apart from it will have a poorer access to the archive. Since the merger of SWF and SDR, the need to more comfortable remote access systems arised, because of the several sites, spread over Baden-Württemberg and Rheinland-Pfalz.

*Searching is time-consuming*. Users find it rather time consuming to retrieve what they are searching for. The reason is partly that not all aspects of the video information is described in the search-able text database, and partly that the user have to alternate between a terminal for database searching and the archive and its viewing station for playing back the video itself.

Almost this leads to the fact, that there is still a *limited access* to information, that is not disclosed within computerised catalogues. At SWR e.g., the FESAD database was introduced in 1985. Much older material is only accessible via a mixture of manual card index, manual notes and, in some instances, personal memory. Different systems are frequently utilised for different types of media, also at SWR. For example, archived films and video, stills, audio and text material have individual and in some parts incompatible information systems.

### 2.2.4 New demands

For a few years now, the footage, which is not needed for a production, is stored in addition to the broadcast versions. In earlier times, not needed footage was thrown away because of the costs of storage and the shortage of storage capacity.

Today it is bitter recognised that the costs of storage capacity may be much less than producing new footage at original locations, e.g. special natural settings; also in a growing number of cases, the state of environmental pollution and destruction determines, that more and more nature documentary films can’t be reproduced as often and when ever one needs it.
Other reasons for this change in storage policy from the journalist’s perspective are

- also financial reasons (urgently needed sujets are to far away, budget is too small, production must go fast);

- and technical reasons (e.g. need for black and white-film of an old gas station, old telephone switching head office or old advertising films).

From the journalists point of view there is also another reason for storing un-used footage. Usually, only a short sequence (about seconds) is used in a programme, where the whole programme is often not longer than one and a half minute (news, magazine item). Whereas the identical shot in the un-edited footage may be (and usually is) very much longer. Or one has more than one time filmed the same picture content in fear of one of the shots may be blured or unusable because of any other reason. If this case does not occur - one has a lot of relative long-running stock footage in high quality. These facts led to the decision at the archives, to store in a minimum number of production cases all un-used or un-edited footage content for later re-use purposes. Of course this leads - in addition to the growing number of programmes - to an explosive growth of archive material.

Taking into consideration the above mentioned developments, both technical and economical changes, one could imagine, how difficult it becomes for traditional TV-archives to satisfy and - furthermore - to guarantee the full-filling of the demands by journalists and programme planners.

Almost from one day to another, the archive sections of public broadcasters became the "cost centres" of the company - with the condition (more or less quoted and expected by the management) to change themselves - whereever possible - into profit centres.

But at the same time, more and technological leading-edge services are expected by our new old "customers", who came in some cases directly to our colleagues to articulate their new service wishes.

Dissatisfying and unreliable sides of text retrieval databases and the constraints of textual content analysis should be replaced by direct on-line access to archived material in preview quality. Journalists expect a higher and more precise number of matches from viewing pictures directly, instead of going the roundabout way via textual description of the footage, which is always a subjective interpreting. In fact, technical language- and terminology-related problems could be solved. Viewing chosen footage content with the possibility of direct comparing and intellectual combining of pictures leads to a more creative way of producing new features.

The next step of the changing demands was the journalists' wish to commercialise their products through the advantages given by modern cross-media- and electronic publishing. Staff of our educational, cultural and scientific desk discovered the practicable way to resell content at the same time as booklet, CD-ROM, video or Internet service. The need for these kind of publishing to the TV-archive is clear: availability of footage in electronic file formats.

With the social and economic thrust to the public broadcasters to produce as efficient as possible, projects were initiated to test the market capacity for offering not yet edited footage instead of selling transmission rights. The seasonable way for distribution in this case is also the Internet. Relating to the last point the journalists think about the world-wide commercialisation of a production before the shooting begins - titles, plot and style of a feature have to be accepted all over the world.

The growing number of channels and distribution techniques creates opportunities for a more creative re-use of footage within new and unusual contexts, and this provides additional outlets for previously under-exploited material. For example, archive news material can be re-used in a variety of ways, ranging from illustration of specific historic events and as stock-shot background footage, to content for quiz shows and humorous programmes.

This implies a very large amount of problems - the IPR-questions increase with the growing multimedia market. Copyright questions of the sometimes years-old footage have to be answered, which makes the search for - in this way - "neutral" pictures more difficult. In addition, creative and innovative use of footage, such as virtual created rooms or mixing of detailed cuttings within trailers makes it more and more impossible to identify and catalogue rights owners.

At least, for some broadcasters it is virtually impossible to generate administration reports on items such as lending, usage and royalty payments for stock footage.

But as sophisticated the new demands were formulated by journalists, as clear was the message: „whatever you will develop - make it as simple as possible; make it like this: typing in a keyword and click ‚Return'." The overall goal is to assist the journalists to find materials of interest with ease and as comfortable as possible, and to provide the users with tools for a more successful search in large video stocks.

## 3 SWR'S RESEARCH PROJECTS

At SWF's TV-archive, the challenge of the digital revolution was faced very early. Since 1992, we are trying to find ways of providing and installing state-of-the-art archive solutions.

Our documentalists have been working this year on four research projects with different points of emphasis together with international broadcasting partners and the software-industry. Three projects are driven under

the funding umbrella of the European Union: Two (Euromedia and VICAR) are ESPRIT-projects, one (PopEye) is a project within the Telematics Application Programme (TAP).

The main focus in our research is developing on-line access to our video archives. We have been extending this technology in three different projects, which differ mainly in their geographic scope. The first ever project, DIVA, was the direct archive access project and ARCHIMEDIA extended its scope by providing access to it via a WAN connection. Euromedia, which has been concluded at the end of November, extended it further to an European level.

The second focus of our work is developing advanced retrieval functionalities for digital video archives. Both PopEye and VICAR concentrate on that issue. While the PopEye-consortium tries to built up state-of-the-art multilingual text retrieval, VICAR (based on pattern recognition) concentrates on retrieval with „query-by-image“ and automation of content analysis (For further information about PopEye and VICAR please read the papers of Wim van Bruxvoort and Ed Tan).

## 3.1 Automatic video analysis

To enable retrieval systems to handle sequential video data streams, they have to be processed. Generally, automatic video analysis includes three steps, i.e. segmentation, content extraction and indexing[4]. The first step in video analysis is usually video segmentation, i.e. partitioning video data into manageable units for easy storage and browsing. For that, it is wise to use shots and sequences as the natural units of a film. There are three ways of selecting key frames already on the market:

1. Showing a pre-definded number of representative key frames out of a shot;
2. Selecting key frames in a pre-defined time intervall, e.g. every 10 seconds. The documentalist can choose the number of key frames presented out of a specific time intervall;
3. The third method is to select the first key frame after the beginning of a shot.

Once a video is partitioned, each unit can be analysed to extract its content. Indexing in this correlation means the construction of data structures that refer to specific contents. Video indexing involves the mapping of video streams to index structures that support searching and grouping of video materials. Automatic indexing should also aim to be semantic, to allow users to search and group videos according to their specific requirements. From the documentalists point of view, video indexing should have two aspects: after extracting useful information from video sequences, classify material extracted from the video (or supplied by the video author, journalist, director) and represent essential information concisely and precisely to the users, or for further processing purposes.

# 4 EUROMEDIA

Euromedia is the third project in the area of digital video applications, in which SWF participated. It started in January of 1996 and was finished in November 1998. It is not the largest, but still the most successfull of our projects[5].

Partners in Euromedia were the broadcast companies BBC (Great Britain), ORF (Austria), SVT (Sweden) and SWR's predecessors SDR and SWF (Germany). The technical group consisted of the software-house TECMATH GmbH, Kaiserslautern (Germany), the European Research Center (CEC) of Compaq (former Digital Equipment GmbH), Karlsruhe (Germany), Helix 5 (the Netherlands) and ITC-IRST (Italy).

As described earlier, in all large archives (and also at SWR) today still a textual description of contents of audio-visual materials is created for documentation and research purposes. These descriptions of contents are retrievable over mainframe (host) system networks in textual form. The moving pictures themselves however remain stored in present practice on conventional carriers such as film, tapes and magnetic tapes. In order to be able to really view the sequences described in the text database, these substrates must be taken physically from archive magazines and displayed on video playing stations. Only in this work procedure the concrete selection of the suitable video sequences takes place.

Euromedia strategies and measures have been developed, in order to overcome these temporal, material and geographical limitations of the current work routine. Central item thereby is a large-scale digital multimedia archive, which is attainable via local and spacious networks. The Euromedia consortium has developed modern tools for automatic and semiautomatic video description, video indexing and automatic analysing of longer video sequences or transmissions into its individual adjustments.

In order to overcome the capacity boundaries, which are still given within digital mass storage in today's technology, applied suitable data compression methods have been considered. The archived video sequences are not stored in transmission quality (broadcast quality), but in capacity and cost-saving opinion quality (preview quality, MPEG I). This ensures rapid option in the available video stocks and

thus the direct and time-saving access to the accurately selected sequences for the production phase.

Euromedia can be seen as the development of a basic digital video archive platform. It is not a finished, individual, final system, but it is like a box of bricks, a record of items, which are assembled. The EUROMEDIA system consists of five software components:

- An object-relational database management system for storage and retrieval of media objects and descriptive data;
- An acquisition client for digitisation of video material;
- An indexing library for automated video analysis of the digitised video material.
- A documentation client for editing the indexing results and entering descriptive data.
- A retrieval client for full-text search through the descriptive data, for viewing the results.

Core of the system is the video analysis, with which a digitised MPEG I-film is divided automatically into its sequences. Main elements of the video-analysis are shot detection and keyframe extraction. Like that it is possible for the first time, to show the entire picture content of a film by *representative* key frames (still images) on only one display screen page, the so-called storyboard. Technically, the storyboard is a list of keyframes linked to transcripts or factual content descriptions via time-code. The storyboard itself is stored as one MJEG-file. The user interfaces of Euromedia are based on a usual webbrowser (Netscape). This way, the journalist has the possibility of investigating picture contents from archive copies on-line on the own workstation (or even at home - in the frequent case he or she is a freelancer), independent from office or archive location.

The result of the video-analysis is not only the storyboard itself; the storyboard is nothing else than the visible part of a list of time-codes. It lists all detected shots with „time-code in“ and „time-code out“. This table can be saved on a floppy disk and exported to non-linear-editing systems, such as AVID Media Composer or Media100. The structure of Euromedia's time-code list is the same as AVID's edit decision list (EDL), the result of the logging process on a Media Composer.

Further more, Euromedia offers a possibility to produce a raw cut versions out of the stored video material. In a special feature, the cutter (or even the journalist) can assemble a first raw cut version out of the offered storyboards - just by a using a „cut-and paste“-functionality. The result is again an EDL-like time-code list.

Thereby a cutter (of the same company or of a private production facility) - independent from the journalist or author of a documentary - can also produce a raw cut version of the new product - in this case, two creative processes will take place at the same time - independent of each other.

In the last years, it was usual that both - the journalist and the cutter - together previewed all the material by digitising it for the AVID and then - only together - produced a raw cut version. Now time and place do not count anymore. Taking this example only, the implementation and use of Euromedia will save two from five days of (at the moment mostly) outsourced editing. Further more, it will save travelling costs and expenses for freelance authors.

For documentation purposes, Euromedia offers two user interfaces for manual textual annotation. When a storyboard is newly created, one has the possibility to shift to a writing mode and click on each of the key frames. Behind each key frame there is a text field, which allows a textual annotation on shot-level (and implicit for later developments on frame level). To ease the change between the textual database and Euromedia, there is also the possibility to show all textual annotations in one global text field (with the impression to be one genuine text).

## 4.2 First applications

A system like Euromedia can be used in different workflow environments.

SWR will run two long-term test implementations. The first implementation will be a part of a programme planning system in combination with a text retrieval database on the basis of Lotus Notes. Within the production department of „ARD Buffet“ (daily live magazine with three different issues) there exists a set of 500 pre-produced short inject films (contributions). One time a week the chief of staff has to decide, which inject films have to be broadcasted in the next seven days. Today, he is doing that by previewing VHS-copies. With Euromedia, the chief of staff gets the comfortable possibility to preview programmes for possible broadcasting directly on his desktop computer, without handling any tape.

The second implementation will be a public website, which can be accessed via http://www.swr-online.de from the beginning of next year. With the help of Euromedia and an AltaVista search engine, a selection of very valuable and visual impressive video material like the broadcast versions of the UNESCO-series Treasures of the World - Heritage of Mankind can be viewed by everybody. A combination with a programme schedule is imaginable.

The results of Euromedia will be marketed by the company Tecmath, Kaiserslautern[6], under the name Media-Archive.

## 5 First experiences with Digital archiving

At SWR we are just in the beginning of digital archiving - but on the international level, there is a broad discussion about the right way and the appropriate technologies to go digital at TV-archives.

On a symposium held at SWR in October 1998, Peter Dusek, the new selected President of the International Federation of Television Archives (FIAT/IFTA), summarised the enormous expectation attitude, with which the responsibles currently are confronted many times even within the TV archive area: "Everybody is talking about digitisation, that is at the moment our actual problem. Many believe that it is already possible today to do research via Internet in all TV archives of the world at the same time. But we are evenly not so far yet."[7]

Investigating the integration of television and computer technology, the very different planning periods of both branches of industry play a substantial role. While the IT industry is planning in really short and medium-term time-periods, the broadcasters have to face the expectation of long-lasting solutions by the their customers, the viewers.

The second main problem, TV-archives are facing today while thinking about digitising their stock footage, is still the financial expenditure. Nobody can really say today, when the complete digitisation of archive stocks will be payable, although the costs of storage capacity are fallen in the last 27 years around on the average 30 per cent per year. But even if it would be payable today, it would take years and years to digitise every item of our archives. And who will guarantee, that there will still be formats like MPEG I and MJPEG on the market then ? „We are not only responsible for certain technology thrusts for today and tomorrow, we are responsible for the materials as such. Not in their physical form, how they are present as tapes, but for contents, which are recorded on these tapes; - for the actual treasure, which is present in the archive, and it doesn't matter in which physical form. And therefore what matters most is to be particularly careful, in the selection of the strategy and the technology, with which one wants to secure this treasure.“[8]

TV-archives really have to make a decision, with which materials they want to start the digitisation. It may be the number of celluloid films that currently disintegrate, but it may also be a number of most needed and most successfull programmes.

Responsibles of the IT industry feel the reserve, that most documentalists have against blind trust in new technologies. Rainer Kellerhals, the Director of the Digital Media section of the Euromedia-partner Tecmath said: "My experience in the field of private TV is: even if one has the financial means and even then, if one really builds up completely new facilities, one falls back within most areas to relatively conventional solutions, because one must have easily the warranty that it works. And one experiments only in quite small areas." TV-broadcasters have to deliver their programmes daily, without IT-caused time-lacks, and that's their most important task.

One big problem, that is not solved by digital television equipment products as they are marketed at the moment is the growing number of cassette formats and standards. The TV-archives do constantly copy to rotational newly released formats, to be still on the edge of technological developments. Most archives have secured their stocks from 2-inch to 1-inch at very high expenditure. Then U-matic and the Betacam SP came on the market. And what today is the Digital Beta, would be tomorrow Betacam SX and DVCpro. Also the Celluloid film plays still a quite substantial role: "Also the film becomes naturally long-term secured. Only we ask ourselves naturally - for how long can we afford at all still this number of copying processes, copying the available stocks according to always new formats and always new requests. I think, we reached a point, where we must consider exactly, what we actually can still carry out.", explained Wolfgang Dehn, head of the Documentation and Archives section Baden-Baden of SWR, the situation. "If I hear, which new formats are introduced to the market, then I do really feel sick", he added, "also the exchange of programmes between ARD and ZDF and also the requests for a world-wide procurement or delivery of programme material brings us therefore increasing difficulties."

What we as documentalists and archivists do really need at the moment is an internationally accepted and standardised format, where a bypass time is ensured for the next years up to the digital, tapeless storage, and please - not still 5 or 6 additional formats.

## 6 Conclusion

The fact that different strategies can also be quite fruitful in the co-operation between computer companies and TV archives, appeared also with the research projects at SWR. "Our technique partners were at the beginning not quite conscious of the complexity of the operational workflows in a

broadcasting corporation. But after initial problems one must say that it became a very profitable co-operation", described Ulla Kreuder (at SWR responsible for Euromedia). "One can learn very much from each other, the technique partners are also ready to deliver much of their know-how. It is important for them, that the user partners understand the technique, which they apply. A very profitable know-how transfer takes place." "My most important experience: software and hardware development are always very strongly shaped of the divergence between desire and reality at the beginning of a project“, said Joachim Haitz, but "the feedback, the mutual fertilisation of practical work and research is the crucial point."

In this sense, Euromedia is the successfull example of the co-operation between public broadcasters and IT companies. It’s open system, the possibilities for mutual applications and the fact, that it can easily be integrated in existing television workflows make it a valuable brick for public broadcasters worldwide.

## REFERENCES


[1] The term 'documentation' has a different meaning in English and German language. The expression 'documentation' at SWR is historically grown and it is also used as description for the whole department. Also ‘documentalist’ as a profession is sometimes unusual,- one can also find ‘librarian’; but the old-fashioned ‘archivist’ at SWR again is only used for the colleagues working in the TV-archive’s storage magazine.

[2] Albrecht Haefner (Suedwestrundfunk): The Role of Archives in the Multimedia Market - the Archive View. An archival look into the next century. Colloque Monte Verità - May 19-21 1997. http://www.srg-sr.ch/memoriav/fra/ahaefner.htm.

[3] Istar Buscher: Multimedia in SWF TV-Archives. Developing state-of-the-art services for journalistic demand. Presented at the International Academy of Broadcasting, Montreux; Seminar: ”Television Archives - Preservation and Creative Use” April 16-17, 1998

[4] Eric A. Domeshek and Andrew S. Gordon: Structuring Indexes for Video Clip. Submitted to IMMI-1, First International Workshop on Intelligence and Multimodality in Multimedia Interfaces: Research and Applications. http://www.ils.nwu.edu/~gordon/papers/immi95.html.

[5] For Euromedia information on the Internet, please have a closer look at: http://www.foyer.de/euromedia/; http://www.cordis.lu/esprit/src/ep20636.htm; http://www.foyer.de/euromedia/presentation/application.html; http://hera.itc.it:3003/~brunelli/Euromedia.html.

[6] TECMATH GmbH & Co. KG, Sauerwiesen 2, 67661 Kaiserslautern, Germany, Digital Media section: phone: +49 / 63 01 / 606-200; fax: +49 / 63 01 / 606-209.

[7] All quotes from original interviews of the author during the symposium „Going digital - what else“ at SWR on October 13-14, 1998 in Baden-Baden.

[8] Herbert Hayduck, multimedia project manager in the TV-archive section of Austria’s ORF.


For further information on SWR’s multimedia activities please have a look at http://www.swr-online.de/.

# Computer vision and image search engines

Arnold W.M. Smeulders, Theo Gevers, Martin L. Kersten
University of Amsterdam
the Netherlands
{smeulders, gevers, mk}@wins.uva.nl

## ABSTRACT

Image search engines call upon the combined effort of computing vision and database technology to advance beyond exemplary systems. In this paper we charter several areas for research and provide an architectural design to accommodate image search engines.

## 1 INTRODUCTION

Image search engines has moved to the center of active research. This new era stems from the fact that practically all imaging devices now deliver their data digitally, including home video. Also, the storage capacity of standard computers has easily surpassed a thousand images. And, finally, Internet has enhanced a visual component in the domain of the computer. No wonder that visual search, query by visual example and visual browsing generate prominent scientific questions.

One means to search for an image is to use a catalogue, as is done in (http://ipix.yahoo.com), or to use textual information. Such caption- and text-based search for pictorial information should always be explored where available, as captions may be very informative about the picture content. But, captions can not capture anything of the image content outside the intention of the writer. Therefore, we make one step beyond text-based search for picture information. We search for pictures by pictorial content.

Current prototype systems include Photobook [16], VisualSeek [22], Virage [15] and QBIC [19]. They provide access to image archives based on contextual information and-or a measure of similarity based on color, shape, and texture features. The features provide the basis for the description of the image content. For a given image feature values are then compared with the feature values in the database to find an image similar to the one shown as an example. The comparison is done on the basis of a similarity measure.

The systems published so far have shown interesting results, but not always it is clear what the result implies. What is a similar image? If the query specification is the image of a sunset and the engine recovers the next image on the same film taken at the same time and the same place is that a good result? If the example shows the picture of a man, and the query returns a woman, is that a good result? So image databases should be developed while also precisely formulating what "similar" is or means.

Also, from a technical point of view, in current systems of image retrieval, the integration of the search engine with a database is poorly developed. In effect, the search engines usually are based on file-based search; the bottom of database sophistication. And, the trustworthiness of the query scheme is usually poor. Most of the time, the search time is linear with the number of pictures in the database, a rather uninteresting performance. So, in order to make real progress with search engines, integration of computer vision technology with database technology should be reached.

In this paper we charter a route towards combining the expertise from computer vision and database research. The approach taken complements the one reported Virage [15] and Grosky1 [1][14] aiming to identify synergy between the two disciplines. In particular, we challenge the computing vision community to develop the database scheme for the discriminative features and effective algorithms to compute them for large collections in real time. From the database community we expect an incremental indexing scheme based on differentiation rather than sorting by commonalties. This provides the framework for handling the intrinsic continuous nature of sensory information, which can subsequently be searched using proximity queries.

The remainder of this paper is organized as follows. In section 2 we review the contributions and challenges of computer vision to deal with image databases. Likewise, in section 3 we touch upon progress made in the database arena. An overview of our experimental

architectural design is given in section 4. Research questions are summarized in section 5.

## 2 COMPUTER VISION SUPPORT FOR IMAGE RETRIEVAL

It is clear that the state of the art in computer vision does not permit the automatic full interpretation of an arbitrary scene. There is no general computational method enabling the analysis of an arbitrary scene, nor is it likely that such an algorithm will be found shortly. The problems are twofold.

The first problem is that an interpretation of image has no unique meaning. The meaning and interpretation of an image cannot be derived from the data alone. Contextual information and knowledge of the world is essential to deliver an interpretation of the picture. For example, an X-ray of the thorax may have "a medical image" as its prime interpretation for the general public. For a family doctor, the same X-ray may have an interpretation "something serious the matter", whereas the radiologist would never use these two interpretations of the image. For the radiologist the interpretation of medical image is so obvious that such a denomination would never occur. Similarly, as soon as the eyes fall on the X-ray, the interpretation fixes at a list of diseases and possible interpretations. The same interplay between picture and interpretation exist for general pictures. The context may induce selection of certain objects crucial in the interpretation of the scene. Such a selection is dependent on the viewer, the context and the purpose to mention a few prime factors. To summarize we have:

Thesis 1: *An interpretation of a picture depends on context, purpose and domain as other connotative factors. The interpretation of a complex scene is rarely unique and not always visual evidence may be present in the actual denomination.*

From a strictly data processing point of view no visual evidence may be present in the image to supporting the best description of the picture. As an example, no sun may be visible to describe a scene by "a sunny day". As a consequence of thesis 1, we get:

Consequence 1: Pictorial search is directed towards finding pictorially visible evidence while a semantic question needs to be directed on a semantic description added to the image.

The second problem is that a picture of one object may be instantiated in a million different data ways. The appearance of an object is depending on the variety in illumination circumstances, the shadows other objects cast on the object, the magnification, orientation and rotation with which the object falls on the field of view, and the projective deformation of the object.

Thesis 2: *The foremost problem of computer vision for image databases is that there are a million different image data arrays which depict the same object.*

The features computed from images are based on mathematical equivalence of illumination and reflectance patterns rather than on semantic equivalence of images. A user of an image database containing an object and a tumbled version of thereof, would expect that both images were retrieved as the semantic is exactly the same. From the computer vision point of view, the images differ in all intensity data values. Where the image data themselves are so variable, depending on accidental aspect as viewer position, and surrounding objects, the search is for other ways to capture the contents of the image irrespective of the pose of the object. This brings us to the definition of query by pictorial example (in the widest form; narrower forms may apply depending on the application domain):

Definition 1: *Query by pictorial example will localize in the database a picture of the same object regardless a difference in viewpoint, illumination, magnification, occluding foreground and occluded background and interfering surrounding objects.*

As an answer to the wide variety of views an object can take, in computer vision model-based image analysis has been defined. In the geometrical case, a deformable model is matched to the image data field. The state vector of the model captures the pose, orientation and the goodness of fit of the object. Preparing a model and developing a robust model match procedure requires considerable development effort specific for each domain. Results of model-based image analysis are usually suited for narrow domains of image images. Geometrical models work for one specific application, one specific sensor and recording circumstances and one specific set of questions. Similarly, the development of symbolic reasoning models requires a considerable amount of work, suited for a small image domain. Constraint resolution as the way to incorporate domain knowledge is unsuitable for broad image databases due to the required complexity of the model. The logical model for the general class of all images would be so big and complex that it would

become unpractical to handle if such a default knowledge model of the world can be defined at all.

Consequence 2: *Geometrical model-based and symbolic model-based reasoning analysis is suited for narrow domains.*

Examples of geometrical model-based domains are atlases containing all brands of roses, or the archive of X-ray photographs of the thorax. Examples of symbolic model-based reasoning are found in map reading, and in design based analysis.

Theses 1 and 2 express the impossibility of finding a generic solution to all vision problems by one completely specified algorithm. Still, enough progress has been made in computer vision in recent years in accessing the visual content of limited domains under limited circumstances in a wealth of useful and necessary techniques. Consequence 1 restricts pictorial search to pictorial evidence, and the second consequence defines the notion query by pictorial example. To avoid overconcentration on the many different images of one object, we conclude with:

Thesis 3: *From thesis 1 and definition 1 may be concluded that algorithms mapping similar images in an equivalent class is most crucial in performing query by example. However, in the end the focus should be on the discriminatory component of feature comparisons.*

## 2.1 KEY NOTIONS OF COMPUTER VISION FOR IMAGE RETRIEVAL

In this section, we address selected notions of computer vision in view of their potential role in image search engines.

Definition 2: *(Complete) segmentation assesses which subset in an image represents an object in the scene.*

An object may be a large variety of things: a toy, a tree in a forest, a pond, a tumor region, ink in the water but also a movement pattern in video such as waving goodbye and parking a car. An object's segmentation result is the complete and usually contiguous representation of the object in the data field. Many different computational techniques for segmentation exist. No single one of them is capable of handling a reasonable set of real world images. Segmentation is a complex task because objects may be partially occluded from sight by the presence of other objects, or hard to distinguish from the surrounding objects. A less difficult task is weak segmentation

Definition 3: Weak segmentation delivers a subset of the object's image.

Weak segmentation is apt when a salient subset of the object's image suits to identify the object from the rest of the world. In a weak segmentation of a solid object patches may cover the complete object but not necessarily so. The subset may consist of patches that are disconnected but even then all disconnected patches fall completely within the object's image. The limit case of weak segmentation occurs when the result is a set of isolated points, in our case hopefully distinctive for the object.

A weak segmentation result may be a sufficient basis to identify the presence of an object in an image by its prominent features. Where a house may be recognized as such by its roof, weak segmentation could delivers the set of visible, non-shadowed portion of roof tiles. Weak segmentation is useful for image retrieval when patches in the picture are conclusive to identify an object partially occluded behind another object. To give one more example, soccer players of one team can all be identified in the image by searching for patches with specific color combinations. To enhance the difference with the fans in the audience wearing the same clothes the patch may be extended to include some of the green for the grass they invariably have around them. Weak segmentation is the appropriate solution when an unknown fraction in the image is covered by occlusion. We will employ weak segmentation later on.

Complete or weak segmentation delivers a set in the image, from which features values may be derived. The set may be the entire image indicative for global properties, for example the average gray value or the most prominent hue in the image. The set may also be the complete image of an object after successful segmentation. In this case average qualities of the object's image are available from its color, texture, or shape. If the set is a result of weak segmentation, only the local features as color, texture, and, differential geometry are available as a weak segmentation only guarantees the pixels are from the same object. It does not guarantee the outline or extent of the patch is stable or will always follow the outline of the object.

For the purpose of retrieval by image content we are rarely interested in the statistics of all intensity or color values of all data points in the image as the image may be an accidental view of more than one object. We rather imply the storage of selected aspect values of the

object's image. These aspect values are computed as features of salient patches to be addressed by value.

Definition 4: *A salient patch is a patch of an object's image highly distinctive for that object.*

The distinction may be relative to random patches taken from that object (internally distinctive) as well as relative to patches taken from other objects (externally distinctive). When recognizing objects from a limited set we are mostly interested in external distinction, when we are interested in cognition of the object we are mostly interested in internally distinctive compared to patches taken from the universal set of images.

Feature values from salient patches are combined to identify the object. To that end the salient features are stored in sparsely occupied histograms serving as indices. Note that the histogram of an image or an incomplete segmentation might contain the feature values of the (salient) patches of more than one object. The histogram may contain feature values of other portions in the image, as well as some portions of the histogram may be missing when part of the object is occluded from sight. Subsequent processing steps should be capable of handling such littering of the histogram characterizing the object.

Following this strategy silently implies:

Hypothesis 1: *In a sparsely populated high dimensional feature space it is computationally easier to discard dissimilar objects than finding the best similar object.*

A key issue in image retrieval is invariance of features.

Definition 5: *A feature of an object is invariant if and only if the value of the feature remains the same regardless a change in the recording circumstances (different scale, location, viewpoint, illumination) or in the scene (occlusion, background, reflection from surrounding objects).*

Invariant features are the answer to the variations as discussed in thesis 2. Variations in illumination and the influence of the light source in the data scene require the consideration of some form of invariant features, insensitive to the undesired changes in the scene. As an example, when the object is at an unknown distance to the sensor, scale, translation and rotation invariant features may be in order to enable search for the same object. For outdoor scenes, the most difficult invariance to handle computationally originates from the great number of viewpoints one can take of an object. This requires viewpoint invariant features in order to recognize the object in the image (and not the image itself) as identical. The sensor- and scene-induced variance is so big compared to the object discriminatory variance that:

Consequence 3: *Sensor- and scene-induced variance should be dealt at the source, in the first processing steps.*

In the MPEG-7 standard this consequence has been recognized and incorporated. Paradoxically, with invariance a warning for their use should be posted.

Consequence 4: *The specification of the smallest applicable class of invariance is a necessary element of the specification of the search task.*

In the design of features, invariance should be as light as possible as all unnecessary invariance reduces the discriminatory power of the feature. Also, in its use, if illumination invariance is not necessary -when all pictures are standardized recordings of paintings- that invariance should not be included in the feature query set as it reduces the selective power.

## 3 ANATOMY OF A VISUAL SEARCH ENGINE

To make the state of the art in computer vision-based query engines more concrete consider the PictoSeek system [18] as a typical example for pictorial search engines. The various systems as cited above differ in the user interface, the type and implication of their internal feature set and in the way the result is presented but not in their general system architecture.

The system consists of the following computational blocks.

To define the object of query, an image is recorded or an image is selected from a repository. The aim to find a similar image in the database. Note that "similar image" may imply a partially identical image (as in the case of finding stamps), or a partially identical object in the image (as in the case of a stolen goods database), or a similarly styled image (as in the case of a fashion design support system). Some of the existing systems characterize the query image by parts with a typical average color. In the QBIC system [19], the user is free to sketch a region in the image with the preferred color. In the Picasso system [4], a sketch of the query object is given and the spatial arrangement of object is taken into account. The PictoSeek and Virage systems let the user select an example image to search for.

The essence of the query image is captured in a set of features after weak segmentation. These features

may cover each aspect of the image data, a measurement of intensity, shape, color or texture, movement or model adherence. A key issue is that the collection of features to use for querying is selective and an essential part of the query formulation. Feature extraction may be preceded by image preprocessing steps, model-matching or symbolic analysis of the image as well as a segmentation or a weak segmentation step. In all cases of image retrieval the process results in a condensation of the visual information in feature sets. Usually the features are calculated after a weak segmentation. That is the feature values are computed from a few salient patches in the image. The Photobook [16] and QBIC systems as typical examples concentrate on RGB-color features of selected regions in the image data field. The PictoSeek system concentrates on color features measured from salient patches in the image. The salient patches are the result of a weak segmentation procedure on the color shape pattern in the image. The size of the patch is reduced to one point or a pair of two close points. The color features are salient points and point pairs enable the identification of colored objects from just a few data points and their color values in the image.

The images in the set to be queried have been indexed by the same features during insertion of the image into the system, computed by the same processing steps to arrive at an identical feature description for each image. Processing may be different to normalize for a different sensor and different recording circumstances. Where possible, the restoration step will be specific for each different setting of the recordings. Moreover, as the algorithm reflects a computational method, the rule should be that the feature set derived from the image is independent from the implementation but refers to the essential sensory aspect of the object. The feature set is stored as an index for similarity comparison at run time.

As discussed before, attention is to be paid to the desired classes of invariance. For each image retrieval query a proper definition of the desired invariance is essential. A concise list of the most important invariance properties is:

Is the search for objects in different orientations and scales?

Is the search for objects in a large variety of scenes?

Is the search for objects in other kind of light?

Is the search for objects from different viewpoints?

Is the search for an object irrespective occlusion?

Note that these classes of invariances can each be turned off or on. As an example, in the search on a database for stamps, the viewpoint invariance will best be switched of as the recording of stamps is usually in frontal view only. This holds also for art. For the case of real world data the viewpoint invariance is a desirable property of the query as it does not ask for the object to be in precisely the same view. In the current state of the art of query engines, the explicit mentioning of invariance receives little attention. Invariance is usually handled by making a system specific for one application such as stamps or art. In large databases, the availability of a brand of invariances at the time of the query definition is essential. In the PictoSeek system both viewpoint invariant color and shape features, as well as illumination invariant features are included. The desired type of invariance determine the brand of features used in the query.

The features are computed for each of the salient patches in the image and captured in sparsely occupied histograms. These histograms indicate the presence of color and shape characteristic for the object. To permit faster access, in the PictoSeek system the histograms of each image is accessed via a hash table. The color histogram of each salient point pair (thus containing a color along each of the two axes) is summarized in 6 x 6 bins = 36 values. If a bin contains the color, the hash table sets the corresponding bit. The hash table thus is 36 bits wide, permitting a compromise between distinction of colors and speed of access. Other visual engines contain similar hash tables, but details are not always publicly available.

The actual query consists of a similarity search for the element in the queried set closest to the query image. As both the query images as the data set is captured in feature values, the similarity function operates between the feature sets. Again, the make the query useful, attention has to be paid to the selection of the similarity function. For the salient point sets, a similarity function is required which encompasses missing points in order to make the search occlusion invariant. In [21] (Santini), it is proposed to define perceptual similarity rather than mathematical similarity. In Picasso [17] the use of color features by their perceptual impression is proposed. These are important extensions of the available similarity functions.

After the query, the result is usually ranked in order of descending similarity. The query may be repeated with an image selected from the result set to achieve a form of visual browsing. The user interface can be made more intelligent [20] by relevance feedback.

For the PictoSeek system, on a 500 consumer object database, the viewpoint invariant color feature set with an EXOR similarity function results in 98% correct retrieval of a different picture of the same object. That is, different recordings at different camera positions of one object result in identifying the images as identical.

The recall rates appear to be robust against up to 60% occlusion of the object related to the spatial distribution of the patches over the object's view. They are also robust to a change in viewpoint up to 75 degrees. From a database point of view this is a marginal data set, but the point of demonstration was in the recall rate not (yet) in scalability. These are encouraging results, but there is a snag in these and almost all other reported figures in literature in the fact that many depends on the composition of the database and the suitability of the feature set for that specific domain and query. In fact, theses 1 and 2 in the section above guarantee that objective performance evaluation of image databases is an art on its own for which a simple standard solution does not exist.

## 4 DATABASE SUPPORT FOR IMAGE RETRIEVAL

Research in database management has reached a state where relational database systems are readily available to manage large amounts of data. The pervasive use and effectiveness of a DBMS can be attributed to the following:

Fact 1: *A database is built on a concise data model.*

This data model provides a high-level abstraction of the data items, their relationships, and their properties using a closed mathematical framework. As a consequence a strong asset of a database that it can be kept in a known state. All admissible input states are known a priori. This helps to ensure integrity.

Fact 2: *A database is based on a calculus and algebraic query language.*

Such an algebraic foundation guarantees a computational complete framework to retrieve and manipulate database portions without concern about their algorithmic behavior.

Fact 3: *A database serves physical independence.*

Physical independence permits a user to ignore the physical storage layout and the details of the algorithmic layers for its maintenance. The mapping from a query expressed against the logical data model is automatically compiled into the most efficient storage realization. The availability of such conceptual layers provides the means to optimize storage and querying at levels of abstraction. There is no need to reconsider the design of the entire database system when concentrating on solving complex queries. The modularity has been an important factor in the development of the database as an established field.

Fact 4: *A database functions on the basis of a closed world assumption.*

When the database is in a guaranteed state, and the query is within the list of admissible queries, a negative search result will carry important information: the requested is not contained in the database. The trustworthiness of such a negative answer has high significance, which cannot be easily guaranteed for sensory data indexed using the feature sets. As the image data are projected on sets of features (a non-reversible reduction of the image data), the object may be undetectable by the indexing features while it still is in the database somewhere. This leads to logically false responses to a query. It is not the result of the DB but rather of the way the information s being compressed during the indexing.

Despite the many relational and object-oriented database management systems produced over the last two decades, no system has been produced that solves all data management problems incurred. The database community estimates that at least 80% of all data still resides outside the confines of a DBMS, i.e. in bulk stores (audio, video, images) and files (word processing, consumer use).

The difficulty in extending the database-technology to sensory data is threefold. In the first place there is a problem with the size and storage structures of these new data types. This area of research quickly generates new solutions for spatial data structures. In the second place there is a difficulty in the specific type and language of query which come along with audio, video, images (and free text). And finally, there is the difficulty in the access to the content of these items.

Fact 5: *Query optimization aims at reducing evidently expensive solutions, not necessarily finding the most efficient solution.*

The size and storage structures and the query language are close to the database paradigm. Current technology enables a user to introduce new atomic types together with its operators that suit the application. Commercial systems such as Oracle and Informix already provide a step in this direction using their DataCartridge and DataBlade technology, respectively. They provide libraries of data structures with operations to support geometrical algorithms, for GIS applications, and image manipulation. Unfortunately, the extension modules provided are just

a first-generation solution. The modules have to be defined by someone fluent in database technology, because interaction with the various database components is tricky. Full use of the DBMS is limited as well, because the query optimizers are generally not equipped to exploit the properties of the user defined enhancements.

The difficulty in the access to the content is well outside the current database-paradigm. Sensory and text items do not permit an algebraic query set. The natural way to approach them is by example, a set of positive and negative examples, or other means of association. They also do not provide a concise data model. The variety and the interpretation of the content cannot be separated from the question. As mentioned above, images come with a multitude of interpretations. Often, only portions identified with a segmentation algorithm carry properties for retrieval. Where the database technology in its development has profited from the closed search space spanned by the data dictionary and the query set, such search spaces are not clearly defined for sensory (and free text) domains.

Construction of a sizeable image archive with its complementary query language is still an open research issue. Problems faced are with the type scheme, dealing with images requires a rich typing scheme surpassing the capabilities of object-relational systems, and with the computational scheme, query processing calls for proximity queries rather than yes/no decisions on objects under consideration.

This leads to overstressing the capabilities of a standard database schema. In a relational system each object, i.e. a fixed and limited set of attributes relevant for a large collection of similar objects. Alternatively, the information is encoded into a relational table and let the user interpret the results. The situation in object-oriented systems is not much better. Although they provide for a richer data model, it is not possible to partially include an object into the hierarchy or to let an object participate in multiple classes at the same time.

But even if we restrict our image archive to those cases where we can describe the properties in a database table (or class hierarchy), the computational model underlying a query language interpreter is too strict. In the DBMS field, a query predicate can be evaluated against a database with absolute precision. The predicate holds or is false. There is no middle way, nor a ranking scheme. Given incompleteness of the information available to classify an object this computational model is bound to fail.

What is needed in the handling of sensory data is: a proximity-based computational model where the DBMS returns the answers together with a value between 0 and 1 to indicate confidence that the query predicate holds, as well as index and parameter domains which ensure selective precision. If sensory data are categorized in qualifiers "large", "small", "yellow" rather than the physical measures not only the discriminatory power is lost to access 100,000+ databases, but also such qualifiers have no meaning without the question. "Yellow" is only "yellow" if "orange" or "ocher" is excluded from the query.

## 4.1 SYSTEMS UNDER DEVELOPMENT

In the large scale AMIS-project coordinated by the University of Amsterdam with participation of the CWI's database group, University of Utrecht's spatial data structures and the University of Twente with Quality of Service management as well as query optimizations. The envisioned architecture for experimentation consists of three layers of activity: storage and WWW access, query and feature detectors, and application layer.

The storage layer is build around the Monet extensible database system maintained at CWI. It provides access to both multi-media data stored locally and accessible through their URLs, and the multi-media archive maintained in the form of a large collection of CD-ROM's at University of Twente.

The input side is dealt with using a feature detector engine, which uses black & white box feature detectors to derive static feature values (vectors) from multi-media objects. Their results are kept around in feature indices for query support. The output side deals with effective support for querying the multi-media database. Both in terms of ranked responses using the feature indices and by exploitation of the inner structure of Web pages.

The top layer contains a number of applications to highlight and exploit various aspects of the platform. A simple search-engine like interface is provided to gain direct access using selections on the features maintained. An administration interface provides access to the database internals and statistics. A simple data entry form can be used to register new feature detectors and sources of information to index as soon as possible. University of Amsterdam's PictoSeek provides an engine to search the database using image filters and characteristics. The PictoSeek system can be viewed at http://www.wins.uva.nl/research/isis/ PicToSeek.

Next to the AMIS-project, in the companion digital media warehouse project, the middle layer comprises two components, roughly dealing with input and output. In this project, the collection of search techniques is extended with demonstrator video and audio service

demos, as well as novel ways to query semi-structured data, e.g. the Web at http://www.cwi. nl/~acoi.

## 5 RESEARCH AGENDA

To make true progress in image databases requires a clear delineation of the research problems. The following issues form the core.

Discrimination on the basis of image content is the sole means to locate images of interest.

The expressiveness of a visual imprint of an object cannot be captured in verbal or categorical expressions. As a consequence, the query should be (partly) specified in visual means.

A critical issue is the definition of useful image features to serve as an index, or rather a range of features expressed in the same data structures. In the paper we have made an attempt to order some of the features by classes of use and data types. This is the best opportunity to make contact to expand the connection further into the database paradigm.

An image may be present in the archive in 1 of a 1,000,000 different incarnations. Features, indices as well as similarity functions have to be able to deal with that. This is the prime research question in image databases as it implies a new view on indices and similarity measures, as well as a new view on image retrieval algorithms. For example, the fact that sensory similarity measures are a mathematical rather than a logical formalism requires new solutions.

To keep the query specific for not only the query object, but also the desired classes of invariance determine the implementation of the actual search. A new research topic is query optimization for a wide variety of invariances, i.e. query classes. In the current practice, the topic is underrated and usually solved by a system specific choice for one set of queries suited for one domain, e.g. art or trademarks.

Indexing an archive is never complete due to the open-ended list of possible queries formulated after the archive was defined. This requires dynamic solutions for the data dictionaries, search strategies as well as index optimizations.

Similarity retrieval is a sorting operation not a selecting operation to reduce the outcome set to one or N viable objects. The selection operation should be considered separately. Similarity is an essential part of dealing with sensory data.

Another essential element of sensory data is the handling of incomplete information. Part of the object may be out of sight while its presence should be detected still.

And, finally, when image databases work, the question how to integrate with other modalities such as free text, categorical information and sound returns on the agenda.

## REFERENCES


[1] W. Grosky, R. Mehrotra, Special Issue on Image Database Management, Computer, Vol. 22, No. 12, 1989.

[2] IFIP, Visual Database Systems I and II, Elsevier Science Publishers, North-Holland, 1989 and 1992.

[3] Image Databases and Multi-Media Search, (eds. A.W.M. Smeulders and R. Jain), Series on Software Engineering and Knowledge Engineering, Vol. 8, World Scientific, ISBN 981-02-3327-2, 1997.

[4] R. Jain, NSF Workshop on Visual Information Management Systems, SIGmod Record, Vol. 22, No. 3, 57-75, 1993.

[5] H. Levkowitz, G. T. Herman, GLHS: A Generalized Lightness, Hue, and Saturation Color Model, Graphical Models and Image Processing, Vol. 55, No. 4, 271-285, 1993.

[6] V. E. Ogle, M. Stonebraker: Chabot: Retrieval from a Relational Database of Images, IEEE Computer, Vol. 28, No. 9, 1995.

[7] S. Sclaroff, L. Taycher, M. la Cascia: ImageRover: A Content-based Image Browser for the World Wide Web, In: Proceedings of IEEE Workshop on Content-based Access and Video Libraries, CVPR, 1997.

[8] C. Frankel, M. Swain, Athitsos: Webseer: An Image Search Engine for the World Wide Web, TR-95-010, Boston University, 1995.

[9] S. A. Shafer: Using Color to Separate Reflection Components, Color Res. Appl., 10(4), 210-218, 1985.

[10] Proceedings of Storage and Retrieval for Image and Video Databases I, II, and III, Vol. 1,908; 2.185; and 2,420; W. Niblack and R. Jain, (eds.), SPIE, Bellingham, 1993, 1994 and 1995.

[11] Proceedings of Visual Information Systems: The First International Conference on Visual Information Systems, Melbourne, Victoria, Australia, 1996.

[12] Proceedings of Visual Information Systems: The Second International Conference on Visual Information Systems, San Diego, USA, 1997.

[13] A. del Bimbo, M. Mugnaini, P. Pala, F. Turco: PICASSO: visual querying by color perceptive regions. In: Proceedings of Visual Information Systems, San Diego, USA, 1997, 125 - 131.

[14] W. I. Grosky: Managing Multimedia Information in Database Systems. CACM 40(12): 72-80 (1997)

[15] Amarnath Gupta, Simone Santini, Ramesh Jain: In Search of Information in Visual Media. CACM 40(12): 34-42, (1997).

[16] A. Pentland, R. W. Picard, S. Sclaroff, Photobook: Tools for Content-based Manipulation of Image Databases, International Journal of Computer Vision, 18(3), 233-254, 1996.

[17] A. del Bimbo, M. Mugnaini, P. Pala, F. Turco, L. Verzucoli: Image retrieval by color regions. In: Image Analysis and Processing, Springer Verlag 1131, 180 - 185.

[18] T. Gevers, A.W.M. Smeulders, PicToSeek: A Content-based Image Search Engine for the World Wide Web, Proceedings of Visual Information Systems, San Diego, USA, 1997, 93-100.

[19] M. Flickner et al, Query by Image and Video Content: the QBIC system, IEEE Computer, 28(9), 1995.

[20] R. Schettini, A. Della Ventura, M. T. Artese: Color specification by visual interaction. The visual Computer vol 9-6, 143 - 150, 1992.

[21] S. Santini, R. Jain: Visual navigation in perceptual databases. In: Proceedings of Visual Information Systems, San Diego, USA, 1997, 101 - 108.

[22] J. R. Smith, Chang S.-F., VisualSeek: A Fully Automated Content-based Image Query System, In Proceedings of ACM Multimedia, 1996.

# Retrieving Pictures for Document Generation

Kees van Deemter
Information Technology Research Institute (ITRI)
University of Brighton
Lewes Road, Watts Building
Brighton, BN2 4GJ, United Kingdom

*Email:* Kees.van.Deemter@itri.brighton.ac.uk

## Abstract

This paper sketches how picture retrieval can be used by a document generation system for the inclusion of 'photographic' pictures in the documents generated by the system. The exposition is based on the What You See Is What You Meant (WYSIWYM) approach to knowledge editing and document generation (Power and Scott 1998, Scott et al. 1998). The paper describes an algorithm that makes use of a library of pictures, each of which is associated with a set of logical representations to facilitate retrieval. The paper introduces the twin notions of pictorial underspecificity and pictorial overspecificity, and focuses on the problems stemming from these two phenomena.

**Keywords:** Document Generation, Multimedia Information Retrieval, Semantics of Pictures

## 1 Pictures in generated documents

Pictorial illustrations can take many different forms. These differences can be crucial for a program having to interpret or generate them. In the present paper, we will study how a specific kind of picture can be included in generated documents. For present purposes, the documents to be generated are pharmaceutical Patient Information Leaflets (PILs), as exemplified by the leaflets in ABPI (1994). Before sketching a plan of the paper, we will briefly characterize how pictures are used in the PILs corpus.

Around 60% of the leaflets in ABPI (1994) contain pictures. Some pictures (often the ones depicting a medicine or its package) are just photographs; most are schematic, though rather complex sketches. Usually, the sketches depict a person performing an action. The vast majority of pictures are so complex that it would be extremely difficult to subject them to a 'compositional' analysis, describing them as consisting of well-defined syntactic parts each of which has a well-defined semantic interpretation. Consequently, and with a few notable exceptions where a picture has clearly distinguishable parts, it is difficult to see how they could be generated from smaller parts. (See e.g. Wahlster et al. 1993 for a generative approach to 'text + pictures' documents.) Something else than generation is required if we want to include such ('photographic') pictures in generated documents, therefore.

Luckily there also seems to be little *need* for genuine generation in this area, since the total number of different pictures used in the PILs produced by any one pharmaceutical company is very limited. Typically, the same picture is used a number of times, in different leaflets. This suggests that it would be attractive to let a document generation program select pictures from a *library*, in which each picture is coupled with a formal representation to characterize what the picture intends to convey. Taking everything into account, the program would allow an author to specify the content of each leaflet and, for a given item of information, determine whether or not it is in need of pictorial illustration. This idea, which has some similarities with the approach to the semantics of diagrams described in Pineda (1997), will be worked out in the remainder of this paper.

The plan of the paper is as follows. First, in Section 2, the 'What You See Is What You Meant' (WYSIWYM) approach to knowledge editing and document generation is introduced (Power et al. 1997, Scott et al. 1998). In Section 3, we sketch how WYSIWYM may be extended to deal with

pictorial information. In Section 4, some problems are discussed that need to be solved before the ideas in the previous section may be applied. These problems are associated with the concepts of *under*-and *over*specification, which the present approach reveals to be central to the way in which pictures convey information. Section 5 discusses how the formal representations in a pictorial library may be created. Section 6, finally, deals with some remaining problems and draws conclusions.

## 2 WYSIWYM FOR TEXT GENERATION

Many applications of Artificial Intelligence require editing of information expressed in a knowledge representation formalism of some sort. Expert systems are an obvious example; others are systems for generating documents, and for encoding design specifications. With most currently available support tools, knowledge editing has to be performed by knowledge engineers who are familiar with the representation formalism; the knowledge cannot easily be modified by domain experts or other interested parties. Much recent research has sought to simplify knowledge editing, e.g. by graphical browsers, or by input in controlled languages.

Elsewhere (Power et al. 1997, Scott et al. 1998), a new knowledge-editing method called 'WYSIWYM editing' has been introduced and motivated. WYSIWYM editing allows a domain expert to edit a knowledge base (KB) reliably by interacting with a *feedback text*, generated by the system, which presents both the knowledge already defined and the options for extending and modifying it. Knowledge is added or modified by menu-based choices which directly affect the knowledge base; the result is immediately displayed to the author by means of an automatically generated natural language feedback text: thus 'What You See Is What You Meant'.

WYSIWYM is a potentially powerful tool in many domains including some that are not commonly thought of as involving the editing of a knowledge representation formalism. Examples are the design of question-answering systems (where WYSIWYM assists the user in the construction of a query) and interface design for complex technical equipment (where WYSIWYM can assist the user to enter a command that is to be carried out by the system). Applications of various kinds are currently being investigated in a number of projects at the ITRI in Brighton. Among the domains currently studied are software documentation, pharmaceutical information and maritime law.

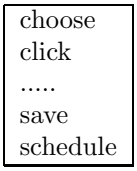
In the present paper we will be concerned with applications of WYSIWYM to document generation. What this means is that the knowledge base created with the help of WYSIWYM is used as input to an natural language generation (NLG) program, producing as output a document of some sort, for the benefit of an end user. Note that this type of WYSIWYM system makes use of two different NLG systems. At present, these two systems have been implemented as different modes of the same NLG program, one of which produces feedback texts (for the author) and the other output texts (for an end user). Let us take the DRAFTER system as an illustrative example. By interacting with the feedback texts generated by DRAFTER, the author defines a procedure for performing a task, e.g. the task of saving a document in a word processor. When a new knowledge base is created, DRAFTER assumes that its root will be some kind of procedure. A procedure instance is created, and assigned an identifier (for internal use only), e.g. `proc1`. The definition of the concept `procedure` specifies that every procedure has two attributes: a goal, and a method. The goal must be some kind of action, and the method must be a list of actions. This information is conveyed to the author through a feedback text

> Achieve **this goal** by applying *this method*.

with several special features.

- Undefined attributes are shown through anchors marked by the use of boldface or italics.
- A **boldface** anchor indicates that the attribute is obligatory: its value must be specified. An *italicized* anchor indicates that the attribute is optional.
- All anchors are mouse-sensitive. By clicking on an anchor, the author obtains a pop-up menu listing the permissible values of the attribute; by selecting one of these options, the author updates the knowledge base.

Although the anchors may be tackled in any order, we will assume that the author proceeds from left to right. Clicking on **this goal** yields a pop-up menu that lists all the types of actions that the system knows about:




(to save space, some options are omitted), from which the author selects 'save'. Although apparently selecting a word, the author is really selecting an option for editing the knowledge base. The program responds by creating a new instance, of type `save`, and adding it to the knowledge base as the value of the `goal` attribute on `proc1`:

```
procedure(proc1).
goal(proc1, save1).
save(save1).
```

From the updated knowledge base, the generator produces a new feedback text

> Save **this data** by applying *this method*.

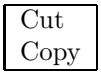
including an anchor representing the undefined `actee` attribute on the `save1` instance. Note that this text has been completely regenerated. It was not produced from the previous text merely by replacing the anchor **this goal** by a longer string. By continuing to make choices at anchors, the author might expand the knowledge base in the following sequence:

- Save the document by applying *this method*.
- Save the document by performing **this action** (*further actions*).
- Save the document by clicking on **this object** (*further actions*).
- Save the document by clicking on the button with **this label** (*further actions*).
- Save the document by clicking on the Save button (*further actions*).

At this point the knowledge base is potentially complete (no boldface anchors remain), so an *output text* can be generated and incorporated into the software manual.

> To save the document, click on the Save button.

To delete information, the author opens a pop-up menu on any span of the feedback text that presents a defined attribute. For instance, the span 'the document' presents the `actee` attribute on the instance `save1`. Clicking on this span in the feedback text, the author obtains the menu




If 'Cut' is selected, the instance that is currently the value of the `actee` attribute is removed to a buffer, leaving the attribute undefined. The resulting feedback text introduces an anchor in place of 'the document'.

> Save **this data** by clicking on the Save button (*further actions*).

When the buffer is full, the pop-up menus that open on anchors contain a 'Paste' option if the instance in the buffer is a suitable value for the relevant attribute.

One WYSIWYM application that is currently under development has the creation of Patient Information Leaflets (PILs) as its domain. The present, limited version of this system allows authors to enter information about possible side-effects of taking a medicine. The dialogue starts by a very general feedback text saying 'There is **a situation**',[1] whereupon the author can choose to expand the anchor '**a situation**' as either an atomic or a logically complex statement. Atomic statements can refer to people (e.g., a doctor or a patient), ordinary objects (e.g., a pill or a cream), or actions (e.g., one person giving medical treatment to another), among other things. The interface allows authors to create small KBs containing information expressible by feedback sentences such as *'If you are either pregnant or allergic to penicillin, then tell your doctor'*. It is this 'PILs' version of WYSIWYM that we will have in mind in the remainder of this paper.

## 3 An extension: WYSIWYM for text plus pictures

This should suffice for a brief introduction to WYSIWYM and its use for the generation of textual documents. In the remainder of this paper, we will be concerned with an extension of current WYSIWYM systems, where WYSIWYM is used to create output documents that contain pictures as well as words. In accordance with what we find in the vast majority of pictures in the PILs corpus, we will assume that pictures never add informa-

[1] Later versions of the PILs system are likely to use more informative, domain-dependent feedback texts as a starting point of the dialogue. These texts may also be geared to one particular section of the leaflet, e.g. 'If you suffer from **this condition**, then there is **this risk**'.

tion that is not expressed in the text.[2] Once we are able to deal with this relatively simple situation (in which we will speak of *illustrative* pictures), cases where pictures do add information can be handled provided we can devise a mechanism for determining what information should be expressed by means of text and what information by means of pictures. Before we move on, let us look at an example. The text of the example is as follows:

> **1.** Unscrew the cap and squeeze a small amount of ointment, about the size of a match-head, on to your little finger.
> **2.** Apply ointment to the inside of one nostril.
> **3.** Repeat for the other nostril.
> **4.** Close your nostrils by pressing the sides of the nose together for a moment. This will spread the ointment inside each nostril.

Clause **1** is illustrated by an image depicting the squeezing of ointment (Figure 1). Clause **2** is illustrated by a picture showing a finger entering the *left* nostril (Figure 2), while clause **3** is illustrated by a similar picture involving the *right* nostril (Figure 3):

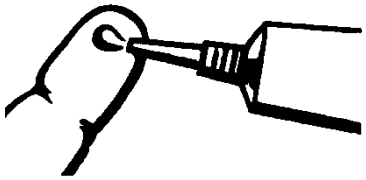


Figure 1: Illustration for clause **1**

Figure 2: Illustration for clause **2**

[2]The main source of exceptions to this generalization, in the PILs corpus, are cases where a picture makes more precise some quantitative information (e.g., 'Apply *a blob* of cream to the affected area of the skin') in the text. Other, far less frequent exceptions include cases where the text refers explicitly to what the picture expresses, e.g. 'Hold your inhaler as depicted in Fig.1'.

Figure 3: Illustration for clause **3**

Let us sketch informally how WYSIWYM could be extended in such a way that pictures of this kind can be included in output texts. The idea is to allow the author to indicate, for a given (mouse-sensitive) stretch $s$ of text, whether or not he or she would like to see $s$ illustrated. If yes, then the system will search its pictorial library to find a picture that matches the meaning of $s$. Determining this on the basis of the pictures alone would be extremely difficult. It would be possible to index the pictures making use of keywords, possibly making use of an existing classification scheme such as Iconclass (van de Waal 1985). What we propose in the present paper, however, is to develop and use a more precise, logically oriented classification scheme that will be outlined below. If a nonempty set of matching pictures is found, one picture is chosen randomly from this set. If the set is empty, the system tells the author that no illustration is currently possible, which can be a reason for expanding the library. Note that more sophisticated extensions of WYSIWYM could be built if the system *itself* was allowed to decide when to use pictures but once more, we will simplify and assume that it is always the author who decides. In what follows we will assume that the author always activates a stretch of text corresponding to *one* instance in the knowledge base. More specifically (but less crucially), we will look at cases in which the instance in question is an *action*.

In most work on picture retrieval, the library is assumed to be extremely large, often containing more than a million pictures (e.g. the picture collections of Keystone (1,300,000) and the Evening Standard (1,500,000)). The situation facing a pharmaceutical company that wants to author a new PIL leaflet is strikingly different. Based on the information in the PILs compendium (ABPI 1994), it seems unusual for one company to use more than 100 different pictures. Clearly, the task of annotating (i.e., designing formal representatons for) 100 pictures is quite feasible, especially

if it is taken into account that effective software tools can be used to do the work (see Section 5). The idea of the representations is to let them make explicit what information each picture *intends to convey*. Details that are there purely for embellishment are omitted. Likewise, things that are only depicted because it would be difficult *not* to depict them (such as the exact size or position of an object, or sometimes even properties like the gender or facial expression of a person) are omitted. In the cases that we will be concerned with, this means that the representations will focus on how an action of a given type should be carried out. More on the question of what information can be left out will be said in Section 5.

It has been observed that pictures tend to express 'vivid' information,[3] which can be expressed by a conjunction of positive literals (Levesque 1986). In what follows, we will use a notational format that matches this observation, using a fragment of predicate logic that is easily translated into the semantic networks used in existing WYSIWYM systems. Thus, we will write $\varphi(x, y)$ in the representation associated with a picture to assert that (the picture is intended to convey that) *there are* $x$ and $y$ such that $\varphi(x, y)$ is true. Thus, all variables in these representation are interpreted as if they were governed by an existential formula taking scope over at least the entire representation. (See Section 4.2 for details.) Assume $\mathfrak{S}$ is the 'activated' part of the database, that is, the part of the database for which a pictorial illustration is requested. In accordance with what has been said earlier, we assume that $\mathfrak{S}$ denotes an action. The representations in the KB of a WYSIWYM can be rendered in logical notation as follows:

$$\begin{array}{l}\text{TYPE}_0(e)\ \&\\ \text{ROLE}_1(e) = x_1\ \&\ ...\ \&\ \text{ROLE}_n(e) = x_n\ \&\\ \text{TYPE}_1(x_1)\ \&\ ...\ \&\ \text{TYPE}_n(x_n),\end{array}$$

where each of $e, x_1, .., x_n$ is either a variable or a constant.[4] This notation reflects the reliance, in the semantic nets used in DRAFTER-II, on *instances*, *types*, and *attribute/value pairs*. Each instance has one (1-ary) property, called its type, and can be the argument of any number of attributes, whose values are instances again. In the present logical notation, instances are rendered as variables or constants (depending on whether the instance is 'generic' or 'individual', cf. Sowa 1984), while types are denoted by the predicates $\text{TYPE}_i$; attributes are denoted by the functions $\text{ROLE}_i$. In the case of an *action* $e$, its type, $\text{TYPE}_0$, corresponds roughly with the meaning of a verb (e.g., it is of type *Squeeze* or *Apply*), saying what kind of action $e$ is; the attributes ($\text{ROLE}_1..\text{ROLE}_n$ above) applied to $e$ can be identified roughly with the semantic roles familiar from functional approaches to grammar (e.g. Fillmore 1977): $e$'s Actor, Actee, Target, etc. The values of these attributes ($x_i$ in the formula above), each of which can be either a variable or a constant, can be of any type $\text{TYPE}_i$ (e.g., a person, a medicine, or even an action) and each of them can have other attributes, and so on.

For example, suppose $\mathfrak{S}$ equals the following representation in the KB:

$$\begin{array}{l}Squeeze(e)\ \&\\ Actor(e) = Reader\ \&\\ Actee(e) = z\ \&\\ Ointment(z)\ \&\\ Quantity(z) = Small\ \&\\ Source(e) = t\ \&\\ Tube(t)\ \&\\ Target(e) = u\ \&\\ LittleFinger(u)\ \&\\ Owner(u) = Reader\end{array}$$

corresponding with part of clause 1. In this representation, $e, z, t$, and $u$ are variables, while *Reader* and (less elegantly, perhaps) *Small* are constants. Now, how can $\mathfrak{S}$ be used to select an appropriate picture? Given the availability of logical representations that capture the intended meaning of the pictures in the library as well as that of the relevant part of the text, at least two rules suggest themselves: one approaching $\mathfrak{S}$ from 'above' and one from 'below':

> Rule **A**: *Use the logically* **weakest** *picture whose representation logically* **implies** $\mathfrak{S}$.
> Rule **B**: *Use the logically* **strongest** *picture whose representation is logically* **implied by** $\mathfrak{S}$.

Logical strength, of course, is determined on the basis of the representations alone. Determining whether $\varphi$ logically implies $\psi$, where each of the two formulas is either an instance in the KB or

[3] Levesque 1986 (p.93) cites as the two main requirements for a KB to qualify as vivid that (1) 'There will be a one-to-one correspondence between a certain class of symbols in the KB and the objects of interest in the world' and (2) An analogous requirement on simple *relationships* of interest in the world.

[4] Present implementations of WYSIWYM do not distinguish between variables and constants, but it will sometimes be convenient to make this distinction in our discussions. Nothing serious will hinge on whether constants are available.

a semantic representation of a picture, is not difficult, given the fact that both are conjunctions of positive literals. (See also Section 5 for some relevant remarks.) Variations on **A** and **B** are possible, but it will be useful to explore each of these two rules in some detail.

## 4 Pictorial Under- and Over-specificity

The term underspecification is widely used in computational semantics, mainly to highlight the capacity of certain representation formalisms to denote several 'meanings' at the same time. (See e.g. van Deemter and Peters 1997 for a survey.) This capacity is useful for the modeling of ambiguous expressions in natural language, when they are ambiguous between these meanings. When applied to natural language itself, the terms 'underspecification' and 'underspecificity' have been used as synonyms for the word 'vagueness'. In conjunction with pictures, we will use the term 'underspecificity' in a closely related sense, to highlight the fact that a picture can sometimes contain less information than might be expected on the basis of its function. (This idea will be made precise below.) Conversely, we will speak of *over*specificity when a picture contains *more* information than might be expected. The two phenomena will be discussed in turn.

### 4.1 The problem of pictorial underspecificity

Suppose Rule **A** is used to choose a picture that matches the activated part $\mathfrak{S}$ of the KB:

> Rule **A**: *Use the logically* **weakest** *picture whose representation logically* **implies** $\mathfrak{S}$.

Let us apply this Rule to the first part of the example. Observe that the rule requires all the relevant information to be expressed in the picture, allowing some additional information. It would, therefore, require the representation of the picture in Figure 1 to contain all the information in $\mathfrak{S}$ including, for example,

1. *LittleFinger(u)*
2. *Owner(u)=Reader*
3. *Actor(e)=Reader*
4. *Ointment(z)*

Looking at the picture, repeated here as Figure 4, it seems clear that none of these facts are actually

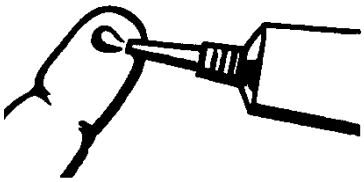


Figure 4: Illustration for clause **1**

expressed by the picture: (1) The finger could be an index finger as well as a little finger; (2) it is impossible to make out whose finger it is; (3) the picture does not say who does the squeezing; finally (4), it seems impossible to make out that the substance squeezed out of the tube is an ointment rather than, for example, a cream. Clearly, a picture can illustrate an item of information $\mathfrak{S}$ without necessarily expressing all the information in $\mathfrak{S}$ . Consequently, **A** is not the rule that we need if we want to find a picture that illustrates a part of the KB.

It is easily verified that Rule **B** does not run into the same problem as Rule **A**. Unlike the other rule, Rule **B** does not require that all the relevant information is expressed by the picture, as long as no *additional* information (i.e., information not in $\mathfrak{S}$ ) is expressed. Consequently, pictorial underspecificity is no problem for Rule **B**. In the case of the example of the previous section, this rule would simply choose a picture that contains nothing that is not in $\mathfrak{S}$ , selecting one that contains a maximum amount of information present in $\mathfrak{S}$ . Suppose, for example, that the library consists of only two pictures, which have the following representations:

**p1**:
$Squeeze(e')$ &
$Actee(e') = z'$ &
$OintmentOrCream(z')$ &
$Quant(z') = Small$ &
$Target(e') = u'$ &
$Finger(u')$

**p2**:
$Squeeze(e')$ &
$Actee(e') = z'$ &
$OintmentOrCream(z')$ &
$Target(e') = u'$ &
$Finger(u')$.

Then Rule **B** will choose **p1** to illustrate $\mathfrak{S}$ , since it is the most specific of the two, while both are implied by $\mathfrak{S}$. If, in addition, the library contained a picture

**p3**:
$Squeeze(e')$ &
$Actee(e') = z'$ &
$OintmentOrCream(z')$ &
$Target(e') = u'$ &
$Finger(u')$ &
$LittleFinger(u')$,

then the choice between **p1** and **p3** would be random.

## 4.2 The problem of pictorial overspecificity

Using Rule **B**, let us move on to the rest of the example. For ease of reading, we repeat the first three clauses:

> **1.** Unscrew the cap and squeeze a small amount of ointment, about the size of a match-head, on to your little finger.
> **2.** Apply ointment to the inside of one nostril.
> **3.** Repeat for the other nostril.

Let us assume that it is the squeezing and the applying, but not the unscrewing that needs repeating (clause **3**). In that case some appropriate feedback texts for (2) and (3) would be

> **2.** Apply the ointment to the inside of one nostril
> **3a.** Squeeze another small quantity of ointment from the tube on to one of your little fingers
> **3b.** Apply the ointment to the inside of the other nostril.

Simplifying slightly, (2) and (3b) correspond to the following statements in the KB:

$\mathfrak{S}_1$ :
$Apply(e_2)$ &
$Actor(e_2) = Reader$ &
$Target(e_2) = m$ &
$Nostril(m)$

$\mathfrak{S}_2$ :
$Apply(e_3)$ &
$Actor(e_2) = Reader$ &
$Target(e_3) = n$ &
$Nostril(n)$ &
$Nostril(m)$ &
$m \neq n$

Before we can study the effects of Rule **B**, we have to make assumptions about what the representations for the pictures are. Suppose the second and third picture (i.e., the ones showing actions involving the left and the right nostril) are associated with the following representations, respectively:

**Picture 2:**
$Apply(e'')$ &
$Target(e'') = a$ &
$LeftNostril(a)$

**Picture 3:**
$Apply(e''')$ &
$Target(e''') = b$ &
$RightNostril(b)$.

Note, incidentally, that these two representations make use of different sets of variables: $e''$ and $a$ versus $e'''$ and $b$, implying that the actions and nostrils involved may be distinct. (In this case, of course, the two *must* be distinct.) Whenever two representations share a variable, the two will be interpreted as speaking about the same object. In other words, we will assume that all our representations, whether they are in the KB or in the pictorial library, are interpreted by the same semantic model and by the same assignment of values to variables. In what follows, crucial use will be made of this assumption.

So, suppose we used the two representations mentioned above. Then Rule **B** fails to select any of these two pictures as applicable to $\mathfrak{S}_1$ , and the same is true for $\mathfrak{S}_2$. The reason is that Rule **B**, while allowing a picture to *omit* any information from the relevant part of the KB, does not allow a picture to express anything *in addition* to what the the relevant part of the KB requires. In the present case, this means that neither $\mathfrak{S}_1$ nor $\mathfrak{S}_2$ can be illustrated by means of a picture whose representation mentions that the nostril being treated is the left nostril, nor by one whose representation mentions that the nostril being treated is the right nostril. The source of the problem is the difficulty of depicting an arbitrary nostril; and even if a picture could do this, it would be even more difficult for the next picture to depict 'the other' nostril (again without saying whether it is left or right). Note that this would not be problematic for Rule **A**, with its tolerance towards additional information, but Rule **B** is not tolerant in this way.

## 4.3 Ways of overcoming both problems

We are looking for a way to deal with pictorial overspecificity (Section 4.2) as well as underspecificity (Section 4.1). Trying to solve the problem of underspecificity within Rule **A** does not seem very promising. Therefore, let us try to stick with Rule **B** and see how the problem may be resolved. One possibility would be to adapt the KB to match the semantic representations of the pictures. This could be done by replacing the property 'nostril' by a disjunction of two properties: 'LeftNostril' or 'RightNostril', along the following lines:

EITHER
(treat left nostril ; treat right nostril)
OR
(treat right nostril ; treat left nostril),

where the semicolon denotes temporal succession. This, however, would be a bad strategy. The resulting representation would be motivated by the issue of pictorial illustration only and would tend to give rise to cumbersome texts. Moreover, this approach would become impractical in the case of instructions that can be carried out in many different orders, and impossible if the number of possibilities is infinite, as in the case of

a. Choose a number;
b. Choose another number.

The only remaining approach seems to be to change the semantic representations of the pictures. This approach will be explored in the next section.

### 4.3.1 A revised scheme for semantic representations

If adding information to the representations in the KB is not a viable way of saving Rule **B** then removing information from the representations of the pictures seems the only option. Let us see how this may be done.

The first step is to keep different representations for the two pictures (i.e., the one where the left nostril is treated and the one where the right one is treated), but to leave out the bit that says which is which:

**Picture(2):**
$Apply(e'')$ &
$Target(e'') = a$ &
$Nostril(a)$

**Picture(3):**
$Apply(e''')$ &
$Target(e''') = b$ &
$Nostril(b)$

The result of this move is that, based on the revised representations, either picture can be used to illustrate $\mathfrak{S}_1$. As a result, an arbitrary one of them will be selected. A small problem with this approach is the fact that in some cases, the system needs to know 'which is which', namely when the activated part of the KB specifies that the right (or the left) nostril must be involved. This can happen in the context of an example, as in

> *'Lie on you side so that the drops will reach the part of your nose that is affected. For example, if the cavities in the right-hand side of your head need treatment, lie on your* right *side for a couple of minutes after using your medicine.'*

The difference between left and right can be more crucial, for example, when the place of the heart is at stake (*'If you feel a pain in your left arm, this can be a sign of a problem with your heart'*). To account for 'specific' cases like this, as well as for the 'nonspecific' cases discussed earlier, the library needs to contain two semantic representations for each picture: one with and one without left/right information:

**Pic(2, nonspec):**
$Apply(e'')$ & $Target(e'') = a$ & $Nostril(a)$

**Pic(2, spec):**
$Apply(e'')$ & $Target(e'') = a$ & $LeftNostril(a)$

**Pic(3, nonspec):**
$Apply(e''')$ & $Target(e''') = b$ & $Nostril(b)$

**Pic(3, spec):**
$Apply(e''')$&$Target(e''') = b$ &$RightNostril(b)$

The notion that one picture may have different semantic representations may have a wider validity, since one picture may be used to make rather different points, as has been pointed out repeatedly by practitioners of picture indexing schemes.

There is a more difficult problem with the approach outlined above, however, which can be il-

lustrated with the original example and especially its clause (3b), which referred to the inside of 'the other nostril'. Suppose the system has chosen the left nostril to illustrate clause (2) (i.e., the information $\mathfrak{S}_1$ ), using Picture(2) and its nonspecific representation. Then, clearly, Rule **B** predicts that $\mathfrak{S}_2$ can be illustrated by Picture(2) as well, so both nostrils (i.e., 'one nostril' and, a bit later, 'the other nostril') threaten to be illustrated by the same picture. This would clearly be inappropriate.

Our solution to this problem is to let illustration affect representation. When a picture comes to illustrate reality, this allows us to identify specific parts of the picture as depicting specific parts of the world. This can be viewed as additional information, causing the two to contain more information than each separately. In our situation, where the world is represented by a formula in the KB and the picture by its semantic representation, a separate KB could be used to record equations that come about as a result of the act of illustration. Each equation would equate a variable in the semantic representation with a variable or constant in the KB. We will explore a slightly different approach, whereby these equations are added to the pictorial representation of the picture. (The additions will remain valid throughout the entire leaflet, but once a new leaflet is initiated, the picture regains its original, smaller representation.) Let us return to our example to show the implications of this idea.

**Claim:** Pic(2, nonspec) cannot illustrate $\mathfrak{S}_1$ and $\mathfrak{S}_2$ in the same document.

*Suppose*, firstly, Pic(2, nonspec) is used to illustrate $\mathfrak{S}_1$ . Then two equations are added to Pic(2, nonspec), resulting in Pic(2, nonspec)$'$ = Pic(2, nonspec) & $e'' = e_2$ & $a = m$.

*Suppose*, after that, Pic(2, nonspec) is used to illustrate $\mathfrak{S}_2$. Then two more equations are added to Pic(2, nonspec)$'$, resulting in Pic(2, nonspec)$''$ = Pic(2, nonspec) & $e'' = e_2$ & $e'' = e_3$ & $a = m$ & $a = n$. Given the statement $m \neq n$ in $\mathfrak{S}_2$ , the combination $a = m$ & $a = n$ is inconsistent. It is easy to see that the same result would be obtained if $\mathfrak{S}_1$ and $\mathfrak{S}_2$ were illustrated in reverse order. □

Note that the above proposal lets the act of illustration change the syntactic form (i.e., the physical shape) of the representation of a picture. More subtly, one might envisage a modeltheoretic approach whereby only those assignments to variables are taken into account that send certain variables (i.e., the ones equated in the other approach) to the same value. Note, incidentally, that coreference can also be enforced by letting different representations share variables, as will be illustrated in section 6.

We have discussed a case in which Rule **B** is unable to decide whether Pic(2, nonspec) or Pic(3, nonspec) should be used and where the situation requires that the system remembers which of the two has been chosen. The mechanism for storing this information, which relies on equating variables, can easily be integrated with the procedure determining that (the representation of) the picture is logically implied by the representation in the KB, since this procedure will work by comparing each of the conjuncts in the two representations and determining whether they are unifiable. To exemplify, let us return to our initial example in Section 4.1, involving $\mathfrak{S}$ and the pictures **p1**, **p2**, **p3**. By unifying variables pairwise, it is easy to prove the following logical implications, thereby establishing that $\mathfrak{S}$ logically implies the representation of picture **p1**:

$$
\begin{aligned}
&Squeeze(e) \models Squeeze(e') \\
&Actee(e) = z \models Actee(e') = z' \\
&Ointment(z) \models OintmentOrCream(z') \\
&Quant(z) = Small \models Quant(z') = Small \\
&Target(e) = u \models Target(e') = u' \\
&LittleFinger(u) \models Finger(u')^5
\end{aligned}
$$

In this example, $e$ is unified with $e'$, $z$ with $z'$, and $u$ with $u'$. In exceptional cases, it will be possible to find different unifiers each of which lead to a logical implication. For example, the representation of the picture may contain conjuncts $Squeeze(e_1)$ and $Squeeze(e_2)$ (i.e., the picture depicts two actions) while the representation in the KB contains the conjuncts $Squeeze(e')$ and $Squeeze(e'')$. In cases like this, the pairing of variables is subject to indeterminism even after it has been decided what part of the KB will be illustrated by which picture.

[5] Cases where a literal occurring in the representation of the picture follows only from a *conjunction* of literals ocurring in the KB can, in principle, arise. (For example, the two literals $OintmentOrCream(z)$ and $OintmentOrGel(z)$ logically imply that $Ointment(z)$ if the KB says that nothing can be a gel as well as a cream.) This complication, however, seems unlikely to arise in practice.

## 5 WYSIWYM FOR THE CREATION OF A PICTORIAL LIBRARY

In Section 3, we briefly mentioned the question of what information can or must be left out in the semantic representation of a picture. This is no easy question, since one and the same property, such as the fact that a person is depicted as a child, can sometimes be essential (e.g. in the pictures that warn against improper use by children) and sometimes inessential (i.e., when *any* person could have been depicted). But, the task of annotating is simplified considerably by the fact that the representations that can be used can be thought of as the well-formed formulas of a precisely defined formal language.

What is this formal language $R$ to which semantic representations of pictures must belong? As a first – and fairly close – approximation, $R$ can be equated with the language defined by the T-box of the KB (henceforth, the language $L$). This approximation arises when it is granted that what can be depicted is the same as what can be described. We have seen that this is not exactly true. In particular, the 'vividness' of pictures implies that pictures never depict logically complex information.[6] Thus, $R$ is at most a ('vivid') subset of $L$.

Other, minor, adaptations may be necessary. For example, we have seen that $R$ may contain certain nonlogical constants that are not defined in the T-box, such as the constant 'cream-or-ointment'. (This constant reflects the fact that it is difficult to depict something as a cream rather than an ointment, or the other way round.) Additions of constants like this are limited in number, are easily made, and need not concern us here. In essence, therefore, one can assume that $R$ is the largest 'vivid' subset of the language $L$.

Granted that $R$ is an easily definable subset of the language defined by the T-box, we can think of *constructing a representation* in $R$ as *creating a* KB of some sort. Consequently, it must be possible to construct such representations by means of a WYSIWYM system, thus causing the WYSIWYM interface to be used for two very different purposes.

[6] Cf. Section 3. In fact, the PILs corpus does contain a few pictures that express logically complex information. All of these depict an action which is 'crossed out' by two crossing lines. These pictures may be analysed as expressing a negation or a prohibition. Exceptions of this kind are so infrequent that we will neglect them here.

Here is a simple example of how a representation may be created with help from WYSIWYM. Consider the representation Pic(2,nonspec). To create this representation, starting from the initial feedback text saying 'There is **a situation**', the anchor **a situation** must be replaced by an atomic statement descibing an action, more specifically the *apply* action. This will lead to a structure in the KB where there is an instance *apply* which is a particular type of *Action*, having four attributes, namely *Actor*, *Source*, *Actee*, and *Target*. The associated feedback text might say 'A person applies **something** to **a body part**. Clicking on **a body part**, a menu will suggest several options including *nostril*. When this is chosen, the system will extend the representation by specifying that the Target of the action is a nostril. Note that, as in any WYSIWYM interface, the person creating this representation has no control over the choice of variables and constants but these choices are irrelevant anyway, except when a term is required to corefer with an earlier-introduced term, in which case WYSIWYM will allow the user to express this requirement.

Before WYSIWYM can be used for the novel purpose of creating semantic representations for pictures, certain adaptations may have to be made. For example, the system might present a formal rendering of the content of the KB, perhaps in the form of logical formulas, in addition to the feedback texts normally produced by the WYSIWYM interface. In addition, the feedback texts will need to be adapted slightly, because they have to describe the bare content of the picture (e.g., a type of action) rather than, for instance, a conditional statement or an injunction to do something. Finally, it would be useful if the system gave additional feedback in cases where no picture could be found or, perhaps, where Rule **B** left several candidates for the system to choose between. There are, however, no problems in principle with extending WYSIWYM to creating semantic representations for pictures in the library.

## 6 Conclusion

Document generation is not normally associated with retrieval. The present paper, however, has outlined an approach to document generation that makes crucial use of picture retrieval when it comes to including pictures into the documents generated. It should be stressed that the resulting retrieval task is quite different from the task facing most picture retrieval systems. It has been ar-

gued forcefully that, in the case of large libraries, it is difficult to annotate the pictures with sufficient information to make a successful search possible (Enser 1995). In a typical document generation task, such as the task of generating PILs leaflets, the search space is much smaller than in most other retrieval tasks (see Section 3). As a result, more elaborate indexing techniques can be used including ones based on formal logic, and it is such techniques that are explored in this paper.

It will be clear that there are many issues that need to be resolved before the approach outlined here can be operational. For example, the issue of locating pictures on a page needs to be addressed and the same is true for the related issue of document transparency. Textual and graphical information must be coordinated in such a way that cross-modal coreference relations are easily understood by a reader of the generated document, for example. (See e.g. Andre and Rist 1995.)

A possibly unexpected feature of the approach outlined in this paper is that it focuses on individual pictures, without proposing a special mechanism for dealing with sequences. This promts the question of how various kinds of coherence between elements of a sequence may be guaranteed. We will briefly discuss a few problems in this area and indicate tentatively how they may be resolved.

Perhaps the most pressing problem with coherence within a sequence has to do with reference. Suppose one were to illustrate the earlier-mentioned statements $\mathfrak{S}_1$ and $\mathfrak{S}_2$ through a sequence of two pictures, the first of which (illustrating $\mathfrak{S}_1$) depicts *a man* and the second (illustrating $\mathfrak{S}_2$) *a woman* as the 'owner' of the nose. This would destroy the notion that the two pictures intend to illustrate how one person can go though the actions detailed in $\mathfrak{S}_1$ and $\mathfrak{S}_2$. Yet, nothing currently prevents this type of incoherence. There is, however, a way in which this problem may be avoided, namely by exploiting the possibility that the representations for two pictures share a variable. Since the variables in all representations (in the KB as well as in the library) are interpreted by means of *one* value assignment, this means that the pictures will be interpreted as expressing information about the same object. Let us exemplify how this fact may be exploited to solve the problem just described.

Suppose the library contains 3 pictures, $x$, $y$, and $z$, while each of the first two depicts a woman, and $z$ depicts a man. Suppose gender is not an attribute of persons, according to the T-box of the KB, so gender cannot be used to distinguish $x$ and $y$ from $z$ in the KB. (If gender entered the representations of $x$, $y$, and $z$, then Rule **B** would not allow these pictures to illustrate a statement that did not itself involve gender.) Then we can try to encode the difference between the pictures by letting $x$ and $y$'s representations, $\text{Pic}_x$ and $\text{Pic}_y$, share a variable that $z$'s representation, $\text{Pic}_z$, does not share. Thus, for example, we might have the following library:

**Pic**$_x$**:** $Squeeze(e)$ & ... & $Actor(e) = a$
**Pic**$_y$**:** $Apply(e')$ & ... & $Actor(e') = a$
**Pic**$_z$**:** $Apply(e'')$ & ... & $Actor(e'') = b$,

where the variable $a$ is shared between the first two representations. Suppose the relevant part of the KB consists of these two statements:

$\mathfrak{S}_1$ : $Apply(e_1)$&...&$Actor(e_1) = Reader$
$\mathfrak{S}_2$ : $Apply(e_2)$&...&$Actor(e_2) = Reader$

Now suppose the picture $x$ illustrates $\mathfrak{S}_1$. As a result, $Pic_x$ becomes $\text{Pic}'_x$:

**Pic**$'_x$**:**
$Squeeze(e)$ &... & $Actor(e) = a$
& $e = e'$ & $a = Reader$.

In order to prevent the picture $z$ (instead of the intended picture $y$) from illustrating $\mathfrak{S}_2$, we can introduce a special convention saying that, in the choice between pictures (more precisely, the choice between their representations) selected by Rule **B**, preference is given to the ones that lead to *fewest new equations*. This idea is akin to the idea, in abductive theories of language interpretation, that those interpretations are preferred that require one to make the fewest assumptions. The convention suggested here would predict a preference of $y$ over $z$ as an illustration for $\mathfrak{S}_2$, once $x$ has been used to illustrate $\mathfrak{S}_1$. The preferred illustration leads to the new equation $e_2 = e'$ only (note that the equation $a = Reader$ is already in place), while the less preferred one leads to the new equations $e_2 = e''$ and $b = Reader$. The preferred illustration gives rise to a sequence of illustrations that is coherent in the sense that it uses similar pictures for the illustration of similar events.

Note that the approach suggested here would not have the desired effect if $z$, instead of one of the

other two pictures, had been chosen to illustrate $\mathfrak{S}_1$ before the choice of an illustration for $\mathfrak{S}_2$. To maximize the chances of finding suitable illustrations, it might be wise to give precedence to illustrations that make use of variables that occur in many other representations as well, thus favouring $x$ and $y$ over $z$. Alternatively, one could shift to an approach based on finding illustrations for entire *sequences* of pictures. This approach might ultimately be more accurate because, in some cases, it does not make sense to illustrate only part of a series of things (e.g. actions) while an approach based on individual pictures cannot anticipate whether illustrations can be found for other parts of the same series.

Another problem in the area of pictorial coherence is similar to the previous one except that it does not involve reference but the way in which pictures depict. Pictures depicting the same state of affairs may differ in terms of their size, colour, or pictorial 'style'. In the PILs domain, for example, some companies have different sequences of pictures in their libraries that express exactly the same information, presumably to suit different sizes of leaflets. Sometimes there are other differences as well, involving the amount of detail, the width of brush strokes, etc. As in the case of the man and the woman (see above), the approach sketched in the paper can lead to incoherent sequences of pictures. This problem can be tackled by separating the *selection* process from the *presentation* process. Selection would continue to abstract away from the way in which the pictures depict, but instead of selecting an arbitrary picture having an appropriate representation, a set of pictures would be selected, each of which has the same representation. The choice of the most appropriate picture from this set can either use an extension of the semantic representations described in this paper (see van Deemter 1998 for some ideas) or by means of some different procedure.

## 7 Acknowledgment

Thanks are due to Richard Power for extremely helpful discussions on the topic of this paper.

## 8 References

[ABPI **1997**] The Association of the British Pharmaceutical Industry, *1996-1997* ABPI *Compendium of Patient Information Leaflets.*

[**André & Rist 1995**] E.André and T.Rist, "Generating Coherent Presentations Employing Textual and Visual Material", *Artificial Intelligence Review* **9**: 147-165.

[**van Deemter and Peters**] van Deemter, K. and Peters, S. (Eds.) (1996), *Semantic Ambiguity and Underspecification*, CSLI Publications, Stanford, Ca.

[**van Deemter 1998**] K.van Deemter "Representations for Multimedia Coreference", in Proc. of ECAI workshop on Combining AI and Graphics for the Interface of the Future. Brighton, Aug.1998.

[**Enser 1995**] P.G.B.Enser, "Progress in Documentation; Pictorial Information Retrieval", *Journal of Documentation*, Vol.51, No.2, June 1995, pp.126-170.

[**Levesque 1986**] H.J.Levesque, "Making Believers out of Computers", *Artificial Intelligence* 30, pp.81-108.

[**Pineda & Garza 1998**] Pineda L. A., Garza G., "A Model for Multimodal Reference Resolution", in Procs of Workshop on Referring Phenomena in a Multimedia Context and Their Computational Treatment, E.Andr (Ed.), Meeting of ACL-SIGMEDIA, Madrid, (1997), pp. 99–117.

[**Power & Scott (1998)**] R.Power and D.Scott, "Multilingual Authoring using Feedback Texts", in Proc. of COLING/ACL conference, Montreal 1998.

[**Rist and E. André 1991**] T.Rist and E.André, "From Presentation Tasks to Pictures: Towards a Computational Approach to Graphics Design", in: ECAI92, pp. 764-768.

[**Scott et al. 1998**] D.Scott, R.Power, and R.Evans, "Generation as a Solution to its own Problem", in Proc. of 9th International Workshop on Natural Language Generation, Niagara, Canada, Aug. 1998.

[**van de Waal 1985**] H. van de Waal (completed and edited by L.D. Couprie, R.H. Fuchs, E.Tholen, G. Vellekoop, a.o.), "ICONCLASS; An iconographic classification system." , Amsterdam 1973-1985 (17 vols). ISBN 0-7204-8264-X. See also the WWW at `<http://iconclass.let.ruu.nl/home.html>`.

[**Wahlster et al. 1993**] W.Wahlster, E.Andr'e, W.Finkler, H.-J. Profitlich, and Th.Rist, "Plan-based Integration of Natural Language and Graphics Generation", *Artificial Intelligence* 63, p.387-427.

# The THISL Spoken Document Retrieval System

Steve Renals and Dave Abberley
University of Sheffield
Department of Computer Science
Sheffield S1 4DP, UK
{s.renals,d.abberley}@dcs.shef.ac.uk

## ABSTRACT

THISL is an ESPRIT Long Term Research Project focused the development and construction of a system to items from an archive of television and radio news broadcasts. In this paper we outline our spoken document retrieval system based on the ABBOT speech recognizer and a text retrieval system based on Okapi term-weighting . The system has been evaluated as part of the TREC-6 and TREC-7 spoken document retrieval evaluations and we report on the results of the TREC-7 evaluation based on a document collection of 100 hours of North American broadcast news.

**Keywords:** Multimedia Information Retrieval; Spoken Document Retrieval; Speech Recognition; Broadcast Data.

## 1 INTRODUCTION

THISL is an ESPRIT Long Term Research project in the area of speech retrieval. It is concerned with the construction of a system which performs good recognition of broadcast speech from television and radio news programmes, from which it can produce multimedia indexing data. The project is concentrating on British and American English applications, with work in progress on a French language system. In particular, the main goal of the project is to develop a system suitable for a BBC newsroom application. The resulting system may be regarded as a "news-on-demand" application in which specific portions of a broadcast may be retrieved in response to a spoken request from the user.

There are two principal approaches to the task of spoken document retrieval. The *phone-based* approach processes the audio data with a lightweight speech recognizer to produce either a phone transcription or a some kind of phone lattice. This data may then be directly indexed or used for word spotting. The *word-based* approach applies a complete large vocabulary speech recognition system to the audio track to produce a word-level transcription; at this point the problem may be treated as standard text retrieval (modulo speech recognizer errors).

The phone-based approach is not restricted by a fixed vocabulary. Since the archiving process only involves phone recognition there is less computational overhead for archiving — although, as discussed in section 6, this may only reduce the computation by around a factor of 2 compared with large vocabulary continuous speech recognition. The phone-based approaches are typically based on indexing overlapping sequences of $n$ phones, where $n$ typically takes values 3 or 4. Ng and Zue [13] have shown that for small spoken document collections this technique can produce average precisions close to that of the reference text, if the perfect phone transcription is known. Working with the output of an automatic phone recognizer results in a relative degradation in performance of around 30%. This approach has also been adopted by Schauble and coworkers [18] and by Smeaton et al [21] who have performed experiments on the TREC-6 spoken document retrieval evaluation and an application based on an archive of RTE news bulletins.

An alternative phone-based approach uses a word spotter, which enables pronunciation constraints to be incorporated. The phone recognizer may be used to produce a phone probability matrix or a phone lattice (graph), which can be used as the input to a word spotting algorithm. Although such algorithms can run many times faster than real-time [4], they are limited by a linear dependence on the size of the archive. However the process may be made considerably more efficient by using an index based on phone sequences to preselect the areas of speech over which the word spotter should be applied. This approach was first suggested by Dharanipragada and Roukos [7] and has also been used by Kraaij et al [12], and may be regarded as a rescoring of the top ranked documents returned by a purely phone-based system. This approach can result in many false alarms; these may be

reduced by including possible "confuser words" in the word-spotter.

In the THISL project we have adopted a word-based approach, similar to that employed by several other groups (eg [2, 9]). This approach requires more computation than phone-based approaches, since a full large vocabulary decoding needs to be applied to the entire archive. However, it enables the constraints of the pronunciation dictionary and language model to be applied. and text retrieval is more robust when applied to words than phone n-grams. Aside from computational considerations, the most frequently cited drawback of this approach is the problem of out-of-vocabulary words. We do not believe that this is a significant problem, and is certainly outweighed by the advantages of the word-based approach. Indeed, of the ad-hoc topics used in the past five TREC evaluations (TRECs 3–7), 9 out of 900 query words were out of vocabulary relative to the 65,000 word vocabulary used in the experiments reported in this paper. This 1% out-of-vocabulary rate corresponds with that we typically observe when recognizing broadcast news data.

In this paper we present the THISL system for spoken document retrieval which is based on the ABBOT large vocabulary continuous speech recognition (LVCSR) system [17] and well-understood probabilistic text retrieval techniques. In section 2, we outline the ABBOT LVCSR system, focusing on those features that make it particularly appropriate for spoken document retrieval. Section 3 describe the text retrieval methods that we used, and we discuss two possible enhancements: the use of multiple transcriptions (section 4) and the use of query expansion (section 5). A series of experiments were carried out as part of our participation in the TREC-7 spoken document retrieval track, and these are described in section 6. Finally, section 7 discusses some conclusions and outlines our current and future research directions in this area.

## 2 SPEECH RECOGNITION USING ABBOT

We have used the ABBOT LVCSR system developed at the Universities of Cambridge and Sheffield [17]. The four principal components of a probabilistic LVCSR system are the signal processing module (which typically transforms the time domain waveform into a sequence of acoustic feature vectors), the acoustic model (which models phones in terms of the acoustic features), the pronunciation dictionary and the language model (which gives a probability of occurance of any sequence of words).

### 2.1 CONNECTIONIST ACOUSTIC MODEL

ABBOT differs from most other state-of-the-art LVCSR systems in that it has an acoustic model based on connectionist networks [3]. Although this model may still be interpreted as a type of HMM, it differs from traditional HMMs by directly estimating the posterior probability of each phone given the acoustic features, rather than the likelihood of that phone generating the acoustics. Posterior probability estimation may be performed by a connectionist network (or set of networks) trained to classify phones. In ABBOT, a set of recurrent networks [16] is used. Direct estimation of the posterior probability distribution using a connectionist network is attractive since fewer parameters are required for the connectionist model (the posterior distribution is typically less complex than the likelihood) and connectionist architectures make very few assumptions on the form of the distribution.

Currently, the acoustic model used in the THISL system consists of two recurrent networks with 53 context-independent phone classes (plus silence). One network estimates the phone posterior probability distribution for each frame given a sequence of 12th order perceptual linear prediction features [8]. The other network performs the same distribution estimation with features presented in reverse order (since recurrent networks are time-asymmetric) and the two probability estimates are averaged in the log domain. Each network contains 384 state units, resulting in a total of about 350 000 acoustic model parameters. The 54 context-independent phone models may be expanded to a set of context-dependent phone models, the context classes being arrived at via a decision tree algorithm. A context class network is used for each context-independent phone class, which (when combined with the context-independent phone probabilities) results in a context-dependent phone probability [11]. In the experiments reported in this paper, the system was trained on 100 hours of broadcast news data released by the Linguistic Data Consortium[1]. Twenty-four hours of this data is not transcribed (commercials, local news, etc.), and further sixteen hours was discarded as being below a confidence threshold before training after computing the average log likelihood per frame during a Viterbi alignment. This system is a simplified version of that used by the CU-CON group in the 1997 DARPA evaluation of broadcast news speech recognition systems (hub 4) [5].

The British English system, used for the BBC application, is currently trained on about twenty-four hours of acoustic training data collected and tran-

[1] The first 100 hours of the so-called Hub 4 training data — see the LDC website at http://www.ldc.upenn.edu/

scribed by the BBC Research Department.

## 2.2 Pronunciation Dictionary

The pronunciation dictionary specifies the finite set of words that may be output by the speech recognizer, and gives at least one pronunciation (ie phone sequence) for each. In the current system, for American English, a dictionary of 65 532 words is used, with a total of about 72 000 pronunciations. The figures are similar for the British English system.

## 2.3 Language Model

The role of the language model is to estimate the probability $P(w_1 w_2 \dots w_n)$ of a string of words $w_1 w_2 \dots w_n$. This may be decomposed as:

$$P(w_1 w_2 \dots w_n) = P(w_1)P(w_2|w_1)\dots P(w_n|w_{n-1}\dots w_1). \quad (1)$$

If it is assumed that the probability of a word is dependent only on the two preceding words, then (1) may be approximated as:

$$P(w_1 w_2 \dots w_n) = P(w_1)P(w_2|w_1)\dots P(w_n|w_{n-1}, w_{n-2}). \quad (2)$$

This is referred to as a trigram language model. Simple maximum likelihood estimation of trigram probabilities of the form $P(w_n|w_{n-1}, w_{n-2})$ will result in zero probabilities for any trigrams that do not occur in the training data. To prevent this a smoothing technique must be employed. Backing-off involves a portion of the probability mass being reserved for unseen trigrams; this probability mass is split using bigram estimates, and the process may be continued recursively [10].

For the experiments reported here a backed-off trigram language model containing 65 532 unigrams, 7.1 million bigrams and 24.0 million trigrams was estimated from a variety of sources including 132 million words of transcribed broadcast news data and 153 million words of newswire data. The vocabulary was selected by including all the words from the transcription of the acoustic training data, made up to 65,532 words using the most frequent words extracted from the broadcast news text corpus (ignoring common misspellings and obvious text processing errors).

## 2.4 Search

The search problem in speech recognition may be posed as follows: what is the most probable sequence of word models (or phone models) given the observed acoustics, the acoustic model, the language model and the pronunciation dictionary? Potentially the search space is huge: for example, in the system described above anyone of 65 532 words could start each 16ms. To efficiently evaluate this search space, the Abbot system employs a start-synchronous stack-based search, with substantial pruning of improbable hypotheses [14]. In particular, the search algorithm makes direct use of the posterior probability estimates produced by the neural network acoustic model by pruning all those phones which have an estimated local posterior probability below a threshold. On average, this enables about 70% of the phonetic search space to be pruned at any one time, with a minimal increase in search error.

# 3 Text Retrieval

In our initial work on spoken document retrieval [1], we used the PRISE text retrieval system[2] developed by NIST. More recently we have developed an Okapi-style testbed system "textbook" probabilistic system, using a stop list, the Porter stemming algorithm and the Okapi term weighting function. Specifically we used the term weighting function $CW(t,d)$ for a term $t$ and a document $d$ given in [15]:

$$CW(t,d) = \frac{CFW(t) * TF(t,d) * (K+1)}{K((1-b) + b * NDL(d)) + TF(t,d)}. \quad (3)$$

$TF(t,d)$ is the frequency of term $t$ in document $d$, $NDL(d)$ is the normalized document length of $d$:

$$NDL(d) = \frac{DL(d)}{\overline{DL}}, \quad (4)$$

where $DL(d)$ is the length of document $d$ (ie the number of unstopped terms in $d$). $CFW(t)$ is the collection frequency weight of term $t$ and is defined as:

$$CFW(t) = \log\left(\frac{N}{N(t)}\right) \quad (5)$$

where $N$ is the number of documents in the collection and $N(t)$ is the number of documents containing term $t$. The parameters $b$ and $K$ in (3) control the effect of document length and term frequency as usual.

A number of experiments were conducted on a locally derived set of development queries to decide on a suitable stop list and to test the behaviour of running with and without stemming. These experiments clearly indicated that stemming substantially improved the average precision of the system, and that good performance was achieved using a 379 word

[2] http://www-nlpir.nist.gov/õver/zp2

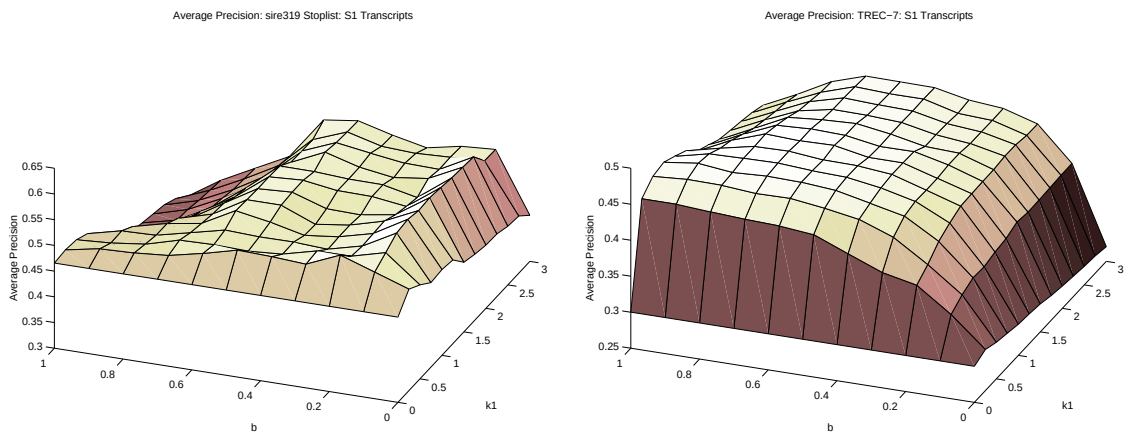



Figure 1: Plot of average precision against term weighting parameters $b$ and $K$ for TREC-7/SDR local development queries (left), and TREC-7/SDR evaluation queries (right).

stop list (based on the 319 word stop list used by the University of Glasgow at TREC-6 [6]).

Since the task of spoken document retrieval is a little different to text-based ad-hoc retrieval, we investigated the effect of varying the parameters $b$ and $K$ in the term weighting function (3). The results for the development set are shown in figure 1, along with post-evaluation results for the TREC-7 SDR queries. We note that in the development queries there is a ridge of high average precision along $K = 0.25$, which corresponds to a decrease in the significance of TF compared with CFW, which is not present in the evaluation queries. There is also a maximum around $(b,K) = (0.5,1.0)$, for both sets of queries, which (fortunately) were the parameter settings used in our TREC-7 SDR experiments reported in section 6.

The reason for the different behaviour of the two query sets is not clear. Although it may be due to the relatively small task size (around 3000 spoken documents), we also note that our local development queries had many fewer relevant documents per query compared with the evaluation queries (4.5 vs. 17). Support for the latter hypothesis is given by the fact that the parameter landscape for the known-item TREC-6/SDR queries (ie 1 relevant document per query) is most similar to the development set.

## 4 MULTIPLE TRANSCRIPTIONS

A number of researchers (eg [6, 20]) have taken advantage of the availability of multiple sets of speech recognition transcriptions and merged them to produce improved information retrieval performance. This method was successful because although speech recognizers make errors, different speech recognizers are likely to make different errors. Thus if an important query word has been missed by one recognizer, another one might recognize it correctly so that it does not get omitted from the index.

As mentioned in section 2, the ABBOT acoustic model is based on multiple recurrent networks, which are averaged together at the acoustic frame level. However, it is possible to run separate decodings based on the individual recurrent networks and to merge them together at the transcription level. Experiments were run on the TREC-6 known-item retrieval task using the 379 word stop list but no query expansion. Table 4 shows the results in terms of word error rate (WER), term error rate (TER) and the various TREC-6 IR performance measures. (See section 6 for a definition of TER.)

The table indicates that merging the RNNs at the acoustic probability level (S1) produces better WER/TER and IR performance than either of the individual networks. Despite the inevitably higher TER, merging multiple transcripts seems to produce slightly better IR results than taking their union. The detrimental effects of merging may be partially offset by term frequency weighting. In these experiments, neither merging technique produced clearly better IR performance than the single best set of transcripts (S1), except for the percentage of queries for which the answer was not found.

The results from these experiments are somewhat inconclusive: it is possible that multiple transcripts could be used to enhance retrieval performance but these benefits have yet to be demonstrated unequivocally, and must be offset against the considerable extra resources required to produce the multiple transcriptions (which is why the experiments were not re-

| Transcripts | WER | TER | Mean Rank | Mean Reciprocal | Percentage at Rank 1 | Percentage Not Found |
|---|---|---|---|---|---|---|
| R1 | – | – | 5.85 | 0.8509 | 78.7% | 0.0% |
| S1 | 38.8% | 55.4% | 11.72 | 0.7776 | 74.5% | 2.1% |
| Forward net | 43.2% | 63.3% | 14.33 | 0.6996 | 61.7% | 2.1% |
| Backward net | 41.7% | 61.4% | 17.96 | 0.7091 | 63.8% | 4.3% |
| Merged fwd+bwd | – | 135.9% | 14.51 | 0.7414 | 68.1% | 0.0% |
| Union fwd+bwd | – | 90.3% | 18.45 | 0.7477 | 68.1% | 0.0% |
| Merged S1+fwd+bwd | – | 228.5% | 14.40 | 0.7793 | 72.3% | 0.0% |
| Union S1+fwd+bwd | – | 95.9% | 19.77 | 0.7434 | 68.1% | 0.0% |

Table 1: Use of multiple transcriptions derived from ABBOT on the TREC-6 known-item retrieval task. R1 are the reference transcripts, S1 are the transcripts produced by ABBOT using frame-level merging. Forward and backward are the decodings produced by the nets in isolation. The term 'merged' implies the concatenation of two or more sets of transcripts whereas the term 'union' implies the union of sets of transcripts — multiple occurrences of the same term are discarded.

peated on TREC-7 data).

## 5 QUERY EXPANSION

If a relevant document does not contain the terms that are in the query, then that document will not be retrieved. The aim of query expansion is to reduce this query/document mismatch by expanding the query using words or phrases with a similar meaning or some other statistical relation to the set of relevant documents. Such a process may have increased importance in spoken document retrieval, since the word mismatch problem is heightened by the presence of errors in the automatic transcription of spoken documents.

An obvious danger in using relevant documents retrieved from a database of automatically transcribed spoken documents is that the query expansion may include recognition errors. This was an experience reported by the INQUERY group in the TREC-6 SDR evaluation [2]. To avoid this problem we retrieved relevant documents from another collection of newswire text. The query expansion algorithm was then applied to the top *n* documents retrieved from that collection. The resulting expanded query was then applied to the collection of spoken documents.

We used an algorithm based on the local context analysis algorithm of Xu and Croft [22]. The initial query $Q$ is applied to the secondary query expansion collection. The *nr* top ranked documents are regarded as relevant; the algorithm is not discriminative so no non-relevant documents are required. A query expansion weight, $QEW(Q,e)$ is defined as follows:

$$QEW(Q,e) = \sum_{t \in Q} CFW(t) * \log\left(\frac{\log(AF(e,t)) * CFW(e)}{\log(nr)} + \delta\right) \quad (6)$$

The potential query expansion terms $e$ are simply those terms in the relevant documents. The term $AF(e,t)$ measures the term frequency correlation of two terms $e$ and $t$ across collection of documents $d_i$:

$$AF(e,t) = \sum_{i=1}^{nr} TF(e,d_i) * TF(t,d_i). \quad (7)$$

The *nt* possible expansion terms with the largest weights are then added to the original query, weighted as $1/rank$.

In practice the values of *nr* and *nt* are maximum limits, since we threshold so that only those documents with a score greater than 0.8 times the score of the top-ranked document are considered, and only those terms with $QEW(Q,e)$ greater than an empirically-determined threshold are added.

In this work we used the June 1997–February 1998 LA Times/Washington Post portion of the TREC/SDR 1998 LM text corpus as the query expansion database. This corpus contains about 13 million words and about 22,000 documents. The parameters *nr* and *nt* are clearly dependent on the size of the query expansion collection. Experiments to investigate the dependence on these parameters were carried out on our local development queries, and the results are shown in figure 2. From this we chose parameter values $(nr,nt) = (8,10)$. Figure 3 shows the performance of query expansion using a newswire corpus versus expanding on the target recognizer transcripts.

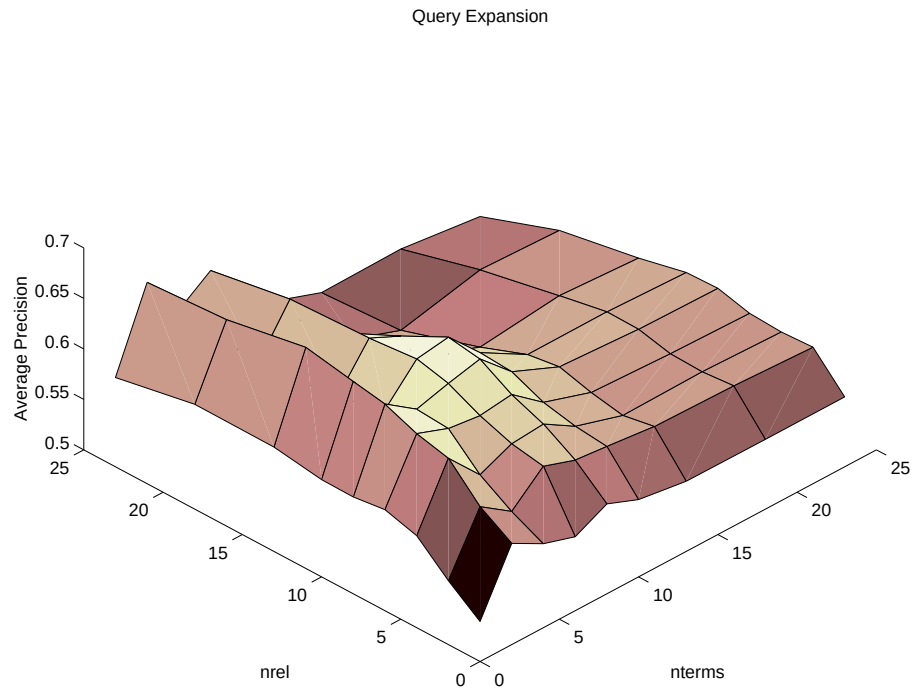



Figure 2: Effect of the query expansion parameters *nr* (maximum number of relevant documents to consider) and *nt* (maximum number of terms to add) on the average precision for our local development queries using ABBOT speech recognizer output. The Jun 1997 – Feb 1998 LA Times/Washington Post portion of the 1998 TREC-7/SDR language model corpus was used as the query expansion collection.

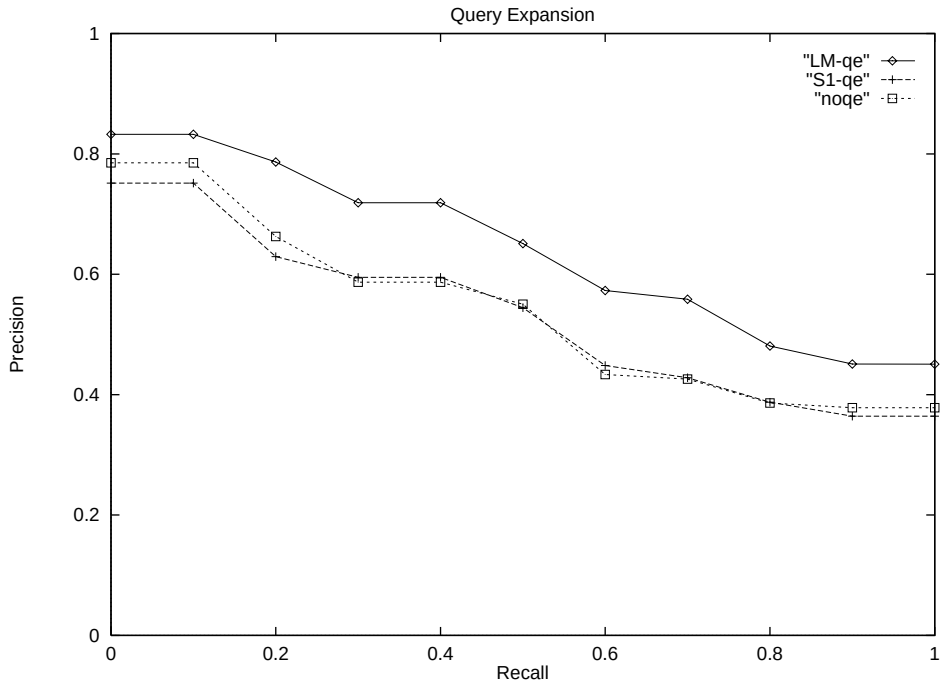



Figure 3: Effect of query expansion on retrieval of recognizer output for local development queries. Query expansion was performed on (1) LA Times/ Washington Post newswire text (LM-qe); (2) the recognizer transcripts that made up the test collection (S1-qe); and (3) no query expansion (noqe).

| Condition | WER | TER | Retrieved | Relevant | Rel. Retrieved | AveP | R-P |
|---|---|---|---|---|---|---|---|
| R1 | – | – | 17613 | 390 | 364 | 0.4886 | 0.4583 |
| S1 | 35.9% | 52.2% | 18312 | 390 | 360 | 0.4599 | 0.4485 |
| B1 | 35.2% | 49.5% | 18093 | 390 | 355 | 0.4355 | 0.4562 |
| B2 | 47.8% | 68.3% | 18671 | 390 | 354 | 0.3529 | 0.3347 |
| CR-CUHTK | 24.8% | 34.0% | 18105 | 390 | 365 | 0.4711 | 0.4469 |
| CR-DERASRU-S1 | 66.2% | 109.3% | 17844 | 390 | 334 | 0.3780 | 0.4164 |
| CR-DERASRU-S2 | 61.5% | 93.7% | 17973 | 390 | 344 | 0.4047 | 0.4016 |
| CR-DRAGON-S1 | 29.8% | 49.2% | 18252 | 390 | 361 | 0.4613 | 0.4372 |

Table 2: Summary of TREC-7 Spoken Document Retrieval track results for different recognizer conditions, evaluated in terms of word error rate (WER), term error rate (TER) defined in the text, average precision (AveP) and R-precision (R-P). **R1** refers to the reference transcripts; **S1** refers to THISL speech recognition described in the paper; **B1** and **B2** are baseline recognition runs with different levels of pruning using CMU Sphinx-III at NIST; **CR-CUHTK** refers to Cambridge University (HTK) speech recognition; **CR-DERASRU-S1** and **CR-DERASRU-S2** refers to DERA/SRU speech recognition; **CR-DRAGON-S1** refers to Dragon Systems speech recognition.

Note that expanding on the recognizer transcripts is worse than no query expansion.

# 6 Experiments

## 6.1 TREC-7 Spoken Document Retrieval

In this section we report the results carried out by the THISL group as part of the TREC-7 Spoken Document Retrieval track. This track involved a collection of 2868 news stories totalling 74 hours of broadcast audio (segmented from a total corpus of 100 hours, and not including commercials, local news, etc.). The audio track was presegmented into stories by hand, and both recognition and retrieval was performed with knowledge of this segmentation. Twenty-three queries were provided by NIST, along with pooled relevance judgments after the evaluation.

## 6.2 Speech Recognition Results

Using the system described in section 2 we were able to recognize the 74 hours of broadcast news audio date in about seven times real time on an Ultra-1/167MHz (512-1024 Mb RAM), with the computation split approximately equally between the recurrent network-based acoustic model and the LVCSR search algorithm. This implies that there was only a factor of two overhead in performing a word level transcription using 65K vocabulary and trigram language model, compared with phone recognition. However, the memory demands of LVCSR are substantial — our decoder requires a machine with 512Mb RAM — whereas the phone recognizer (essentially the recurrent networks) could run in a couple of megabytes. Running at this speed required a higher degree of pruning, resulting in a relative search error (ie, error resulting from incorrect pruning of the search space) was 10–20%.

The overall average word error rate (WER) of the THISL speech recognition system in this evaluation was 35.9%. We can also use an error metric conditioned on the text retrieval system. The *term error rate* (TER) [9] is given by the following formula:

$$TER = \frac{\sum_{t \in T} \mid R(t) - H(t) \mid}{T} \times 100\% \qquad (8)$$

where $R(t)$ and $H(t)$ represent the number of occurrences of *term* $t$ in the reference and hypothesised transcripts respectively. The set of terms $T$ is calculated after the transcripts have been stopped and stemmed but without taking account of term order. Thus TER gives a more accurate measure than WER of the erroneous terms which will be processed during IR. Additionally, calculating WER is meaningless for merged transcripts (section 4), but TER still provides some information about transcript quality. In conjunction with our submitted system, using a 379 word stop list and Porter stemming the THISL speech recognition system returned a TER of 52.2%.

## 6.3 Query Processing

Before inputting them to the text retrieval system, the queries were put through several pre-processing operations to normalize their appearance: punctuation was removed, all text was converted to lower case and possible abbreviations/acronyms were expanded to cover alternative transcription possibilities, eg “AIDS” was expanded to “aids a. i. d. s.”. No multiwords or phrases were used in the recognition or retrieval pro-

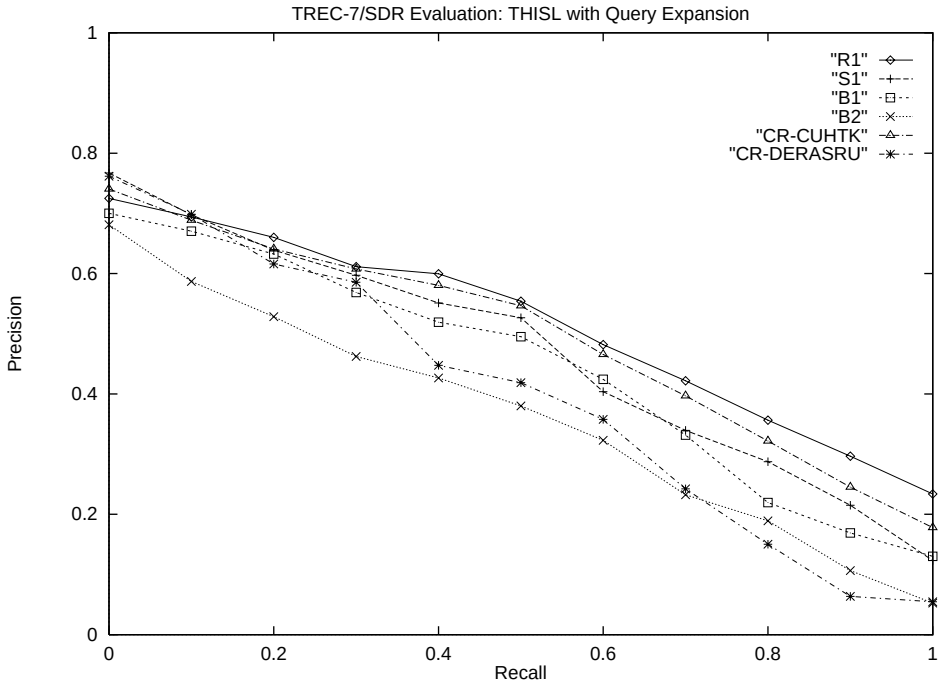



Figure 4: Recall-precision curves of the THISL system running on various transcripts submitted for TREC-7/SDR.

cess. There were three OOV query words: *Montserrat*, *Trie* & *vs.* (versus). We have previously used a word-spotting system for OOV query words [1], but in these experiments it was not used.

## 6.4 SDR RESULTS

As well as performing retrieval on the output of our own recognizer, the TREC-7/SDR evaluation permitted retrieval from the transcripts output by the recognizers of other participants. We ran on the recognition output generated by the Cambridge University HTK group, Dragon Systems and DERA/SRU, as well as on the reference transcripts and the output of two baseline recognizers run by NIST. The results were evaluated by word and term error rate and the usual TREC measures of average precision and R-precision, and are shown in table 2.

The recall-precision curves resulting from these runs are shown in figure 4. Figure 5 shows the effect of query expansion on recall and precision for the R1 and S1 conditions. Results for the other speech recognizers are not shown to avoid cluttering the graph, but the effect of query expansion follows a similar trend for those.

Figure 6 shows the relative change due to query expansion for each of the twenty-three queries. As can be seen, query expansion resulted in an improvement or no significant change in average precision for most queries. An example of a query for which the query expansion algorithm proved effective:

**60:** What information is available on the activities and motivation of intrusive photographers, i.e., the so-called paparazzi?

**Original Query:** activ avail paparazzi photograph intrus motiv call (AveP = 0.5630)

**Expansion Terms:** spencer ritz gambino merced editor trespass tabloid (AveP = 0.8589)

A query for which query expansion failed was the following:

**62:** Find reports of fatal air crashes.

**Original Query:** air fatal crash (AveP = 0.3520)

**Expansion Terms:** auto aviat safeti vehicl occup bag jour util (AveP = 0.1893)

## 7 CONCLUSIONS

We have reported on the development of a spoken document retrieval system for a broadcast news application. Beyond using the straightforward use of a text retrieval system to index the output of a speech

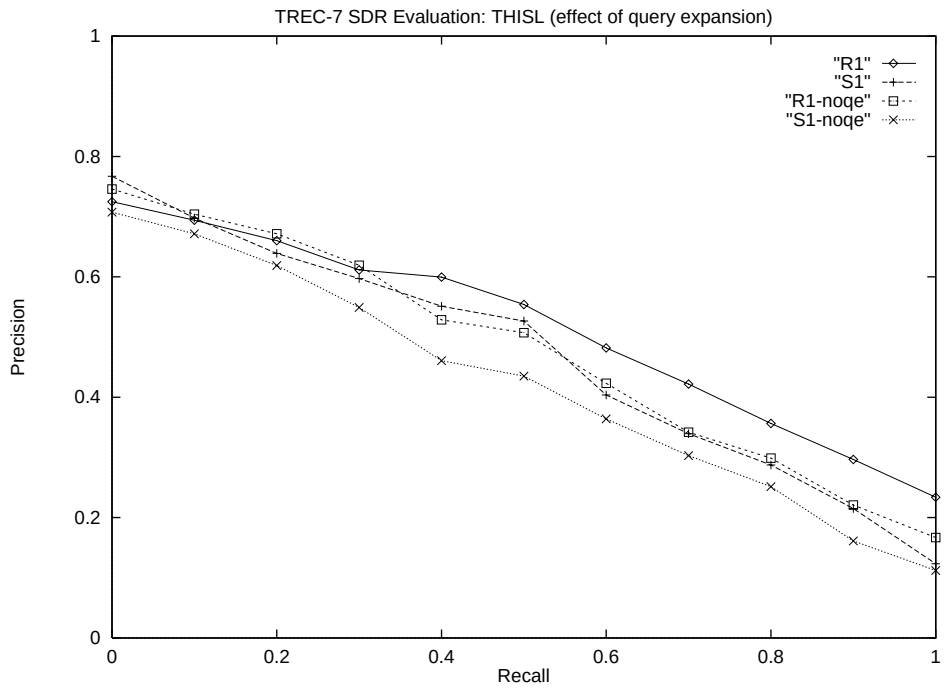



Figure 5: Effect of query expansion on recall-precision for evaluation R1 and S1 conditions (post-evaluation experiment).

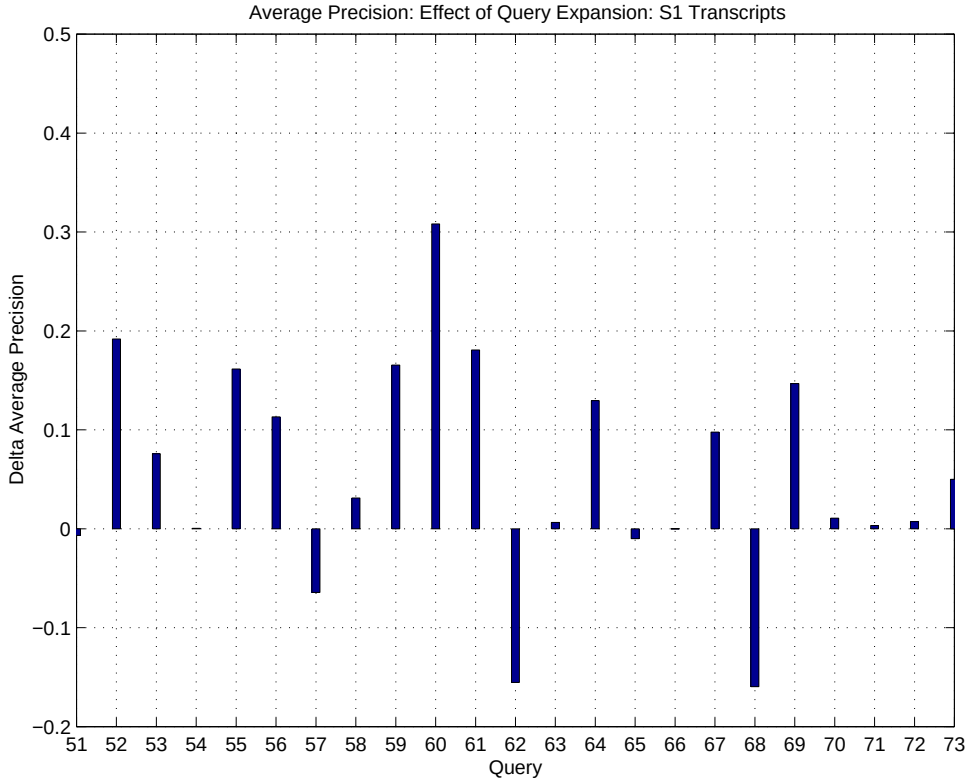



Figure 6: Query-by-query effect of query expansion in terms of change in average precision compared with no query expansion.

recognizer, we have investigated the use of multiple transcription information and query expansion. Document level merging of the possible multiple transcriptions produced by the ABBOT system was not successful in terms of improvements on the TREC-6 known item task. Query expansion, using a secondary collection of newswire data, proved to result in a consistent improvement in average precision of around 10%.

On the 100 hour TREC-7/SDR spoken document collection, our results have indicated that speech recognition systems with word error rates in the region 25–40% are adequate for this task, with only a small degradation from the reference transcripts. There is a correlation with the recognizer word error rate, but there is no clear linear relation between recognition and retrieval performance.

These experiments must be accompanied by the caveat that, in text retrieval terms, we have been working with a very small collection — less than 3000 documents — and experiments to simulate larger collections (eg by corrupting text with a similar number of insertions, deletions and substitutions that a speech recognizer would create) have indicated that difference in average precision between collections of reference transcripts and recognizer output increases with collection size [19]. Although computationally expensive, larger scale experiments in spoken document retrieval are important to test whether this simulated behaviour is accurate. The proposed TREC-8 SDR evaluation, based on 632 hours of broadcast news is a step towards this, as will be the final THISL system based on a large archive of BBC broadcast news.

## ACKNOWLEDGMENTS

This work was supported by the ESPRIT Long Term Research Projects THISL (23495) and SPRACH (20077). This work has benefited from collaboration with the partners of the THISL and SPRACH projects, in particular Tony Robinson (Cambridge University and SoftSound) and Gary Cook (Cambridge University).

## REFERENCES

[1] D. Abberley, S. Renals, and G. Cook. Retrieval of broadcast news documents with the THISL system. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pages 3781–3784, Seattle, 1998.

[2] J. Allan, J. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. Swan, and J. Xu. INQUERY does battle with TREC-6. In *Proc. Sixth Text Retrieval Conference (TREC-6)*, pages 169–206, 1998.

[3] H. Bourlard and N. Morgan. *Connectionist Speech Recognition—A Hybrid Approach*. Kluwer Academic, 1994.

[4] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Spärck-Jones, and S. J. Young. Open-vocabulary speech indexing for voice and video mail retrieval. In *Proc. ACM Multimedia 96*, pages 307–316, Boston, 1996.

[5] G. D. Cook and A. J. Robinson. The 1997 Abbot system for the transcription of broadcast news. In *Proceedings of the 1998 Broadcast News Transcription and Understanding Workshop*, 1998.

[6] F. Crestani, M. Sanderson, M. Theophylactou, and M. Lalmas. Short queries, natural language and spoken document retrieval: Experiments at Glasgow University. In *Proc. Sixth Text Retrieval Conference (TREC-6)*, pages 667–686, 1998.

[7] S. Dharanipragada and S. Roukos. A fast vocabulary independent algorithm for spotting words in speech. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pages 233–236, 1998.

[8] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Amer.*, 87:1738–1752, 1990.

[9] S. E. Johnson, P. Jourlin, G. L. Moore, K. Spärck-Jones, and P. C. Woodland. The Cambridge University Spoken Document Retrieval System. In *TREC-7 Workshop notebook*, 1998.

[10] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. Acoustics, Speech and Signal Processing*, 35:400–401, 1987.

[11] D. J. Kershaw, M. M. Hochberg, and A. J. Robinson. Context-dependent classes in a hybrid recurrent network-HMM speech recognition system. In *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1996.

[12] W. Kraaij, J. van Gent, R. Ekkelenkamp, and D. van Leeuwen. Phoneme-based spoken document retrieval. In *Proc. TWLT-14*, 1998.

[13] K. Ng and V. Zue. Phonetic recognition for spoken document retrieval. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pages 325–328, 1998.

[14] S. Renals and M. Hochberg. Start-synchronous search for large vocabulary continuous speech recognition. *IEEE Trans. Speech and Audio Processing*, in press.

[15] S. E. Robertson and K. Spärck-Jones. Simple proven approaches to text retrieval. Technical Report TR356, Cambridge University Computer Laboratory, 1997.

[16] A. J. Robinson. The application of recurrent nets to phone probability estimation. *IEEE Trans. Neural Networks*, 5:298–305, 1994.

[17] T. Robinson, M. Hochberg, and S. Renals. The use of recurrent networks in continuous speech recognition. In C. H. Lee, K. K. Paliwal, and F. K. Soong, editors, *Automatic Speech and Speaker Recognition – Advanced Topics*, chapter 10, pages 233–258. Kluwer Academic Publishers, 1996.

[18] P. Schauble. *Multimedia Information Retrieval*. Kluwer Academic Publishers, 1997.

[19] M. A. Siegler, M. J. Witbrock, S. T. Slattery, K. Seymore, R. E. Jones, and A. G. Hauptmann. Experiments in spoken document retrieval at CMU. In *Proc. Sixth Text Retrieval Conference (TREC–6)*, pages 291–302, 1998.

[20] A. Singal, J. Choi, D. Hindle, and F. Pereira. AT&T at TREC-6: SDR track. In *Proc. Sixth Text Retrieval Conference (TREC-6)*, pages 227–232, 1998.

[21] A. F. Smeaton, M. Morony, G. Quinn, and R. Scaife. Taiscéalaí: Information retrieval from an archive of spoken radio news. In *Proc. Second European Digital Libraries Conference*, 1998.

[22] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proc. ACM SIGIR*, 1996.

# Phoneme based Spoken Document Retrieval

Wessel Kraaij, Joop van Gent and
Rudie Ekkelenkamp
TNO-TPD
Stieltjesweg 1, 2600 AD Delft
The Netherlands
{kraaij,gent,ekkelen}@tpd.tno.nl

David van Leeuwen

TNO-HFRI
Kampweg 5, 2769 DE Soesterberg
The Netherlands
vanleeuwen@tm.tno.nl

## ABSTRACT

Since speech recognition technology has become more and more mature, retrieval of spoken documents has become a feasible task. We report about two cases, which aim at scalable and effective retrieval of broadcast recordings. The approach is based on a hybrid architecture, which combines the speed of off-line phoneme indexing and precision of wordspotting while maintaining a scalable architecture, which allows for frequent updates of the database where out-of-vocabulary (OOV) words are abundant. A pilot experiment has been done on a small database of recordings of a Dutch talkshow. A more extensive evaluation took place in the framework of the Spoken Document Retrieval track of TREC7 on English broadcast news.

**Keywords:** Spoken Document Retrieval, Speech Recognition, Radio broadcast databases

## 1 INTRODUCTION

This document describes ongoing experiments at TNO in the area of spoken document retrieval. This relatively new field poses new challenges to classical information retrieval because from an IR perspective the data to be indexed is highly corrupted. SDR is also quite demanding for Speech Recognizers. The source material has a high variety of speakers, low signal/noise ratio, there are many proper names (which are very important for retrieval) and the collections grow rapidly. So a SDR system should be very robust.

### 1.1 EARLY EXPERIMENTS WITH SPEECH RETRIEVAL

The idea for the approach described in this paper was inspired by the success of an earlier project called Talking Heads, which was carried out in 1995 by the independent Dutch research organisation TNO-TPD and the telecom company KPN Research.

The idea in Talking Heads was to combine a proprietary full text retrieval system called MOOI with speech recognition. MOOI was developed earlier by TNO and allowed users to retrieve information from a textual database in two steps, in the following way.

A textual query entered via the keyboard was matched with an index consisting of noun phrases, using proprietary fuzzy matching technology based on trigrams, called ISM ([12])The noun phrases were derived automatically from the database using simple forms of syntactic analysis. This first retrieval step would thus yield a list of noun phrases ranked according to “relevance” to the query, whereby relevance was defined in terms of the fuzzy matching algorithm. In a second step users could select one of the noun phrases from the retrieved list, and thus retrieve a list of all documents that contained the selected phrase.

This method worked well in the sense that the intermediate step in which the phrases were shown to the user allowed for an improvement in precision over one step approaches and a high flexibility on the query side, because the system could come back with good results in cases where the query contained typo’s, misspellings or morphological/syntactic variants of relevant

phrases in the index. The system also gave good results when applied to corrupted data such as OCR text. TNO had also experimented with a "phonetic" variation of the fuzzy matching algorithm: here *triphones* (trigrammes constructed from phonetic representations of the phrases and the queries) were used.

Talking Heads used both the two step retrieval strategy and the fuzzy matching algorithm in its "phonetic mode" to allow for retrieval of textual information using spoken queries. The front-end of the Talking Heads system consisted of the Hidden Markov Toolkit (HTK) system configured as a phone recognizer. It was tuned such that the phonetic representation of the speech data could be used for trigramme matching. Spoken queries could thus be transformed to so-called graphone strings. Graphones were phonetic representations of character strings. The graphone alphabet was developed earlier by TNO and optimised for soundex-like retrieval on Dutch texts, and later on tuned to the Talking Heads demonstrator. The (back-end) textual database was indexed with phrases, with the use of the MOOI system. These phrases were converted to a graphone representation using a rule based transfer module. The output of the HTK was matched with the graphone representation of the phrases using a combination of ISM and the Levenstein edit distance metric. For the purpose of the project a special speech model was developed for circumstances of casual use of speaker independent microphone speech.

The result was a system that translated spoken queries to graphone strings and matched them with a graphone representation of the noun phrase index. The system would thus yield a list of noun phrases on the screen, and the idea was that by using just a touch screen and a microphone the user could have full interaction with the system. In the project a mouse was used instead of a touch screen. The envisaged applications were either robust electronic information services in public buildings such as department stores, or professional electronic assistants in situations where users would need hands free information access, such as assembly or maintenance, like in a garage.

Despite of the relatively poor speech model that was used the results were quite promising. In most of the retrieval sessions arbitrary users could get access to relevant information within 6 interactions, i.e. 1 or 2 retries for query entry, and 1 or 2 retries for phrase selection. In the other cases the system just could not interpret the query.

The main problem within Talking Heads was the relatively poor speech model, that was based on insufficient training, particularly with respect to its speaker-independence. The project time and budget did not allow for a more profound training though. The results lead the research team to the conclusion that it would be worth trying a single speaker approach which would involve significantly less training. This approach would be uninteresting for the desired applications, but useful for retrieval on a database with spoken information uttered by one or just a couple of speakers, like for instance radio news items. A concurrent advantage would perhaps be the better (studio) quality of the speech signal. Within talking Heads this idea of using *triphone matching* for retrieval on speaker-dependent speech data could not be worked out anymore, but it made its way into the follow-up projects DAS+ and DRUID which provided the framework for the experiments which will be presented in this paper. DAS+ is a project for the Dutch broadcasting company TROS. Together with systems house VDA, TNO has built a prototype retrieval system for their Radio broadcast archive. DRUID is a multimedia research project funded by the Dutch government and industry via the Telematics Institute. One of the topics in DRUID is speech recognition. In order to test the DAS+ prototype on a larger scale, we participated in the SDR track of TREC7. The paper is organised as follows, this section continues with a description of our approach. Section 2 describes the pilot experiments in the DAS+ framework and section 3 presents the TREC7 SDR experiments. Section 4 gives conclusions and ideas for further work.

## 1.2 SPOKEN DOCUMENT RETRIEVAL (SDR)

With the recent rapid improvements in Speech Recognition technology, retrieval of Spoken Documents has become feasible. The SDR task is similar to the Text Retrieval task, but the document collection consists of Audio recordings containing spoken material. A typical SDR system applies large vocabulary continuous speech recognition (LVCSR) as a preprocessing step in order to use standard text retrieval techniques. A second option is to convert the audio signal into a phoneme sequence, which has certain advantages like a flexible vocabulary. The latter approach was also chosen used for the experiments, which are reported in this paper. Both conversion approaches have one common feature i.e. recognition errors which are quite frequent even with broadcast quality speech. The error rate for phoneme transcripts is higher however, because the models take less context into account.

The spoken material in SDR applications ranges from Radio News bulletins, radio talkshows to sound clippings from video broadcasts each posing their particular difficulty to SR systems. Unlike traditional SR applications where precision is of primary importance like *command & control* or *dictation*, recognition errors do not necessarily invalidate spoken document retrieval, because the goal of SDR is not to retrieve transcripts, but to retrieve and play audio segments that are relevant to the users query. If we manage to locate relevant fragments, we can just play the original recording instead of displaying the corrupt transcript. There is an analogy to the retrieval of OCR text. The latter case is likely to be less hampered by corruption because of the quite high redundancy in written text. In Radio news bulletins the redundancy is probably less high, so the need for error-tolerant search techniques is more urgent.

Depending on the application type, Recall might be more or less important, in most cases Precision is important. A final requirement for SDR systems is scalability, off-line recognition and indexing time should be at least one order of magnitude faster than real time. Retrieval response time should be low, which necessitates an architecture based on indexing instead of linear search.

The majority of SDR applications consist of a combination of an LVCSR system with a classical IR system. This means that the audio material is simply converted to a textual transcript that is input for the IR system. Successful prototype systems exist like Informedia [10] with spoken input. However, although the vocabulary of these systems is quite large, the vocabulary is fixed. Secondly, the majority of these recognizers have been trained for American English only. One approach to building an SDR application for Dutch could be to train an LVCSR system for Dutch. This is quite an effort, however, because large annotated corpora are required to train the acoustical models. We think, however, that we would encounter two more fundamental problems with a LVCSR-only based SDR system for Dutch. Firstly there's the out-of-vocabulary (OOV) problem which is quite prominent in the news domain (proper names). Secondly, the morphology of Dutch is more complex than English, requiring a much larger vocabulary for the same coverage. For the majority of Germanic languages (English being an exception) compounds are written as single orthographic units, which means that in order to be recognized, they have to be included in the lexicon of the recognizer. As compounding is a highly productive process, the OOV problem is more severe.

## 1.3 PHONEME BASED APPROACH

For DAS+ we have chosen to experiment with phone based retrieval. This choice was mainly motivated by the following arguments

- No language model required
- Off-line recognition time is much faster than LVCSR (simpler search algorithm)
- Less sensitive to the OOV problem
- Reuse of robust indexing strategies for OCR text (triphones instead of trigrams)

Pilot experiments showed that the retrieval results with triphone matching produced results with a

rather poor precision. Because experiments with a wordspotter configuration using Abbot had shown quite impressive precision, we decided to add a word-spotting step as a refinement step on the result set of the triphone search. The 2-stage search strategy has the following advantages:

- Retrieval based on triphone matching is fast, but not very precise because of the high phone error rate.
- A word spotter based on on-line phone lattice spotting is much more precise, but also slower due to the linear search process.

A more detailed description of the system will be given in section 2.

## 1.4 EVALUATION METHODS

In order to assess the retrieval quality of SR components and SDR systems as a whole, different methodologies can be applied. All methods presuppose a 'test corpus'. A classical IR test corpus consists of a collection of documents, a collection of queries and a set of relevance judgements. For a good evaluation one would like to test on several test collections, and preferably on test collections of considerable size. Such a test would produce results on 'average precision' and 'precision at cut-off levels'. Unfortunately these test collections do not yet exist in the SDR domain (an exception is the test corpus of the SDR track of TREC7 which has been constructed this year). A simpler poor mans solution to evaluation (which was used at TREC6), is to perform *Known Item Retrieval*. This procedure works as follows. First a set of unique documents is selected (this is in fact a non-trivial task). Queries are constructed from these documents. A perfect SDR system will return the document, which was the source for the query at first rank in the result list. Three measures can be derived from this evaluation method: (a) mean reciprocal rank (b) percentage of queries which retrieved the the known item in first position (c) cumulative percentage of queries that retrieve the known item by rank. A disadvantage of the method is that it does not say a lot about Precision and Recall.

When comparing results of different groups it's quite important to compare the characteristics of the test collection. Does it contain read speech or spontaneous speech, one speaker or multiple speakers? What methodology or rationale has been used to segment the audio files into separate 'audio documents'? Segmentation could for example be based on fixed time frames, on speaker pauses, on lexical or visual clues determining story boundaries in news shows (CNN). The segmentation methods have a profound effect on the characteristics of a collection, and of course, also on the usability of a system.

## 1.5 RELATED WORK

There are two European groups that have been working on phoneme based SDR for several years and which inspired our work. In 1995 Wechsler and Schauble [14] started experiments at ETH-Zürich with a HTK based phone recognizer. They performed tests on a German corpus consisting of 4 hours of broadcast news, segmented into overlapping windows of 20 seconds. Queries are titles from news stories taken from the similar period. Experiments included a comparison between bi-, tri- and tetra grams of phones. Trigrams performed 225% better (average precision) than bigrams. Tetragrams were slightly worse. They also experimented with a probabilistic method to cope with recognition errors. Phone transcripts are scanned for sequences similar to the query via a fuzzy matching procedure, bounded by a maximal edit distance. The probability that these near matches are in fact correct hits is estimated on the basis of, among others, the confusion matrix of the recognizer. The method yielded an improvement of 32 % with respect to the trigram baseline. A variant where the fuzzy matching is applied on triphones has been tested on the TREC6 SDR corpus[9]. Results are impressive, the mean reciprocal rank is improved with 63% (0.20→0.43), given a phone error rate of 55%. The method can be simplified as: select candidate hits (slots) by dynamic generation of variants of partly matching triphones within a certain maximum edit distance. Subsequently estimate

the probability of each hit, rank them and select the top 100 hits, assuming that the rest are false alarms. The approach is essentially a word-spotting architecture, based on linear search. The dynamic programming and probability re-estimation technique defines a hypothesis space which is analogous to the phone lattice structures which can be produced with Abbot. However the ETH method is probably significantly slower than the phone lattice based word spotting approach in our experiments. The phone lattice files (produced off-line) are simply searched by a finite state automaton. No dynamic programming is needed to find candidate hits.

At Cambridge University, James [4] evaluated a number of SDR configurations on a small corpus (2h27') of broadcast news. The test collection further consisted of 40 queries and relevance judgements. His best results are produced with a hybrid system: a word recognizer in combination with a (phone lattice based) wordspotter. The approach has been extended and tested in the Video Mail Retrieval project [5][15] resulting in an average precision of 85% of the baseline (retrieval on manual transcripts).

An operational Broadcast news archive has been built at Dublin City University[8]. The SR system is based on the HTK, which is trained on triphones. This significantly reduced the error rate. The computational complexity of the training process was reduced by a smart conflation of similar phones. Results in TREC6 were good.

The SDR track of TREC6 [11] showed a diverse spectrum of hybrid and fault tolerant approaches based on exploiting N-best SR output or generating confusion variants on the fly. CMU's run with an N-best recognizer was promising, Clarit's query expansion with confusion variants was not so successful, maybe due to lack of adequate term re-weighting, ETH's system performed disappointing due to high phone error rate. IBM presented an approach, which has inspired us: the LVCSR based architecture is complemented by a word spotter fall back strategy to find OOV words. Because the spotter is relatively slow, it is only started on a pre-selection of the collection as determined by an n-gram retrieval run on a phone representation. The University of Sheffield did tests with the Abbot system, configured as word recognizer. OOV words were spotted, with limited effect because the Abbot pronunciation dictionary was missing only one word.

Conclusion: Phoneme based approaches are feasible, though extra care for term weighting schemes is required. In an environment with only a few OOV words, word based recognizers are much more effective. In operational systems, hybrid approaches are probably the best option.

## 2 SYSTEM ARCHITECTURE

The TNO spoken document retrieval system is based on the ABBOT Large Vocabulary Continuous Speech Recognition (LVCSR) system [7]developed by Cambridge University, Sheffield University and SoftSound. Abbot is complemented by a set of Indexing and Fuzzy matching modules, developed at TNO-TPD.

The following figure shows the architecture of the TNO system as used in the TREC7 experiments. For the pilot experiment, which is described in section 3 a more simplistic term weighting strategy was used.

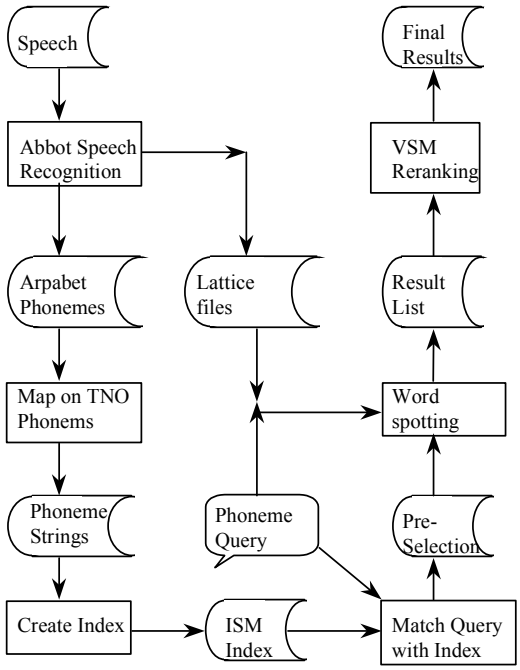



Figure 1: System architecture

In the left branch, audio files are converted to phone strings and phone lattice files. The phoneme transcripts are segmented and indexed by triphone sequences. There are two posibble retrieval strategies:

- A query is mapped to a phoneme representation based on the CMU dictionary. This phoneme representation is matched on the triphone index using the fuzzy matching tool ISM.
- The phoneme representation of the query is input for a word spotter, that searches the phone lattice representation of the collection.

These techniques will be discussed in some more detail in the following sections.

## 2.1 FUZZY MATCHING ON PHONEME TRANSCRIPTS

Abbot is configured as a phone recognizer (instead of a continuous word recognizer), in order to generate phone [1] transcripts of the speech documents. These are in turn converted to phoneme strings by segmenting the phones on pause symbols and mapping the phone symbols onto the characters a-z and A-Z. The phoneme strings are input for a fuzzy index based on phoneme trigrams (ISM index). For retrieval a fuzzy match is carried out between a phoneme representation of the query and the phoneme trigram index resulting in the top N documents which contain phrases similar to the query. The phonetic representation of the topic is determined by using the Carnegie Mellon Pronouncing Dictionary [CMU, 1995]. OOV words have been ignored.

[1] A *phone* is an acoustical realisation of a sound. A *phoneme* is a conceptual representation of a sound. There can be several phone realizations for a single phoneme in a language, for instance the `t' in `top' is aspirated, while the one in `stop' is not.

## 2.2 WORD SPOTTING WITH *TF.IDF* TERM WEIGHTING

**Off-line processing**
First, Abbot generates phone lattices, by reducing the acoustical input to posterior probability vectors of all phones in the phone set, of each 16 ms time frame. These lattices can be used to do both phone recognition (see 2.1) and on-line word spotting.

**On-line processing**
For word spotting, a phonemic representation of all query words is made. The words are mixed with single phones in a finite state grammar, and the query terms are spotted in the phone lattices using the finite state grammar decoder of Abbot. This is effectively a linear search. Finally all documents with query term hits will be scored and ranked via a standard *tf.idf* term weighting strategy.

# 3 PILOT EXPERIMENT

A first series of experiments with phoneme based SDR was performed within the DAS+ project. The test collection consisted of 1380 seconds of "Kamerbreed", a talkshow about news topics. This collection is quite different from the typical SDR test collections, which are typically dominated by, read speech and well-directed interviews. Our test collection was characterized by spontaneous speech, with interruptions and discussions. Part of the material was transcribed manually and segmented into utterances, separating different speakers and not splitting semantical units.

## 3.1 TRAINING THE PHONE RECOGNIZER

For acoustical training of the phone recognizer, approximately 4 hoursof speech data was available. This material was collected from 24 one-hour broadcasts of the radio programme “Kamerbreed.” In order to achieve high recognition accuracy, speaker-dependent models were made. Therefore, only models for two speakers were used, these speakers have the

function of interviewer in the radio show. At a later stage we plan to make speaker-independent models. The training material was transcribed manually at the word level. Because of the spontaneous character of the material, several notation conventions were introduced. Examples are:

1. special symbols for "breath" (an audible breath) and "eh" (a spoken filler word).
2. Hesitations, repetitions, and talking errors were spelled out phonetically in case they could not be transcribed as words.

For all words transcribed in the training material, a pronunciation in terms of phones was looked up in the CELEX dictionary. For words not in this dictionary, pronunciations were produced manually. For the latter class of pronunciations, suggestions were made available automatically by a compound splitter, because for Dutch, OOV words are often compounds.

For training of the acoustical parameters of the recognition system, an earlier version of the system was used as a bootstrap model. Training was performed after forced alignment of the phone transcription with the acoustical data. Acoustical training included the updating of the weights in the recurrent neural network, which forms the heart of the Abbot recognition system. Also the phone Markov models, which model phone durations, were re-estimated. The performance of the speaker-dependent models was measured from the development test set, giving phone error rates of 34% and 39% for the two speakers. Approximately half of the errors are deletions. Subjective comparisons of the acoustic material and the reference phone transcription suggest that the deleted phones are indeed never pronounced.

## 3.2 EXPERIMENT

First, Abbot was used to produce both a phone sequence representation and a phone lattice representation of each segment. The test collection was quite small: 134 segments. 74 queries were constructed from these segments in order to do a *known item retrieval* evaluation. The queries were converted to phoneme sequences using a grapheme to phoneme converter (G2P) The G2P uses a standard phoneme dictionary (CELEX) a large list of proper names and a simple algorithm to decompose compounds. We did experiments with triphone fuzzy matching on the phone transcripts (ISM [12]), and with Abbot word-spotting using the phone lattice files. In the latter case the phoneme representation of a query term is used to construct a finite state automaton to search through the lattice. The word spotter runs did not use any form of term weighting, straight *tf* was used as document score, queries were 2.18 words long on average.

The following experiments have been performed: (a) baseline: manually transcribed data (b) triphone matching (ISM) on phonemes (c) word-spotting on phone lattices. We tested a set of minor variations of the wordspotter configuration (i) one word-spot run per query term (ii) one word-spot run per query (iii) addition of phrases. The rationale for these experiments is that the false alarm rate is lower for longer words (phrases) and that a larger vocabulary for the wordspotter increases accuracy, because overlap between spotted words is impossible.

## 3.3 RESULTS

### 3.3.1 Wordspotter versus triphone matching

Different word spotter configurations were tested. We counted the number of queries that did not retrieve the known item at first rank. As a measure for discriminatory power, we also computed the mean size of the first rank.

| test on 74 queries | # failures | # rank 1 |
|---|---|---|
| Ws1: 1 word/run | 5 | 3.27 |
| Ws2: stopword removal | 4 | 3.57 |
| Ws3: phrases in single run | 6 | 3.28 |
| Ws4: concatenated phrases | 7 | 3.27 |
| Ws5: 1 run/query | 5 | 3.18 |
| Triphones (ISM) | 25 | 1.39 |

**Table 1: Results of wordspotting and triphone**

**matching**

**Table 1** shows the tremendous difference between wordspotting and triphone matching. On a queryset of 74 queries ISM fails about 1 out of 3 times to retrieve the known item at first rank. The Wordspotte, however, fails only 4 times. The word spotting runs can probably be improved further by a proper weighting scheme, which uses the confidence information, which in principle is available from the spotter. Small modifications that improved results were stopword removal (test2) and starting the wordspotter with the full set of query terms (test5). The experiments with phrases did not yield uniform improvements. A more sophisticated weighting scheme could probably help.

### 3.3.2 ISM as pre-selection for the wordspotter

We also experimented with a hybrid architecture where triphone matching was used as a first step to reduce the amount of data to be searched by the word spotter. The word spotter in this test is the word spotter with stopword removal.
We experimented with several methods to make the preselection:

1. Take all documents from the first stage
2. Take the N best hits
3. Take the hits with a score higher than δ
4. Take the N best ranks

| INPUT | # failures | # rank 1 | # refs searched | (reduction factor) |
|---|---|---|---|---|
| N=all hits (1) | 6 | 2.18 | 27 | 5.0 |
| N= 20 hits (2) | 8 | 2.16 | 17.5 | 7.7 |
| N= 10 hits (2) | 11 | 1.82 | 9.6 | 13.9 |
| N= 1 hit (2) | 33 | 1.0 | 1.0 | 134 |
| δ=0.20 (3) | 21 | 1.79 | 3.46 | 38.7 |
| δ=0.40 (3) | 53 | 1.28 | 0.54 | 247.9 |
| δ= 0.50 (3) | 62 | 1.41 | 0.32 | 413 |
| Best 3 ranks (4) | 8 | 1.6 | 9.7 | 13.7 |
| Best 2 ranks (4) | 14 | 1.4 | 4.4 | 30.2 |
| Best rank (4) | 27 | 1.1 | 1.56 | 85.5 |

**Table 2: ISM + wordspotter hybrid system**

The third column in **Table 2** gives the average number of segments per query which is presented to the wordspotter, the next column gives the corresponding reduction factor. The table shows that the 2-stage architecture has a perfect means to trade time for the percentage of known items retrieved by the wordspotter. Figure 2 shows that the %found@1st_rank measure increases sharply in the initial segment of the plot. A threshold of e.g. N=7 seems quite reasonable for this (small) collection, limiting the search time for the spotter with a factor 10.

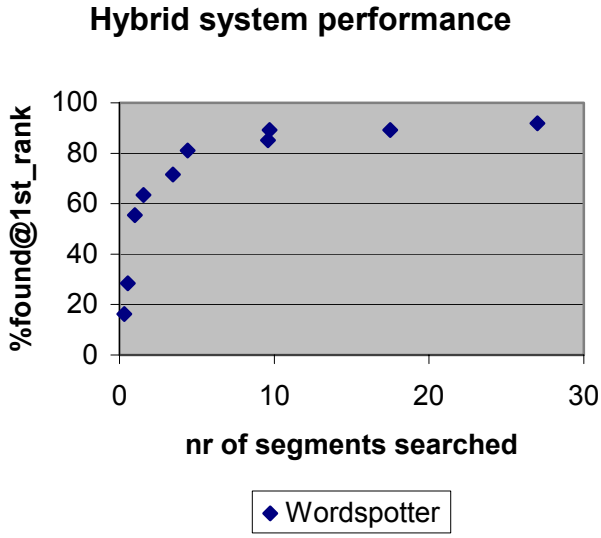



Figure 2: Trading speed for Recall

## 3.4 CONCLUSIONS

The various wordspot architecture variants do not differ much when we look at the %found@1st_rank measure. A more precise comparison can only be made when we include the confidence values of the wordspotter into the ranking algorithm. There is already evidence, however, that removing stopwords is quite useful, and that treating proper names (e.g 'Wim Kok' the Dutch prime minister) as a single acoustic unit probably improves precision. When we compare the wordspotter results with fuzzy matching on triphones, the difference is much more marked. The experiments with a hybrid approach show that ISM can be used as a coarse first step to select a list of candidate hits which in turn is

searched with the high quality word spotter. The hybrid system is much more precise than the 1st step alone, and faster than wordspotting on the full database. If a user wants more Recall, he can simply start the 2nd reranking step on a bigger fraction of the results of the 1st step. The wordspotter works in approximately 40 times real-time on a Sun ultrasparc 300Mhz.

# 4 TREC 7 SPOKEN DOCUMENT RETRIEVAL TRACK

This section reports about experiments in the SDR track of TREC7, carried out at TNO-TPD and TNO-HFRI. The corpus for this track consisted of 100 hours of American Broadcast news for training and 100 hours for evaluation. We only used the latter part, which consists of 2866 documents with an average of 268 words. The segmentation of the corpus has been done manually, in order to produce separate stories. There were 23 queries.

Again we used the Abbot SR system[7], configured as a phone recogniser. Because we lacked time, we could not train Abbot for American English. Tony Robinson from the University of Cambridge kindly provided the acoustical models (the weights of the recurrent neural network) needed to carry out phone recognition and word spotting. For retrieval we experimented with several approaches: (a) Fuzzy matching on a phoneme representation of the database and (b) Phone lattice based word spotting, with a quite standard *tfidf* term weighting strategy.

## 4.1 TRAINING

The phone models were trained at Cambridge University on the 100 hours SDR training set (the 1996 Broadcast News speech corpus , CSR-V: Hub 4). Training was done on two different subsets:

(1) All training data stripped from commercials (about 70 hours)
(2) Only studio quality material (for F0 and F1 conditions) each producing their own model.

Training was performed forwards and backwards in time, effectively resulting in 4 models. At recognition time the log probabilities of the 4 models were averaged.

We did an initial evaluation of the quality of the phone recognizer on 2.2 hours of the TREC7 Broadcast News testset and found a phone error rate of 44% (21% substitutions, 21% deletions and 2% insertion, phone set of 55 phones)

## 4.2 OFFICIAL RUNS

We used a single strategy for the R1, B1 and B2 tasks. A vector space index was built on the transcripts and the topics were matched with this index. The weighting scheme used is okapi9, as used in the PRISE engine from NIST[6].okapi9 defines the tf component as

$$\frac{tf}{tf + \log(1 + dl/avdl)}$$

where *dl* is the document length and *avdl* represents the average document length. This resulted in the following average precision values for the tasks R1, B1 and B2.

| Run Type | AVP | %change |
|---|---|---|
| **R1**: Reference Retrieval using human-generated "perfect" transcripts | 0.3970 | 0 |
| **B1**: Baseline Retrieval using medium error (35%WER) recognizer transcripts | 0.3533 | -11 |
| **B2**: Baseline Retrieval using high error (50%WER) recognizer transcripts | 0.2833 | -29 |
| **S1**: Full SDR based on wordspotting | 0.0436 | -89 |
| **S1_fixed:** Unofficial bugfix run | 0.1219 | -69 |

Table 3: Results of the SDR TREC7 runs

For the S1 run we submitted a run based on the method described in 2.2. The only conceptual difference with the DAS+ pilot set-up on the IR

side was that we intended to do a more sophisticated term-weighting strategy. This strategy was based on the retrieval model developed within TwentyOne at University of Twente[2][3], which was used to modify the ranking based on a simple count of the number of spotted words

After receiving the relevance judgements some unofficial runs were done for the S1 task. It turned out that there were some major errors in the system. Some of these errors have been solved now (cf. 4.3) and the best run for the S1 task using the word spotting approach has an average precision of 0.1219 (runtag: S1_fixed).

## 4.3 DISCUSSION

The baseline runs show that the average precision decreases steadily with increased word error rate. But at a 50%word error rate, the performance is still quite reasonable. The S1 results were quite disappointing. We have identified a series of possible causes. First of all, due to lack of time no phoneme or phone lattice transcript of the training set was available for the S1 task. The R1 run was used as a substitute relevance judgements file. It turned out to be very hard to tune the system with these judgements. Post-hoc analysis of our S1 run revealed some problems:

- **Errors in term weighting:** termweighting was effectively inactive.
- **Document length of word spotted document**. The ranking of the documents was suboptimal because we didn't know the document length of the spoken documents. In the official S1 run we used the number of spotted words as document length. This turned out to be a bad measure for the document length. In our unofficial run we used the length of the phonetic representation of the document. This dramatically improved the performance.
- **False alarms for short words.** Another big problem for word spotting was the high false alarm rate for short words. For example: the word "gun" has been spotted 14.000 times while in the transcriptions it only occurs about 100 times. This degraded the performance dramatically. In the future the confidence value of the spotted word should be taken into account to be able to tune for small and large words. Another possibility is to test the reweighting strategies as proposed by ETH[9], or to start the wordspotter with an extra dictionary of say 1k frequent words. This will most probably decrease the false alarm rate of short words. Figure 3 gives a comparison between the number of words found by the wordspotter (grouped by word length) and the true number of occurrences as found in the manual transcripts. It can be clearly seen that especially short words generate a lot of false alarm. The alpha parameter is used by the wordspotter to limit the number of false alarms. Note that stopwords were removed from the query term set, before making this plot!
- **OOV query terms**. There is an important difference in the consequences of OOVs in the conventional word recognition based retrieval and the word spotting based retrieval. For word spotting, only phone representations of OOV query words need to be generated on-line after the query has been made. A fast word spotting search can then be performed. But unfortunately no text to phoneme converter was available for English, so the CMU dictionary (0.4 version) has been used. This means that we could not fully exploit the potential advantage of our phoneme-based approach, which in principle is not vocabulary dependent. Quite a few topic words were not found in the CMU pronunciation dictionary among which crucial terms like: paparazzi, Montserrat and US. Since the spotter skipped these words, the results for the corresponding topics were ruined.

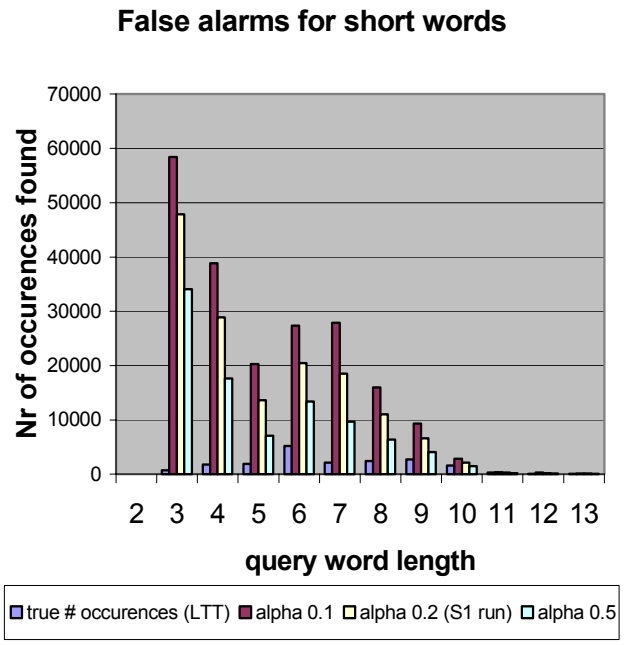



Figure 3: Rough estimate of false alarm rate

## 5 CONCLUSIONS

We have succeeded in building laboratory versions of an application for Spoken Document Retrieval based on phone recognition for Dutch and English. The pilot experiments with Dutch have shown that the 2-stage architecture is quite effective. A coarse 1st stage triphone search proved an effective means to limit the search space for the linearly operating but high quality word spotter. The initial results in the TREC7 SDR revealed a number of errors, some of which have already been corrected, resulting in big improvements. As such, the TREC7 evaluation testbed will be used to test and validate improved version of our applications. The unofficial corrected SDR runs have already shown that phone based retrieval is a feasible and scaleable approach. However for a real test of the architecture we need to do tests with a rule based grapheme to phoneme converter.

## 6 ACKNOWLEDGEMENTS

We would like to thank Tony Robinson from the University of Cambridge for providing us with the acoustical models for American English. Furthermore we would like to thank the DAS+ colleagues and especially Daan Otten and Jurgen den Hartog of TNO-TPD for their help to set up the DAS+ evaluation.

## REFERENCES

[1] Carnegie Mellon Pronouncing Dictionary (cmudict.0.4, 1995). http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[2] Hiemstra, D , A Linguistically Motivated Probabilistic Model of Information Retrieval, *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL2)*,Crete,1998.

[3] Hiemstra, D. and W. Kraaij, TREC working notes: Twenty-One in ad-hoc and CLIR, *TREC 7 working notes*, 1998.

[4] David James, The application of Classical Information Retrieval Techniques to Spoken Documents, *Thesis*, University of Cambridge, 1995.

[5] Jones, Gareth J.F., J.T.Foote, K. Sparck-Jones and S. Young, Retrieving Spoken Documents by Combining Multiple Index Sources, *Proceedings of ACM-SIGIR 1996,* Zürich.

[6] The ZPRISE 1.0 Home page: www-nlpir.nist.gov/~over/zp2.

[7] Tony Robinson, Mike Hochberg and Steve Renals ,The use of recurrent neural networks in continuous speech recognition *http://*svrwww.*eng.cam.ac.uk/~ajr/rnn4csr94/rnn4csr94.html*

[8] Smeaton, A. F., M. Morony, G. Quinn and R. Scaife, Taiscéalái: Information Retrieval from an Archive of Spoken Radio News, *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL2),* Crete, 1998.

[9] Wechsler. M., E. Munteanu and P. Schäuble, New Techniques for Open-Vocabulary Spoken Document Retrieval, Proceedings of ACM-SIGIR 1998, Melbourne.

[10] Hauptmann, Alexander G., en Witbrock, Michael J., Informedia News-On-Demand: Using Speech Recognition to Create a Digital Video Library, *CMU paper* http://informedia.cs.cmu.edu/pubs/aaaiinfo-haupt.pdf

[11] Garofolo, J., E.Voorhees, V. Stanford, TREC-6 1997 Spoken Document Retrieval Track Overview and Results, Harman, Donna (ed*.), Proceedings of the Sixth Text REtrieval Conference (TREC6*), NIST special publication 500-240, 1998.

[12] Heer, T. de, Quasi comprehension on natural language simulated by means of Information Traces, *Information Processing & Management,* 15,89-98, 1979.

[13] Jones, Gareth J.F. en James, David A., A Critical Review of State-of-the-Art Technologies for Cross-Language Speech Retrieval, *AAAI Spring symposium Cross Language Retrieval*, 1997, Stanford.

[14] Schäuble, Peter. Multimedia Information Retrieval*, Kluwer Academic Publishers*, Boston, 1997.

[15] Sparck Jones, Jones, Foote en Young, Experiments in Spoken Document retrieval*, Inofrmation Processing & Managament* **32** 399-419, 1996.

# THE USE OF MMR, DIVERSITY-BASED RERANKING IN DOCUMENT RERANKING AND SUMMARIZATION

Jade Goldstein and Jaime Carbonell
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
jade@cs.cmu.edu, jgc@cs.cmu.edu

## ABSTRACT:

This paper[1] develops a method for combining query-relevance with information-novelty in the context of text retrieval and summarization. The Maximal Marginal Relevance (MMR) criterion strives to reduce redundancy while maintaining query relevance in re-ranking retrieved documents and in selecting appropriate passages for text summarization. Preliminary results indicate some benefits for MMR diversity ranking in ad-hoc query and in single document summarization. The latter are borne out by the trial-run (unofficial) TREC-style evaluation of summarization systems. However, the clearest advantage is demonstrated in the automated construction of large document and non-redundant multi-document summaries, where MMR results are clearly superior to non-MMR passage selection. This paper also discusses our preliminary evaluation of our summarization system.

## 1. INTRODUCTION

With the continuing growth of online information, it has become increasingly important to provide improved mechanisms to find information quickly. Conventional IR systems rank and assimilate documents based on maximizing relevance to the user query [1, 8, 6, 12, 13]. In cases where relevant documents are few, or cases where very-high recall is necessary, pure relevance ranking is very appropriate. But in cases where there is a vast sea of potentially relevant documents, highly redundant with each other or (in the extreme) containing partially or fully duplicative information we must utilize means beyond pure relevance for document ranking.

In order to better illustrate the need to combine relevance and anti-redundancy, consider a reporter or a student, using a newswire archive collection to research accounts of airline disasters. He composes a well-though-out query including "airline crash", "FAA investigation", "passenger deaths", "fire", "airplane accidents", and so on. The IR engine returns a ranked list of the top 100 documents (more if requested), and the user examines the top-ranked document. It's about the suspicious TWA-800 crash near Long Island. Very relevant and useful. The next document is also about "TWA-800", so is the next, and so are the following 30 documents. Relevant? Yes. Useful? Decreasingly so. Most "new" documents merely repeat information already contained in previously offered ones, and the user could have tired long before reaching the first non-TWA-800 air disaster document. Perfect precision, therefore, may prove insufficient in meeting user needs.

A better document ranking method for this user is one where each document in the ranked list is selected according to a combined criterion of query relevance and novelty of information. The latter measures the degree of dissimilarity between the document being considered and previously selected ones already in the ranked list. Of course, some users may prefer to drill down on a narrow topic, and others a panoramic sampling bearing relevance to the query. Best is a user-tunable method that focuses the search from a narrow beam to a floodlight. Maximal Marginal Relevance (MMR) provides precisely such functionality, as discussed below.

If we consider document summarization by relevant-passage extraction, we must again consider anti-redundancy as well as relevance. Both query-free summaries and query-relevant summaries need to avoid redundancy, as it defeats the purpose of summarization. For instance, scholarly articles often state their thesis in the introduction, elaborate upon it in the body, and reiterate it in the conclusion. Including all three in

[1] This research was performed as part of Carnegie Group Inc.'s Tipster III Summarization Project under the direction of Mark Borger and Alex Kott.

versions in the summary, however, leaves little room for other useful information. If we move beyond single document summarization to document cluster summarization, where the summary must pool passages from different but possibly overlapping documents, reducing redundancy becomes an even more significant problem.

Automated document summarization dates back to Luhn's work at IBM in the 1950's [12], and evolved through several efforts including Tait [24] and Paice in the 1980s [17, 18]. Much early work focused on the structure of the document to select information. In the 1990's several approaches to summarization blossomed, include trainable methods [10], linguistic approaches [8, 15] and our information-centric method [2], the first to focus on query-relevant summaries and anti-redundancy measures. As part of the TIPSTER program [25], new investigations have started into summary creation using a variety of strategies These new efforts address query relevant as well as "generic" summaries and utilize a variety of approaches including using coreference chains (from the University of Pennsylvania) [25], the combination of statistical and linguistic approaches (Smart and Empire) from SaBir Research and Cornell University, topic identification and interpretation from the ISI, discourse driven summarization from GE R&D Labs, and template based summarization from New Mexico State University [25].

In this paper, we discuss the Maximal Marginal Relevance method (Section 2), its use for document reranking (Section 3), our approach to query-based single document summarization (Section 4), and our approach to long documents (Section 6) and multi-document summarization (Section 6). We also discuss our evaluation efforts of single document summarization (Section 7) and our preliminary results (Section 8).

## 2. MAXIMAL MARGINAL RELEVANCE

Most modern IR search engines produce a ranked list of retrieved documents ordered by declining relevance to the user's query [1, 18, 21, 26]. In contrast, we motivated the need for '"relevant novelty" as a potentially superior criterion. However, there is no known way to directly measure new-and-relevant information, especially given traditional bag-of-words methods such as the vector-space model [19, 21]. A first approximation to measuring relevant novelty is to measure relevance and novelty independently and provide a linear combination as the metric. We call the linear combination "marginal relevance" -- i.e. a document has high marginal relevance if it is both relevant to the query and contains minimal similarity to previously selected documents. We strive to maximize marginal relevance in retrieval and summarization, hence we label our method "maximal marginal relevance" (MMR).

The Maximal Marginal Relevance (MMR) metric is defined as follows:

Let C = document collection (or document stream)

Let Q = ad-hoc query (or analyst-profile or topic/category specification)

Let R = IR (C, Q, q) – i.e. the ranked list of documents retrieved by an IR system, given C and Q and a relevance threshold theta, below which it will not retrieve documents. (q can be degree of match, or number of documents).

Let S = subset of documents in R already provided to the user. (Note that in an IR system without MMR and dynamic reranking, S is typically a proper prefix of list R.) R\S is the set difference, i.e. the set of documents in R, not yet offered to the user.

$$\text{MMR}(C,Q,R,S) \stackrel{\text{def}}{=} \underset{D_i \in R \setminus S}{\text{Argmax}} [\lambda * \text{Sim}_1(D_i,Q) - (1-\lambda) \underset{D_j \in S}{\text{Max}} (\text{Sim}_2(D_i,D_j))]$$

Given the above definition, MMR computes incrementally the standard relevance-ranked list when the parameter $\lambda=1$, and computes a maximal diversity ranking among the documents in R when $\lambda=0$. For intermediate values of $\lambda$ in the interval [0,1], a linear combination of both criteria is optimized. Users wishing to sample the information space around the query, should set $\lambda$ at a smaller value, and those wishing to focus in on multiple potentially overlapping or reinforcing relevant documents, should set $\lambda$ to a value closer to 1. We found that a particularly effective search strategy (reinforced by the user study discussed below) is to start with a small $\lambda$ (e.g. $\lambda = .3$) in order to understand the information space in the region of the query, and then to focus on the most important parts using a reformulated query (possibly via relevance feedback) and a larger value of $\lambda$ (e.g. $\lambda = .7$).

Note that the similarity metric $\text{Sim}_1$ used in document retrieval and relevance ranking between documents and query could be the same as $\text{Sim}_2$ between documents (e.g., both could be cosine similarity), but this need not be the case. A more accurate, but computationally more costly metric could be used when applied only to the elements of the retrieved document set R, given that $|R| << |C|$, if MMR is applied for re-ranking the top portion of the ranked list produced by a standard IR system.

| **query** : Brazil external debt figure | | | |
|---|---|---|---|
| | | λ | |
| **Article Title** | 1 | 0.7 | 0.3 |
| BRAZIL SEEN AS VANGUARD FOR CHANGING DEBT STRATEGY | 76 | 76 | 76 |
| FUNARO REJECTS UK SUGGESTION OF IMF BRAZIL PLAN | 1308 | 1308 | 1293 |
| ECONOMIC SPOTLIGHT - BRAZIL DEBT DEADLINES LOOM | 1431 | 1431 | 1308 |
| U.S. URGED TO STRENGTHEN DEBT STRATEGY | 104 | **2149** | 133 |
| U.S. URGES BANKS TO DEVELOP NEW 3RD WLD FINANCE | 50 | 104 | 14 |
| FUNARO'S DEPARTURE COULD LEAD TO BRAZIL DEBT DEAL | **2149** | **1388** | **1388** |
| U.S. OFFICIALS SAY BRAZIL SHOULD DEAL WITH BANKS | 1713 | 1293 | 1762 |
| BRAZIL SEEKS TO REASSURE BANKS ON DEBT SUSPENSION | **1388** | 1713 | **2149** |
| BRAZIL SEEKS TO REASSURE BANKS ON DEBT SUSPENSION | **1403** | 50 | **69** |
| BRAZIL CRITICISES ADVISORY COMMITTEE STRUCTURE | 1291 | 133 | 1713 |
| LATIN DEBTORS MAKE NEW PUSH FOR DEBT RELIE | 32 | 1291 | 104 |
| BRAZIL DEBT SEEN PARTNER TO HARD SELL TACTICS | 99 | 99 | 1431 |
| BRAZIL DEBT POSES THORNY ISSUE FOR U.S. BANKS | 54 | 14 | 99 |
| U.S. URGES BANKS TO WEIGH PHILIPPINE DEBT PLAN | 44 | 54 | 1291 |
| U.K. SAYS HAS NO ROLE IN BRAZIL MORATORIUM TALKS | 1293 | 32 | 54 |
| TALKING POINT/BANK STOCKS | 53 | **69** | 44 |
| CANADA BANKS COULD SEE PRESSURE ON BRAZIL LOANS | 1762 | 1762 | 32 |
| TREASURY'S BAKER SAYS BRAZIL NOT IN CRISIS | 133 | 44 | 50 |
| BRAZIL'S DEBT CRISIS BECOMING POLITICAL CRISIS | 14 | **1403** | **1403** |
| BAKER AND VOLCKER SAY DEBT STRATEGY WILL WORK | **69** | 53 | 53 |

Table 1: Initial Relevance Ranking ($\lambda = 1$) vs. MMR reranking ($\lambda = .7$ & $\lambda = .3$)

## 3. DOCUMENT REORDERING

We implemented MMR in two retrieval engines, PURSUIT (an upgraded version of the original retrieval engine inside the Lycos™ search engine), [9] and SMART (the publicly available version of the Cornell IR engine) [1]. Using the scoring functions available in each system for both $Sim_1$ and $Sim_2$, we obtained consistent and expected results in the behavior of the two systems.

The results of MMR reranking are shown in Table 1. In this Reuters document collection, article 1403 is a duplicate of 1388. MMR reranking performs as expected, for decreasing values of l, the ranking of 1403 drops. Also as predicted, novel but still relevant information as evidenced by document 69 starts to increase in ranking. Relevant, but similar to the highest ranked documents, such as document 1713 drop in ranked ordering. Document 2149 's position varies depending on its similarity to previously seen information.

We also performed a pilot experiment with six users who were undergraduates from various disciplines. The purpose of the study was to find out if they could tell what was the difference between the standard ranked document order retrieved by SMART and a MMR reranked order with $\lambda = ..5$. They were asked to perform nine different search tasks to find information and asked various questions about the tasks. They used two methods to retrieve documents, known only as R and S. Parallel tasks were constructed so that one set of users would perform method R on one task and method S on a similar task. Users were not told how the documents were presented only that either "method R" or "method S" were used and that they needed to be try to distinguish the differences between methods. After each task we asked them to record the information found. We also asked them to look at the ranking for method R and method S and see if they could tell any difference between the two. The majority of people said they preferred the method which gave in their opinion the most broad and interesting topics. In the final section they were asked to select a search method and use it for a search task. 80% (4 out of 5) chose the method MMR to use. The person who chose Smart stated it was because "it tends to group more like stories together." The users indicated a differential preference for MMR in navigation and for locating the relevant candidate documents more quickly, and pure-relevance ranking when looking at related documents within that band. Three of the five users clearly discovered the differential utility of diversity search and relevance-only search. One user explicitly stated his strategy:

*"Method R [relevance only] groups items together based on similarity and Method S [MMR re-ranking] gives a wider array. I would use Method S [MMR re-ranking] to find a topic ... and then use Method R [relevance-only] with a specific search from Method S [MMR re-ranking] to yield a lot of closely related items."*

The initial study was too small to yield statistically significant trends with respect to speed of known-item retrieval, or recall improvements for broader query tasks. However, based on our own experience and questionnaire responses from the 6 users, we expect that task demands play a large role with respect to which method yields better performance.

## 4. SINGLE DOCUMENT SUMMARIES

Human summarization of documents, sometimes called "abstraction" is a fixed-length *generic* summary, reflecting the key points that the abstractor -- rather than the user -- deems important. Consider a physician evaluating a particular chemotherapy regimen who wants to know about its adverse effects to elderly female patients. The retrieval engine produces several lengthy reports (e.g. a 300-page clinical study), whose abstracts do not contain any hint of whether there is information regarding effects on elderly patients. A useful summary for this physician would contain *query-relevant* passages (e.g. differential adverse effects on elderly males and females, buried in page 211-212 of the clinical study) assembled into a summary. A different user with different information needs may require a totally different summary of the same document.

We developed a minimal-redundancy query-relevant summarizer-by-extraction method, which differs from previous work in summarization [10, 12, 15, 18, 24] in several dimensions.

- Optional query relevance: as discussed above a query or a user interest profile (for the vector sum of both, appropriately weighted) is used to select relevant passages. If a generic query-free summary is desired, the centroid vector of the document is calculated and passages are selected with the principal components of the centroid as the query.

- Variable granularity summarization: The length of the summary is under user control. Brief summaries are useful for indicative purposes (e.g. whether to read further), and longer ones for drilling and extracting detailed information.

- Non-redundancy: Information density is enhanced by ensuring a degree of dissimilarity between passages contained in the summary. The degree of query-focus vs. diversity sampling is under user control (the $\lambda$ parameter in the MMR formula).

Our process for creating single document summaries is as follows:

1. Segment a document into passages and index the passages using the inverted indexing method used by the IR engine for full documents. Passages may be phrases, sentences, n-sentence chunks, or paragraphs. For the TIPSTER III evaluation, we used sentences as passages.
2. Within a document, identify the passages relevant to the query. Use a threshold below which the passages are discarded. We used a similarity metric based on cosine similarity using the traditional TF-IDF weights.
3. Apply the MMR metric as defined in Section 2 to the passages (rather than full documents). Depending on the desired length of the summary, select a few or larger number. If the parameter $\lambda$ is not very close to 1, redundant query relevant passages will tend to be eliminated and other different, slightly less query relevant passages will be included. We allow the user to select the number of passages or the percentage of the document size (also known as the "compression ratio").
4. Reassemble the selected passages into a summary document using one of the following summary-cohesion criteria:
   - Document appearance order: Present the segments according to their order of presentation in the original document. If the first sentence is longer than a threshold, we automatically include this sentence in the summary as it tends to set the context for the article. If the user only wants to view a few segments, the first sentence must also meet a threshold for sentence rank to be included.
   - News-story principle: Present the information in MMR-ranked order, i.e., the most relevant and most diverse information first. In this manner, the reader gets the maximal information even if they stop reading the summary. This allows the diversity of relevant information to be presented earlier and topic

introduced may be revisited after other relevant topics have been introduced.

- Topic-cohesion principle: First group together the document segments by topic clustering (using sub-document similarity criteria). Then rank the centroids of each cluster by MMR (most important first) and present the information, a topic-coherent cluster at a time, starting with the cluster whose centroid ranks highest.

We implemented query-relevant document-appearance-based sequencing of information. Our method of summarization does not require the more elaborate language-regeneration needed by Kathy McKeown and her group at Columbia in their summarization work [15]. As such our method is simpler, faster and more widely applicable, but yields potentially less cohesive summaries.

Query: Delaunay refinement mesh generation finite element method foundations three dimension analysis; λ = .3

[1] Delaunay refinement is a technique for generating unstructured meshes of triangles or tetrahedra suitable for use in the finite element method or other numerical methods for solving partial differential equations.
[5] The purpose of this thesis is to further this progress by cementing the foundations of two-dimensional Delaunay refinement, and by extending the technique and its analysis to three dimensions.
[15] Nevertheless, Delaunay refinement methods for tetrahedral mesh generation have the rare distinction that they offer strong theoretical bounds and frequently perform well in practice.
[39] If one can generate meshes that are completely satisfying for numerical techniques like the finite element method, the other applications fall easily in line.
[131] Our understanding of the relative merit of different metrics for measuring element quality, or the effects of small numbers of poor quality elements on numerical solutions, is based as much on engineering experience and rumor as it is on mathematical foundations.
[158] Delaunay refinement methods are based upon a well-known geometric construction called the Delaunay triangulation, which is discussed extensively in the mesh generation chapter.
[201] I first extend Ruppert's algorithm to three dimensions, and show that the extension generates nicely graded tetrahedral meshes whose circumradius-to-shortest edge ratios are nearly bounded below two.
[2250] Refinement Algorithms for Quality Mesh Generation: Delaunay refinement algorithms for mesh generation operate by maintaining a Delaunay or constrained Delaunay triangulation, which is refined by inserting carefully placed vertices until the mesh meets constraints on element quality and size.
[3648] I do not know to what difference between the algorithms one should attribute the slightly better bound for Delaunay refinement, nor whether it marks a real difference between the algorithms or is an artifact of the different methods of analysis.

Figure 1: Generic MMR- generated summary of dissertation.

Query: sliver mesh boundary removal small angles; λ = .7

[1] Delaunay refinement is a technique for generating unstructured meshes of triangles or tetrahedra suitable for use in the finite element method or other numerical methods for solving partial differential equations.
[129] Hence, many mesh generation algorithms take the approach of attempting to bound the smallest angle.
[2621] Because s is locked, inserting a vertex at c will not remove t from the mesh.
[2860] Of course, one must respect the PSLG; small input angles cannot be removed.
[3046] The worst slivers can often be removed by Delaunay refinement, even if there is no theoretical guarantee.
[3047] Meshes with bounds on the circumradius-to-shortest edge ratios of their tetrahedra are an excellent starting point for mesh smoothing and optimization methods designed to remove slivers and improve the quality of an existing mesh (see smoothing section).
[3686] If one inserts a vertex at the circumcenter of each sliver tetrahedron, will the algorithm fail to terminate?
[3702] A sliver can always be eliminated by splitting it, but how can one avoid creating new slivers in the process?
[3723] Unfortunately, my practical success in removing slivers is probably due in part to the severe restrictions on input angle I have imposed upon Delaunay refinement.
[3724] Practitioners report that they have the most difficulty removing slivers at the boundary of a mesh, especially near small angles.

Figure 2: Focused-query MMR-generated summary of dissertation.

## 5 SUMMARIZING LONGER DOCUMENTS

The MMR-passage selection method for summarization works better for longer documents (which typically contain more inherent passage redundancy across document sections such as abstract, introduction, conclusion, results, etc.). To demonstrate the quality of summaries that can be obtained for long documents, we summarized an entire dissertation containing 3,772 sentences with a generic topic query constructed by expanding the thesis title (Figure 1). In contrast, Figure 2 shows the results of a more specialized query with a larger λ value to focus summarization less on diversity and more on topic.

The above example demonstrates the utility of query relevance in summarization and the incremental utility of controlling summary focus via the lambda parameter. It also highlights a shortcoming of summarization by extraction, namely coping with antecedent references. Sentence [2621] refers to coefficients "s", "c", and "t," which do not make sense outside the framework that defines them. Such

referential problems are ameliorated with increased passage length, for instance using paragraphs rather than sentences. However, longer-passage selection also implies longer summaries. Another solution is co-reference resolution [25].

## 6. MULTI-DOCUMENT SUMMARIES

As discussed earlier, MMR passage selection works equally well for summarizing single documents or clusters of topically related documents. Our method for multi-document summarization follows the same basic procedure as that of single document summarization (see section 4). In step 2 (Section 4), we identify the N most relevant passages from each of the documents in the collection and use them to form the passage set to be MMR re-ranked. N is dependent on the desired resultant length of the summary. We used N relevant passages from each document collection rather than the top relevant passages in the entire collection so that each article had a chance to provide a query-relevant contribution. In the future we intend to compare this to using MMR ranking where the entire document set is treated as a single document. Steps 2, 3 and 4 are primarily the same.

The TIPSTER evaluation corpus provided several sets of topical clusters to which we applied MMR summarization. In one such example on a cluster of apartheid-related documents, we used the topic description as the query (see Figure 3) and N was set to 4 (4 sentences per article were reranked). The top 10 sentences for $\lambda = 1$ (effectively query relevance, but no MMR) and $\lambda = .3$ (both query relevance and MMR anti-redundancy) are shown in Figures 4 and 5 respectively.

The summaries clearly demonstrate the need for MMR in passage selection. The $\lambda = 1$ case exhibits considerable redundancy, ranging from near-replication in passages [4] and [5] to redundant content in passages [7] and [9]. Whereas the $\lambda = .3$ case exhibits no such redundancy. Counting clearly distinct propositions in both cases, yields a 20% greater information content for the MMR case, though both summaries are equivalent in length.

**Topic:**

<head> Tipster Topic Description
<num> Number: 110
<dom> Domain: International Politics
<title> Topic: Black Resistance Against the South African Government
<desc> **Description**:
Document will discuss efforts by the black majority in South Africa to overthrow domination by the white minority government.
<smry> **Summary**:
Document will discuss efforts by the black majority in South Africa to overthrow domination by the white minority government.
<narr> **Narrative:**
A relevant document will discuss any effort by blacks to force political change in South Africa. The reported black challenge to apartheid may take any form -- military, political, or economic -- but of greatest interest would be information on reported activities by armed personnel linked to the African National Congress (ANC), either in South Africa or in bordering states.
<con> **Concept(s):**
1. African National Congress, ANC, Nelson Mandela, Oliver Tambo
2. Chief Buthelezi, Inkatha, Zulu
3. terrorist, detainee, subversive, communist
4. Limpopo River, Angola, Botswana, Mozambique, Zambia
5. apartheid, black township, homelands, group areas act, emergency regulations

**Query:**

Black Resistance South African Government black majority South Africa overthrow domination white minority government force political change South Africa challenge apartheid military political economic activities armed personnel African National Congress (ANC) South Africa African National Congress ANC Nelson Mandela Oliver Tambo Chief Buthelezi Inkatha Zulu terrorist detainee subversive communist Limpopo River Angola Botswana Mozambique Zambia apartheid black township homelands group areas act emergency regulations

**Query (short version - no concepts):**

lBlack Resistance South African Government black majority South Africa overthrow domination white minority government force political change South Africa challenge apartheid military political economic activities armed personnel African National Congress (ANC) South Africa bordering states

Figure 3: Topic and Query for Tipster Topic 110

[1] [761] AP880212-0060 [15] ANGOP quoted the Angolan statement as saying the main causes of conflict in the region are South Africa's ``illegal occupation'' of Namibia, South African attacks against its black-ruled neighbors and its alleged creation of armed groups to carry out ``terrorist activities'' in those countries, and the denial of political rights to the black majority in South Africa.
[2] [758] AP880803-0080 [25] Three Canadian anti-apartheid groups issued a statement urging the government to sever diplomatic and economic links with South Africa and aid the African National Congress, the banned group fighting the white-dominated government in South Africa.
[3] [756] AP880803-0082 [25] Three Canadian anti-apartheid groups issued a statement urging the government to sever diplomatic and economic links with South Africa and aid the African National Congress, the banned group fighting the white-dominated government in South Africa.
[4] [790] AP880802-0165 [27] South Africa says the ANC, the main black group fighting to overthrow South Africa's white government, has seven major military bases in Angola, and the Pretoria government wants those bases closed down.
[5] [654] AP880803-0158 [27] South Africa says the ANC, the main black group fighting to overthrow South Africa's white-led government, has seven major military bases in Angola, and it wants those bases closed down.
[6] [92] WSJ910204-0176 [2] de Klerk's proposal to repeal the major pillars of apartheid drew a generally positive response from black leaders, but African National Congress leader Nelson Mandela called on the international community to continue economic sanctions against South Africa until the government takes further steps.
[7] [781] AP880823-0069 [18] The ANC is the main guerrilla group fighting to overthrow the South African government and end apartheid, the system of racial segregation in which South Africa's black majority has no vote in national affairs.
[8] [375] WSJ890908-0159 [24] For everywhere he turns, he hears the same mantra of demands -- release, lift bans, dismantle, negotiate -- be it from local anti-apartheid activists or from foreign governments: release political prisoners, like African National Congress leader Nelson Mandela; lift bans on all political organizations, such as the ANC, the Pan Africanist Congress and the United Democratic Front; dismantle all apartheid legislation; and finally, begin negotiations with leaders of all races.
[9] [762] AP880212-0060 [14] The African National Congress is the main rebel movement fighting South Africa's white-led government and SWAPO is a black guerrilla group fighting for independence for Namibia, which is administered by South Africa.
[10] [91] WSJ910404-0007 [8] Under an agreement between the South African government and the African National Congress, the major anti-apartheid organization, South Africa's remaining political prisoners are scheduled for release by April 30.

Fig 4: $\lambda = 1.0$ Multi Document Summarization
[Rank] Document ID [Sentence Number] Sentence

[1] [1] [761] AP880212-0060 [15] ANGOP quoted the Angolan statement as saying the main causes of conflict in the region are South Africa's ``illegal occupation'' of Namibia, South African attacks against its black-ruled neighbors and its alleged creation of armed groups to carry out ``terrorist activities'' in those countries, and the denial of political rights to the black majority in South Africa.
[2] [2] [758] AP880803-0080 [25] Three Canadian anti-apartheid groups issued a statement urging the government to sever diplomatic and economic links with South Africa and aid the African National Congress, the banned group fighting the white-dominated government in South Africa.
[3] [6] [92] WSJ910204-0176 [2] de Klerk's proposal to repeal the major pillars of apartheid drew a generally positive response from black leaders, but African National Congress leader Nelson Mandela called on the international community to continue economic sanctions against South Africa until the government takes further steps.
[4] [8] [375] WSJ890908-0159 [24] For everywhere he turns, he hears the same mantra of demands -- release, lift bans, dismantle, negotiate -- be it from local anti-apartheid activists or from foreign governments: release political prisoners, like African National Congress leader Nelson Mandela; lift bans on all political organizations, such as the ANC, the Pan Africanist Congress and the United Democratic Front; dismantle all apartheid legislation; and finally, begin negotiations with leaders of all races.
[5] [4] [790] AP880802-0165 [27] South Africa says the ANC, the main black group fighting to overthrow South Africa's white government, has seven major military bases in Angola, and the Pretoria government wants those bases closed down.
[6] [11] [334] AP890703-0114 [14] The white delegation chief, Mike Olivier, said the ANC members, including President Oliver Tambo and South African Communist Party leader Joe Slovo, said some white anti-apartheid members of Parliament could make a difference, although the organization believes Parliament as a whole is not representative of South Africans.
[7] [14] [788] WSJ880323-0129 [11] These included a picture of Oliver Tambo, the exiled leader of the banned African National Congress; a story about 250 women attending an ANC conference in southern Africa; a report on the crisis in black education; and an advertisement sponsored by a Catholic group in West Germany that quoted a Psalm and called for the abolition of torture in South Africa.
[8] [12] [303] AP880621-0089 [8] There was no immediate comment from South Africa, which in the past has staged cross-border raids on Botswana and other neighboring countries to attack suspected facilities of the African National Congress, which seeks to overthrow South Africa's white-led government.
[9] [24] [502] WSJ900510-0088 [24] While the membership of Inkatha, the religiously and politically conservative group that is the ANC's chief rival for power in black South Africa, is overwhelmingly Zulu, Inkatha's leader, Mangosutho Buthelezi, has very seldom appealed to sectional tribal loyalties.
[10] [16] [593] AP890821-0092 [11] Besides ending the emergency and lifting bans on anti-apartheid groups and individual activists, the Harare summit's conditions included the removal of all troops from South Africa's black townships, releasing all political prisoners and ending political trials and executions, and a government commitment to free political discussion.

Fig 5: $\lambda = .3$ Multi Document Summarization.
[Rank] [Previous Rank in $\lambda = 1.0$ Version] Document ID [Sentence Number] Sentence

<TITLE>Angola Rejects South African Proposal for Peace Talks
</TITLE>
<TEXT>
[1] Angola has rejected a South African proposal for a regional peace conference that would include Angolan rebels, Angola's official ANGOP news agency reported Friday.
[14] ANGOP quoted the Angolan statement as saying the main causes of conflict in the region are South Africa's ``illegal occupation'' of Namibia, South African attacks against its black-ruled neighbors and its alleged creation of armed groups to carry out ``terrorist activities'' in those countries, and the denial of political rights to the black majority in South Africa.
</TEXT>

Figure 6: Single Document Summary AP880212-0060, 10% of document length.

As can be seen from the above summaries, multi-document synthetic summaries require support in the user interface. In particular, the following issues need to be addressed:

- Attributability: The user needs to be able to access easily the source of a given passage. This could be the single document summary (see Figure 6).
- Contextually: The user needs to be able to zoom in on the context surrounding the chosen passages.
- Redirection: The user should be able to highlight certain parts of the synthetic summary and give a command to the system indicating that these parts are to be weighted heavily and that other parts are to be given a lesser weight.

## 7. EVALUATION OF SINGLE DOCUMENT SUMMARIZATION

An ideal text summary contains the relevant information for which the user is looking, excludes extraneous information, provides background to suit the user's profile, eliminates redundant information and filters out relevant information that the user knows or has seen. The first step in building such summaries is extracting the relevant pieces. We performed a pilot evaluation in which we examined how well a summarization system could extract the relevant sections of documents, through use of a database of assessor marked relevant sentences.

Automatically generating summaries based on a query or high frequency words from the text can produce a reasonable looking summary, yet this summary can be far from the optimal goal of readable, useful, intelligible, appropriate length summaries from which the information that the user is seeking can be extracted. Evaluation of summarization systems can be intrinsic (measuring a system's quality) or extrinsic (measuring a system's performance in a given task) [7].

In the past year, there has been a focus in TIPSTER on both the intrinsic and extrinsic aspects of summarization evaluation [4]. The evaluation consisted of three tasks (1) determining document relevance to a topic for query-relevant summaries (an indicative summary), (2) determining categorization for generic summaries (an indicative summary), (3) establishing whether summaries can answer a specified set of questions (an informative summary) by comparison to an ideal summary. In each task, the summaries are rated in terms of confidence in decision, intelligibility and length. Jing, Barzilay, McKeown and Elhadad [6] performed a pilot experiment (40 sentences) in which they examined the performance (precision-recall) of three summarization systems (one using notion of number of sentences, the other two using numbers of words or number of clauses). They compared the performance of these systems against human ideal summaries and found that different systems achieved their best performances at different lengths. They also found the same results for determining document relevance to a topic (one of the TIPSTER tasks) for query-relevant summaries.

Our approach to summarization is different from Columbia and TIPSTER in that the focus is not on an "ideal human summary" of any particular document cutoff size. An ideal summarization system must first be able to recognize the relevant sentences (or parts of a sentence) for a topic or query and then be able to create a summary from these relevant sentences. Although a list of words, an index or table of contents, is an appropriate label summary and can indicate relevance, informative summaries need at least noun-verb phrases. Thus, we evaluated summarization systems for the first stage - coverage of relevant sentences. Other systems [16, 23] use the paragraph as a summary unit. Since the paragraph consists of more than one sentence and often more than one information unit, it is not as suitable for this type of evaluation, although it may be more suitable for a construction unit in summaries due to the additional context that it provides, for example., paragraphs will often solve coreference issues. One of the issues in summarization evaluation is how to

score (penalize) extraneous non-useful information contained in a summary.

Unlike document information retrieval, text summarization evaluation has not addressed the performance of different methodologies in a manner in which different components can be separated out. Most summarization systems use linguistic knowledge as well as a statistical component [3, 5, 16, 23]. We performed a preliminary evaluation of the certain monolingual query expansion [28] and pseudo-relevance feedback [20]. The evaluation will be by a modified version of the 11-pt average recall/precision. (see Section 7.3).

## 7.1 Query Expansion

We expanded the original queries in three ways: (1) adding the results of the original query passed through WordNet with options for nouns, verbs, adjectives and adverbs; extracting all senses that were presented; and eliminating exact duplicate words. (2) adding the results of the original query expanded by one sense of WordNet, i.e., the first sense presented by WordNet. (3) adding the highest ranked sentence of the document (a form of pseudo-relevance feedback).

## 7.2 Compression

We used a document sentence compression factor based on the number of sentences in the document. Since the sentences selected by the system tend to be longer than the average sentence, the output summary ends up being slightly longer than the actually compression factor, i.e. 25% document length is actually 25% of the sentences in the document and is actually slightly higher than 25% of the characters in the document

## 7.3 Evaluation Code

The 11-pt average precision recall [21] commonly used for document information retrieval evaluation was modified for summarization. Since many documents only have a few relevant sentences, 11-pt curves have a lot of bins (intervals) with missing data items, whereas in the case of document retrieval this would only occur for cases where the number of relevant documents for the query is less than 11. To remedy this situation, we implemented a step function for the precision values. This caused recall intervals that previously had the value of 0 (which would not naturally occur in document retrieval) to be assigned an actual precision value. We used this modified non-interpolated 11-pt average precision recall for our results.

## 7.4 Experiment Design

We created two data sets for our pilot experiments. For the first {110 Set} we took 50 documents from the TIPSTER evaluation provided set of 200 news articles spanning 1988-1991. All these documents were on the same topic (see Figure 3). Three evaluators ranked each of the sentences in the document as relevant, somewhat relevant and not relevant. For the purpose of this experiment, somewhat relevant was treated as relevant and the final score for the sentence was determined by a majority vote. The sentences that received this majority vote were tabulated as a relevant sentence (to the topic). The document was ranked as relevant or not relevant. All three assessors had 68% agreement in their relevance judgments. The query was extracted from the topic (see Figure 3).

In order to evaluate what the relevance loss for a diversity gain in single document summarization, we created summaries for two document length percentages and determined how many relevant sentences the summaries contained.

The results are given in Table 3 for compression factors 0.25 and 0.1. Two precision scores were calculated, (1) that of TREC relevance plus at least one CMU assessor marking the document as relevant (yielding 23 documents) and (2) at least two of the three CMU assessor marking the document as relevant (yielding 18 documents). From these scores we can see there is no significant statistical difference between the $\lambda$=1, $\lambda$=.7, and $\lambda$=.3 scores. This is often explained by cases where the $\lambda$=1 article failed to pick up a piece of relevant information and the reranking of $\lambda$=.7 or .3 might or vice versa. The baseline (baseln) contains the first N sentences of the document, where N is the number of sentences in the summary.

The second data set {Model Summs} was providing as a training set for the Question and Answer portion of the TIPSTER evaluation. It consisted of marked sentences that would contribute to a model summary that contained answers to specific questions. We extracted these sentences to provide our sentence answer file. The query was extracted from the questions.

| Sentence Precision | | | |
|---|---|---|---|
| Percentage of Document Length | λ | TREC and CMU Relevant | CMU Relevant |
| 10% | 1.0 | 0.78 | 0.83 |
| 10% | 0.7 | 0.76 | 0.83 |
| 10% | 0.3 | 0.74 | 0.79 |
| 10% | Baseline | 0.74 | 0.83 |
| 25% | 1.0 | 0.74 | 0.76 |
| 25% | 0.7 | 0.73 | 0.74 |
| 25% | 0.3 | 0.74 | 0.76 |
| 25% | Baseline | 0.60 | 0.65 |

Table 2: Precision Scores

| | Model Summaries | 110 Set |
|---|---|---|
| task | Q & A | indicative summaries |
| number of documents | 48 | 50 |
| source | provided by Tipster | 3 people marked each sentence |
| relevant documents | all | 15 |
| average sentences per document | 22.6 | 25.1 |
| median sentences per document | 19 | 23 |
| maximum sentences per document | 51 | 50 |
| minimum sentences per document | 11 | 11 |
| query formation | provided questions | topic |
| statistics | all documents | 40 documents |
| percent of document length | 19.4% | 24.9% |
| summary includes first sentence | 72% | 47%, 73% (only relevant docs) |
| average summary size (sentences) | 4.3 | 6.1 |
| median summary size (sentences) | 4 | 5 |

Table 3: Data Set Comparison

| 110 Set Comparison | Relevant Documents | Non-Relevant Documents |
|---|---|---|
| number of documents | 15 | 25 |
| average sentences per document | 27.5 | 23.8 |
| median sentences per document | 23 | 23 |
| maximum sentences per document | 51 | 44 |
| minimum sentences per document | 15 | 11 |
| percent of document length | 36.2% | 17.7% |
| summary includes first sentence | 73% | 32% |
| average summary size (sentences) | 10.1 | 3.7 |
| median summary size (sentences) | 7 | 4 |

Table 4: 110 Set - Relevant vs. Non-Relevant Documents with relevant sentences.

The data set statistics are shown in Tables 3 and 4. Note that non-relevant documents (Table 4) still have a high percentage of relevant sentences. 10 documents in the 110 set were non-relevant and had no relevant sentences.

Summaries were compared using the 11-pt non-interpolated average precision recall code as previously described.

The following experiments were performed:

- *query expansion techniques*: evaluate the effects of query expansion methods (pseudo-relevance feedback and expansion using WordNet) (see Section 7.1).
- *compression effects*: examine the effects of sentence compression on the recall/precision values. The output number of sentences is the compression factor multiplied by the number of sentences in the original document. (see Section 7.2)

## 8. RESULTS

From the results in Figures 7 and 8, we can see that the summaries did a reasonable job of obtaining relevant sentences for the initial sentences picked.

Ideally, we would have like to have the initial point of the graphs be at 1.0, meaning that the first sentence(s) selected are always deemed relevant by the judges or one(s) that was part of the model summaries provided by Tipster.

## 8.1 Query Expansion

For the 110 Set (see Figure 7), expansion using pseudo-relevance feedback (prf) slightly helped the score and expansion using WordNet [27] detracted from the score. However, this was a fairly well-formed query. In the case of the Model Summaries where there was not a well formed query, but the query was taken from a series of questions, overall WordNet and pseudo-relevance feedback slightly helped the results (see Figure 8). This indicates that WordNet and pseudo-relevance feedback might significantly help in cases of short queries, by helping to obtain key concepts mentioned elsewhere in the document for the summary. Our results are similar to those obtained for document information retrieval [27]. We plan to explore this further by creating shorter queries and looking at the effects of query expansion. Other areas we plan to explore are the use of the first sentence in the query. (72% of the first sentences were marked relevant - see Table 3) and the addition of the document title.

.

## 8.2 Compression

An important evaluation criteria for summarization is what is the ideal summary output length (compression of the document) and how does it affects the user's task. From Tables 2 & 3, we can see that the summary length or number of relevant sentences chosen per document can vary significantly. In Figure 9, we examine the effect of compression on recall and precision and in Figure 10, we show a plot of F1. This F1 graph indicates that for lower compression (less document percentage chosen) it helps to have the pseudo-relevance feedback (also see in the graphs in that section). Pseudo-relevance feedback helps obtain sentences that might not normally be extracted or might be extracted later. As the number of sentences that are allowed in the summary grows, these sentences have a higher probability of being chosen for the summary and pseudo-relevance feedback might select other sentences that are not as relevant or it might not add anything to what has been selected. We need to do more studying on the effects of query expansion and compression on summarization, as well as see how our preliminary results hold for additional data sets.

# CONCLUSION

We have shown that MMR ranking provides a useful and beneficial manner of providing information to the user by allowing the user to minimize redundancy. This is especially true in the case of query-relevant multi-document summarization in this one data collection. We are currently performing studies on how this extends to several document collections. In the future we will also be investigating how to handle coreference in our system as well as analyzing the most suitable $\lambda$ parameters and clustering the output results.

Text Summarization is still in the infant stage in terms of evaluation. Many document information retrieval results can be applied to text summarization, but as of yet, there has been little extensive evaluation of these techniques. This pilot experiment showed many areas that need to be examined in further detail including: (1) when, where and how to apply query expansion, and (2) the exploration of pseudo-relevance feedback for multiple sentences, including titles and the first sentence in a document. We also plan to (1) compare our summaries to a baseline summary, i.e., the first N sentences of a document, where N is the result of the compression factor applied to the number of sentences, (2) fix the number of sentences for each document as the number of relevant sentences chosen by the judges and investigate the system performance, and (3) evaluate the performance of other information retrieval models such as GVSM and LSI. We are currently in the process of building a more extensive sentence relevance database for further evaluation.

An important question in summarization is what makes a good summary? Our experiments are designed around extracting relevant portions of documents. We must still put together these relevant sections to produce a human understandable, readable and non-redundant summary. The first two criteria may involve coreference resolution and possibly natural language generation. We believe MMR can address the latter. Another question is how to produce a usable summary for a user? Since users' backgrounds, interests and levels of education differ, summaries of the same document may need to include differing information depending on user profiles. We also need to address differing types of summaries, such as time-line summaries (presenting information summarized in time) and novel information summaries based on information not included in a previous collection of summaries.

## REFERENCES


[1] C. Buckley, Implementation of the SMART Information Retrieval System. *Department of Computer Science Technical Report* Cornell University, TR 85-686.

[2] J.G. Carbonell, and J. Goldstein, The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, In *Proceedings of SIGIR 98*, Melbourne, Australia, 24-28 August 1998, p. 335-336.

[3] J. Cowie, K. Mahes, S. Nirenbug, R. Zajac, MINDS -- Multi-lingual Interactive Document Summarization, *AAAI Intelligent Text Summarization Workshop*, p. 131-1328, Stanford, CA March 1998

[4] T.F. Hand, A Proposal for Task-Based Evaluation of Text Summarization Systems In *ACL/EACL-97 Summarization Workshop}*, 31-36, Madrid, Spain., July 1997.

[5] E. Hovy and C.Y. Lin, Automated Text Summarization in SUMMARIST, In *ACL/EACL-97 Summarization Workshop*, 18-24, Madrid, Spain July 1997

[6] H. Jing, R. Barzilay, K. McKeown, M. Elhadad, Summarization Evaluation Methods Experiments and Analysis, AAAI Intelligent Text Summarization Workshop, p. 60-68, Stanford, CA March 1998

[7] K.S. Jones and J.R. Galliers, *Evaluation Natural Language Processing Systems: an Analysis and Review*. New York: Springer 1996

[8] J.L Klavans and J. Shaw, Lexical Semantics in Summarization, In *Proceedings of the First Annual Workshop of the IFIP Working Group FOR NLP and KR*, Nantes, France, April 1995.

[9] G. Kowalski, *Information Retrieval Systems: Theory and Implementation*, Kluwer Academic Publishers, 1997.

[10] J.M. Kupiec, J. Pedersen, J. and F. Chen, A Trainable Document Summarizer, In *Proceedings of the 18th Annual Int. ACM/SIGIR Conference on Research and Development in IR*, Seattle, WA, July 1995, pp. 68-73.

[11] D.D. Lewis, B. Croft, B., and N. Bhandaru, "Language-Oriented Information Retrieval," *International Journal of Intelligent Systems*, Vol 4 (3), Fall 1989.

[12] H.P. Luhn, Automatic Creation of Literature Abstracts, *IBM Journal*, 1958, pp. 159-165.

[13] M.L Mauldin, Retrieval Performance in FERRET: A Conceptual Conference on Research and Development in Information Retrieval, *Proceedings of the 14th International Conference on Research and Development in Information Retrieval*, October 1991.

[14] M.L. Mauldin and J.R. Leavitt, Web Agent Related Research at the Center for Machine Translation. In *Proceedings of SIGNIDR V*, McLean Virginia, August 1994.

[15] K. McKeown, J. Robin, and K. Kukich, Empirically Designing and Evaluating a New Revision-based Model for Summary Generation. In *Information Processing and Management*, 31 (5) 1995.

[16] M. Mitra, A. Singhal and C. Buckley, Automatic Text Summarization by Paragraph Extraction, In *ACL/EACL-97 Summarization Workshop,* 39-46, Madrid, Spain July 1997.

[17] C.D. Paice, Automatic Generation of Literature Abstracts - An Approach Based on the Indification of Self-Indicated Phrases, in *Information Retrieval Research*, R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen and P.W. Williams, editors, Butterworths, London, 1981, 172-191.

[18] C.D. Paice, Constructing Literature Abstracts by Computer: Techniques and Prospects, In *Information Processing and Management*, Vol. 26, 1990, pp.171-186.

[19] G. Salton G and C. Buckley Improving retrieval performance by relevance feedback. *Journal of American Society for Information Sciences*, 41:288-297, 1990.

[20] G. Salton *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley 1989.

[21] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, McGraw-Hill Computer Science Series, 1983.

[22] G. Salton, A. Singhal, M. Mitra,. and C. Buckley, Automatic Text Structuring and Summarization, *Information Processing and Management*, 33(2), 193-208, 1997.

[23] T. Strzalkowski, J. Wang, and B. Wise, A Robust Practical Text Summarization, *AAAI Intelligent Text Summarization Workshop*, p. 26-3, Stanford, CA March 1998.

[24] J.I. Tait, *Automatic Summarizing of English Texts*, PhD dissertation, University of Cambridge, 1983.

[25] TIPSTER Text Phase III 18-Month Workshop, Fairfax, VA 4-6 May, 1988,

[26] C.J. van Riesburg, *Information Retrieval*, London Butterworths 1979.

[27] J. Xu and B. Croft. Query expansion using local and global document analysis in 19th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96), pages 4-11, 1996.

[28] E.M. Vorhees. Using Wordnet to disambiguate words senses for text retrieval. In *Proceedings of ACM SIGIR Conference (SIGIR '93*) pages, 171-180, 1993.

Figure 7: 110 Set: Query Comparison, document length percentage = 25%

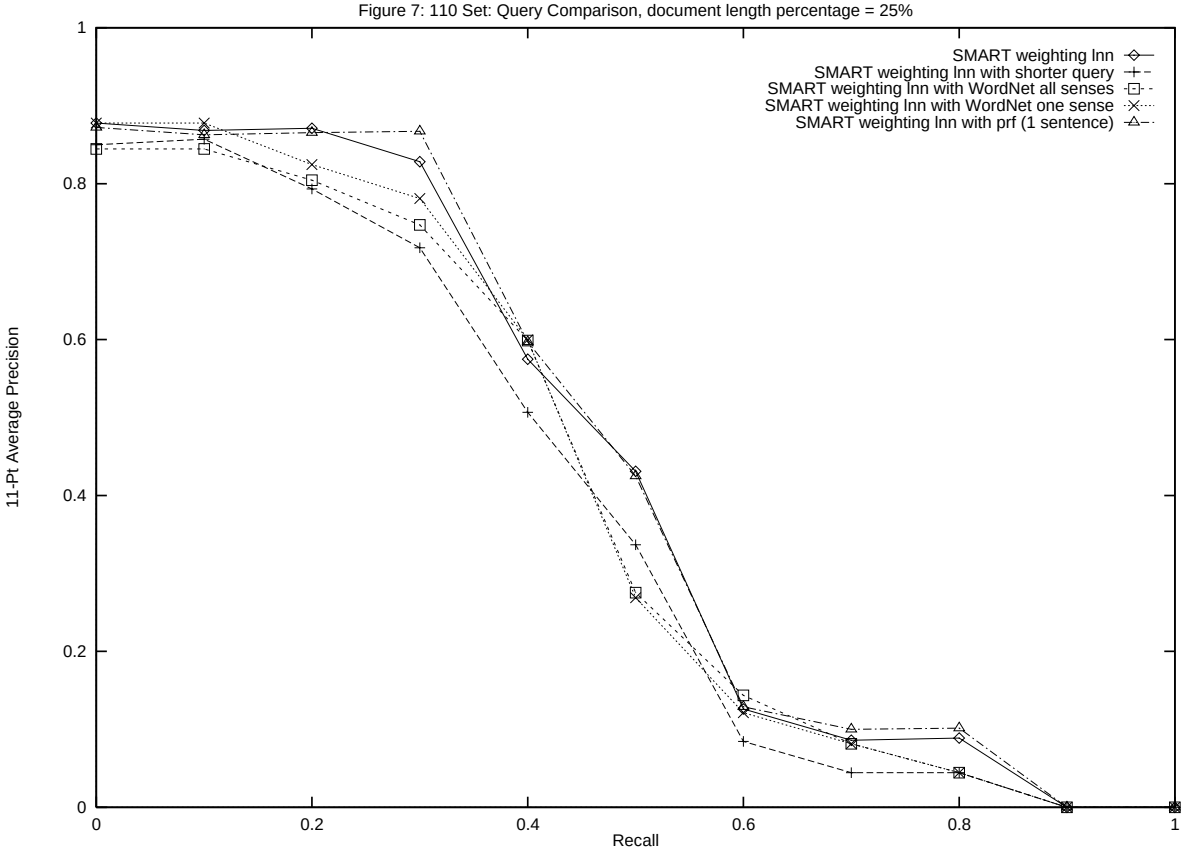



Figure 8: Model Summaries: Query Comparison, document length percentage = 25%

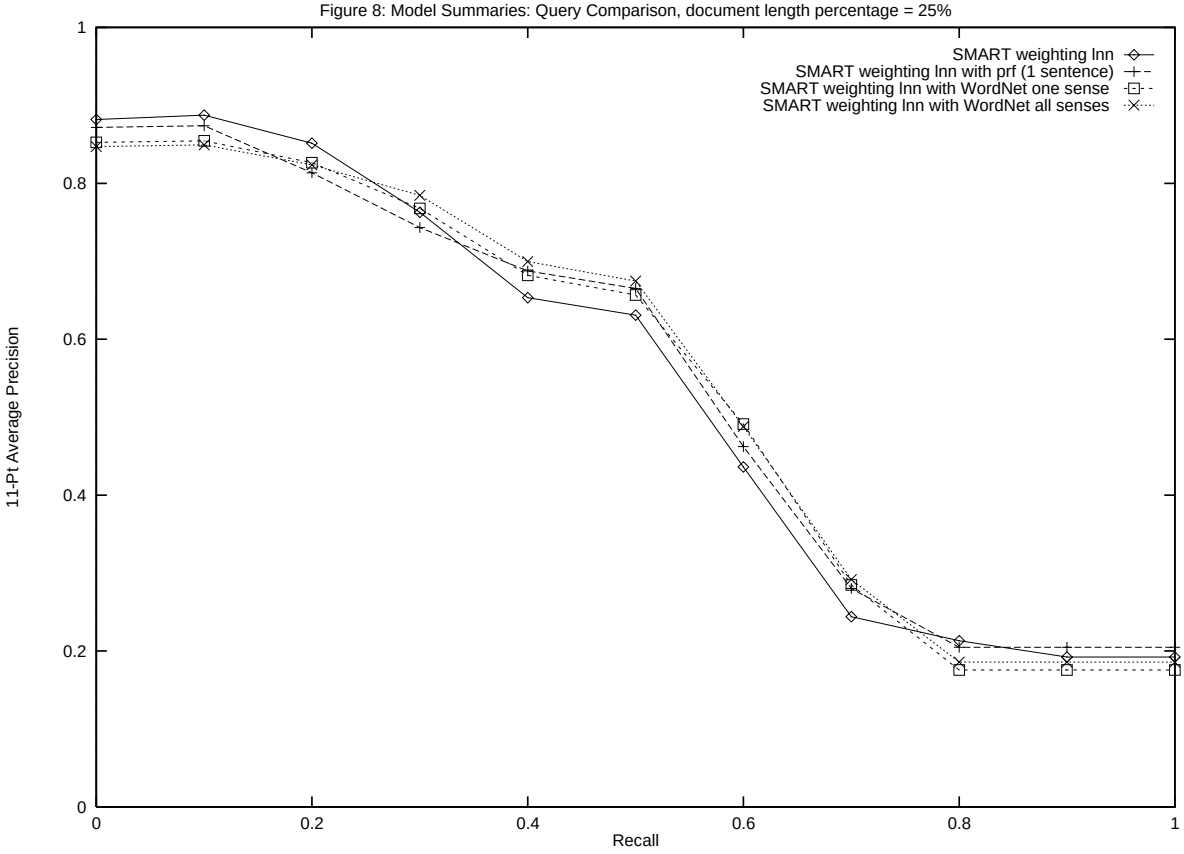


Figure 9: Model Summaries Effects of Compression

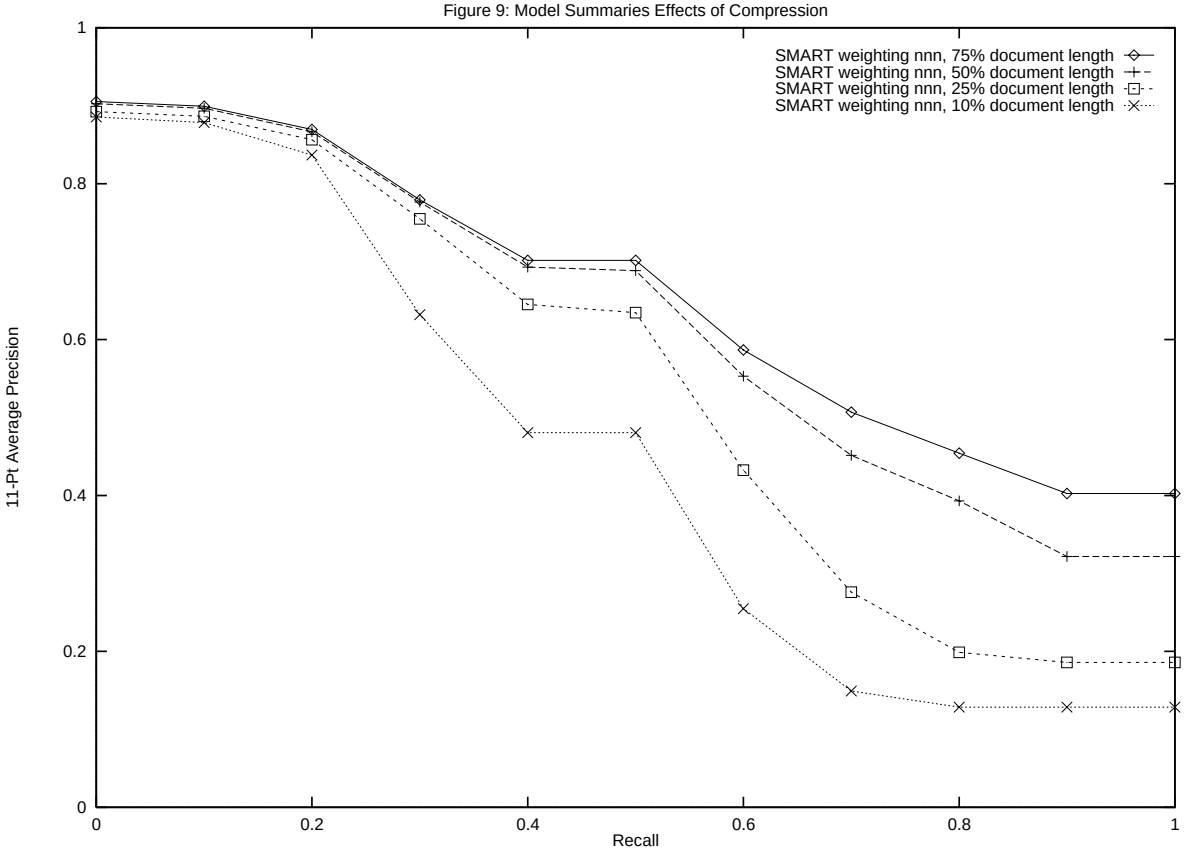
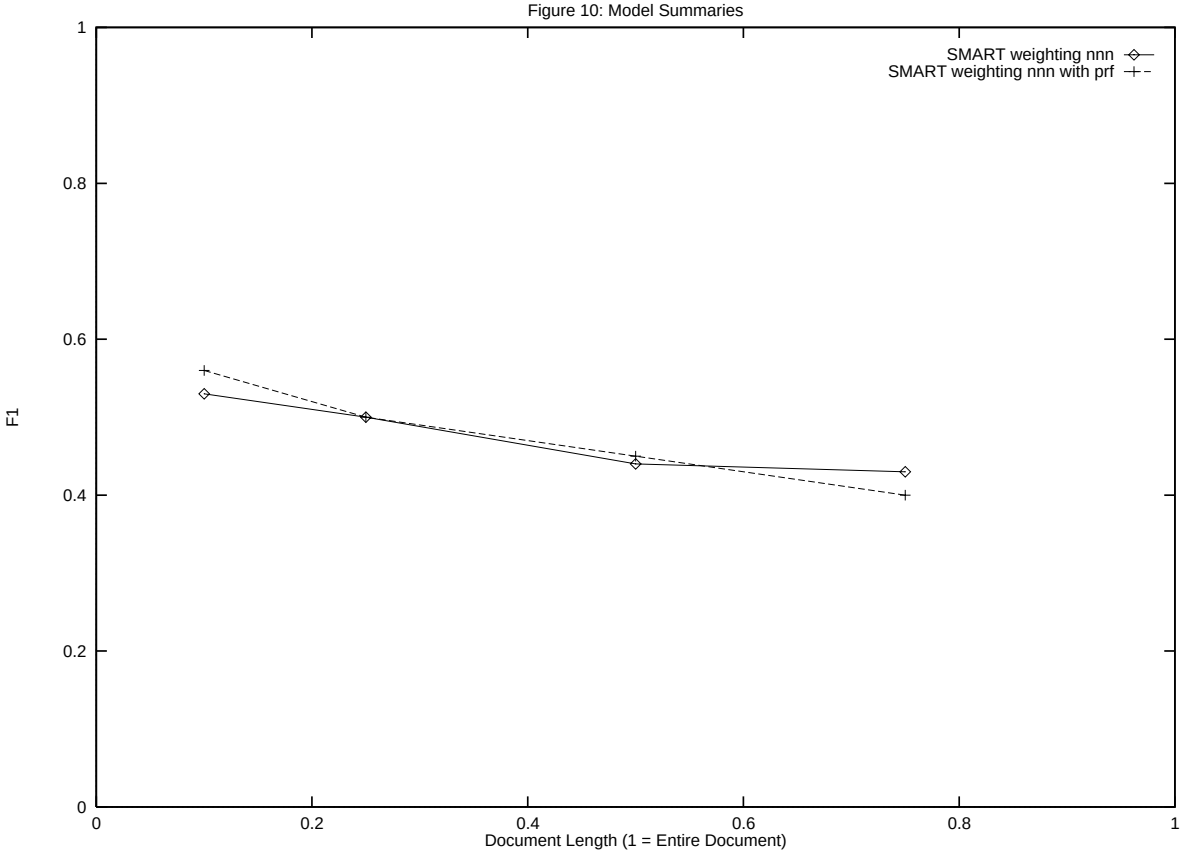



Figure 10: Model Summaries



# Posters and demonstrations

# Evaluation of an automatic abstracting system

Michael P.Oakes & Chris. D. Paice
Computing Department
Lancaster University
Lancaster LA1 1YR, England
oakes,cdp@comp.lancs.ac.uk

## ABSTRACT

The Concept-Based Abstracting system reduces the full text of an academic article to a list of domain-specific roles and suitable fillers. In the domain of agriculture, a role might be SPECIES and its fillers "maize" and "soybean". In this paper we use the measures of strict and premissive recall and precision to compare the machine-generated list of roles and fillers with the corresponding human-selected lists of "ideal" roles and fillers.

**Keywords:** Information Extraction, Text Summarisation, Recall, Precision.

## 1 INTRODUCTION

In this paper we describe a method of evaluating the Concept-Based Abstracting (CBA) system of Paice & Jones [1], where the goal is the automatic abstracting of journal articles, initially in the field of crop protection. The method is to build a set of contextual templates against which the original text is compared. The templates are designed so that they match the text at points of high information content, where inferences can be made about which concepts most truly reflect the content of the document.

Templates are in the form of alternating literals and fillers, where each literal must match some point in the text exactly, and a filler can be any sequence of words occurring between literals. Filler strings found by matching templates are assigned roles specific to each template. These roles correspond to slots in a frame which is used to represent the document as a whole. For example the template "fertilised with ?" would match the phrase "fertilised with procymidone". The interpretation associated with the filler of this template might be "AGENT", and thus the filler "procymidone" would be assigned to this role. This provides evidence that the main chemical agent reported in the text is procymidone. The full set of roles used for crop protection articles is agent (AGE), cultivar (CV), high level property (HLP), influence (INF), laboratory (LAB), low level property (LLP), location (LOC), pest (PES), soil (SOI) and species (SPE).

After the text has been read in, all the fillers found for each role are collated. The substrings may be weighted, since some templates are more reliable than others. If any substring of a filler is found more than once, the weight associated with each instance is combined. The weight for each substring is also enhanced if the substring occurs in the title of the document or can be lexically validated [2] by matching a word or phrase in a lexicon of terms known to be feasible role fillers. The most highly weighted substring found in this way is taken to be the most likely interpretation of a given role.

**Table 1. Sample output from the CBA program: List of the most highly weighted candidate substrings for the role of SPECIES.**

| SPE weight | text | freq | cap | tit | lex |
|---|---|---|---|---|---|
| 26.0 | maize | 13 | 1.0 | 1.0 | 2.0 |
| 14.0 | soybean | 7 | 1.0 | 1.0 | 2.0 |
| 4.0 | irrigated row | 2 | 1.0 | 2.0 | 1.0 |
| 4.0 | irrigated | 2 | 1.0 | 2.0 | 1.0 |
| 3.0 | plant | 3 | 1.0 | 1.0 | 1.0 |
| 1.8 | plant growth | 2 | 1.0 | 1.0 | 0.9 |

Table 1 shows six of the most highly weighted candidate substrings for the role of SPECIES, extracted from a paper entitled "Controlled traffic to increase productivity of irrigated row crops in the semi-arid tropics". In this example, all templates are given a weight of 1, and the weight for each candidate string is the number of times the string was matched in the text (see column "weight"), multiplied by 2 if the role is CULTIVAR or LOCATION and the string commences with a capital letter in the text (see column "cap"), then multiplied by 2 if the string appears in the title of the paper (see column "tit"), and finally multiplied by 2 if the string can be lexically validated, i.e. occurs in a list of words or phrases known to be appropriate to the role in question. The overall weight is multiplied by 0.9 if the string occurs in one of the lexical validation lists appropriate to a different role (See column "lex").

Sometimes more than one filler is appropriate for a given role in a given paper. To decide on the number of candidate strings to be assigned to a particular role, we use the following heuristic:

a) the candidate string should be among the $n$ most highly weighted;

b) the weight of the candidate string must be more than $x$ times the weight of the most highly-weighted candidate string;

c) the weight of the candidate string must be greater than a threshold $y$;

d) no candidate string may be accepted if it is a substring of another candidate string which has been accepted.

Thus if $n = 6$, $x = 0.5$ and $y = 2$, "maize" and "soybean" would both be selected for the role of species from the list of candidate strings given in Table 1.

The production of a coherent abstract from the list of roles and selected fillers for a given paper is achieved using the "Select and Generate" approach [3].

## 2 EVALUATION OF THE CBA SYSTEM

To retrieve the appropriate fillers for each role, two subtasks must be performed: (a) to pinpoint where in a text each relevant concept is mentioned; and (b) to select or construct an appropriate string to represent each concept. This suggests doing evaluation at two levels: (a) where we count even a partially matching string as correct, and (b) where the selected string matches one of the candidates exactly. Even under (b), there may be two or more acceptable ways to express a concept, so as far as possible our evaluation must allow any acceptable alternative expression.

Level (a) may not only apply to partially matching phrases, but also to exactly matching expressions which we describe as "weak", typically because too broad; an example is the selection of the generic term "fungicide" where the specific fungicide in question is "LB-pickel".

For each document an evaluation agenda is drawn up, which contains a series of sections, each relating to a different role such as SPECIES or PEST. The entries in this agenda are agreed upon by two human judges, and are compared with the corresponding machine-generated expressions. Each section comprises a series of role instances, ALL of which we want to identify (this allows for papers which discuss two crops or three influences etc). Each concept instance comprises a list of concept expressions, any ONE of which should be identified.

The human-generated evaluation agenda for the article entitled "Influence of foliar-applied fungicides on seed yield of faba bean (Vicia Faba L.) in northern New South Wales" is shown in Table 2, while the machine-generated list of roles and their fillers is shown in Table 3. In Table 2, the symbol "/", as in "edible oil linseed/linola" denotes that either "edible oil linseed" or "linola" should be accepted as a full match. "Weak" expressions are flagged by attaching a "?" symbol.

When comparing the human-judged evaluation agenda with the machine-generated results (list of roles and their fillers) for the same paper, the degree and type of match is found for each word in the machine-generated output with respect to the human-judged evaluation agenda. Only fillers corresponding to the same role can be compared.

a) strong, exact match: the machine-generated and human-selected strings are identical, and the human-selected string is not flagged with "?";

b) strong, partial match: the machine-generated string is a superstring of the human-

selected string, which is not flagged with "?";

c) strong, partial match: the machine-generated string is a substring of the human-selected string, which is not flagged with "?";

d) weak, exact match: the machine-generated and human-selected strings are identical, and the human-selected string is flagged with "?" to denote a weakly-matching concept;

e) weak partial match: the machine-generated string is a superstring of the human-selected string, which is flagged with "?";

f) weak, partial match: the machine-generated string is a substring of the human-selected string, which is flagged with "?";

g) no match: all other cases.

**Table 2. Evaluation agenda for an article in the domain of crop protection.**

| |
|---|
| AGENT<br>1. mancozeb / fungicide?<br>2. dichlofluanid / fungicide?<br>3. tebucozole / fungicide?<br>4. vinclozolin / fungicide?<br>5. procymidone / fungicide? |
| CULTIVAR<br>1. fiord |
| HIGH LEVEL PROPERTY<br>1. seed yield |
| INFLUENCE |
| LABORATORY<br>1. plot / outdoor field / field |
| LOCATION<br>1. northern new south wales |
| LOW LEVEL PROPERTY<br>1. diseases scored<br>2. area of leaf affected |
| PEST<br>1. chocolate spot / botyris fabae<br>2. rust / uromyces viciae fabae |
| SOIL |
| SPECIES<br>1. faba bean / vicia faba l. |

In our evaluation, strong, exact matches (a) are considered the best, followed by the various combinations of partial and weak matches (b) to (f), followed by no match (g) at all. In making these comparisons, the best-match criterion of Gaussier, Langé & Meunier is employed [4]. Only matches between each machine-generated term and the best matching human-selected term are considered. However, if the best-matching term in the human-selected list can be matched even more strongly with a different machine-generated term, the match with the first machine-generated term is not considered.

**Table 3. Machine generated list of roles and fillers corresponding to Table 2.**

| |
|---|
| AGENT<br>1. liberal<br>2. fungicide<br>3. mancozeb<br>4. seed yield |
| CULTIVAR<br>1. fiord |
| HIGH LEVEL PROPERTY<br>1. seed yield |
| INFLUENCE<br>1. foliar-applied fungicide |
| LABORATORY<br>1. plot |
| LOW LEVEL PROPERTY<br>1. seed weight<br>2. yield |
| LOCATION<br>1. new south wales |
| PEST<br>1. chocolate spot<br>2. disease |
| SOIL<br>1. clay |
| SPECIES<br>1. faba bean<br>2. yield |

In many information retrieval experiments, the effectiveness of the retrieval is measured using the measures of recall and precision. Recall (R) measures the proportion of relevant material actually retrieved in response to a search [5]. Since our ultimate aim is to retrieve ALL the

human-selected terms by machine, and the number of matching terms in both lists is the number of relevant terms actually retrieved, our formula for recall is as follows:

$$R = \frac{number\_of\_matches}{number\_of\_human-selected\_terms}$$

Precision (P) is the proportion of retrieved items which are actually relevant. In our case this is the number of matches (number of relevant items obtained) divided by the total number of items in the machine-generated list (total number of relevant items), i.e.

$$P = \frac{number\_of\_matches}{no.\_of\_machine-generated\_terms}$$

The total number of strong, exact matches is used to derive strict measures of recall and precision, denoting the system's ability to generate appropriate strings for each concept. The sum of the totals for all types of matches (strong or weak, full or partial) is used to derive permissive measures of recall and precision, denoting the system's ability to identify the points in the text in which each concept may be found.

In comparing the data for Tables 2 and 3, strong, exact matches were found for the machine-generated terms "mancozeb", "fiord", "seed yield", "plot", "chocolate spot" and "faba bean". A weak, exact match was found for "fungicide", and a strong partial match was found for "new south wales", which is a substring of "northern new south wales". Thus there were 6 strong, exact matches, and 8 matches in all. The total number of machine-generated terms was 16, and the total number of human-selected terms was 14. Thus strict recall was 6 / 14 = 0.43, permissive recall was 8 / 14 = 0.57, strict precision was 6 / 16 = 0.38 and permissive precision was 8 / 16 = 0.50. Results are averaged over all the texts in a test collection (20 texts have been used in our studies so far). Averaging is done by first summing the numerators and denominators of the R and P formulae over the entire test collection. If desired, separate values can be computed for each role.

## 3 CONCLUSION

The measures described in this paper are continually being used to evaluate new features which might be included into the Concept-Based Abstracting system such as lexical validation, and for the optimisation of the heuristics for deciding how many fillers should be allocated to each role.The measures will next be used to evaluate proposed word frequency-related enhancements to the system.

## ACKNOWLEDGEMENT

The Concept-Based Method for Automatic Abstracting project is supported by the British Library, award number CSA 7406.

## REFERENCES

[1] C.D. Paice & P.A. Jones, The identification of important concepts in highly structured technical papers, *ACM SIGIR '93,* Pittsburgh, PA, USA, pages 69-77, 1993.

[2] M.P. Oakes & C.D. Paice. Term extraction for automatic abstracting. In D. Bourigault, C. Jacquemin & M.-C. L'Homme, editors, *Proceedings of Computerm '98,* First Workshop on Computational Terminology, University of Montreal, Quebec, Canada, pages 91-95, 1998.

[3] P.A. Jones & C.D. Paice. A 'select and generate' approach to automatic abstracting. In A. M. McEnery & C. D. Paice, editors, *14th British Computer Society Information Retrieval Colloquium,* Workshops in Computing, Springer Verlag, London, England, pages 141-154, 1992.

[4] E. Gaussier, J.M. Langé & F. Meunier. Towards bilingual terminology. In *Proceedings of the Joint ALLC/ACH Conference,* Oxford University Press, pages 121-124, 1992.

[5] G. Salton & M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983.

# Sumatra: A system for Automatic Summary Generation

Danny H. Lie
Carp Technologies
Hengelosestraat 174, 7521 AK Enschede, The Netherlands
lie@carp-technologies.nl

## ABSTRACT

This paper describes a system for automatic summary generation called Sumatra. It differs from other systems in being domain independent and, instead of relying on statistical techniques, it uses a Natural Language Processing approach, involving parsing, semantic analysis and text generation. The system has been evaluated by using final exam texts from the Dutch grammar school in summarizing. The main conclusion is that the Sumatra system is adequately capable of extracting the important information elements from a text.

**Keywords:** Language Technology, Automatic Text Summarization, Natural Language Understanding

## 1 INTRODUCTION

This paper describes a system for automatic summary generation called Sumatra. This system has been developed entirely in the Java programming language during a M. Sc. Thesis at the University of Twente in collaboration with Medialab (Origin) [1].

First, a definition of a summary will be given (2). Next, the summarizing strategy used in Sumatra will be described (3). Finally, the evaluation of the system will be described (4) followed by the main conclusions (5).

## 2 WHAT IS A SUMMARY?

Most summary definitions state in more or less words that a summary is simply an abbreviated version of a document. For example, [2] defines a summary as a condensation of the main ideas in an article and [3] defines it as a text reduced to its main points. Most summary definitions – like the ones above – contain a clause stating that a summary only contains the crucial information elements needed to understand the text. This is however not a very practical definition, because a summary needs to contain the information the *user* needs, and this information will not be the same for every user. Therefore, for the purpose of this document, I will define a summary as follows:

> *A summary is a selection from a collection of information elements.* Definition 1

This definition however, does not mention the purpose of a summary. People use summaries in many different ways, and for each use a different type of summary exists [4]. I identify the following different summary types and purposes:

1. An abstract can be used to determine whether a text is relevant to read.
2. A summary or a synopsis can be used to save time by reading the summary instead of the entire text.
3. An overview can be used to better understand a text by hiding irrelevant details and highlighting important relations.
4. An outline can be used to read about some excerpted points that the reader finds relevant.

## 3 SUMMARIZATION STRATEGY

Summarizing can be done on the following three levels:

1. Omitting information.
2. Compacting a sentence by restating it in fewer words.
3. Aggregating sentences to form one sentence that conveys the same meaning.

Dumb strategies only use the first one, while more sophisticated strategies should be capable of performing all three. Many sophisticated strategies however, are domain dependent which limits them to texts about a specific topic.

Unlike most other sophisticated strategies, the Sumatra system uses a *domain independent* strategy, which consists of the following steps:

1. Parse the text and build a semantic structure from it.
2. Prune this semantic structure.
3. Generate a new text from the semantic structure.

The Sumatra system uses the following architecture to perform these steps:

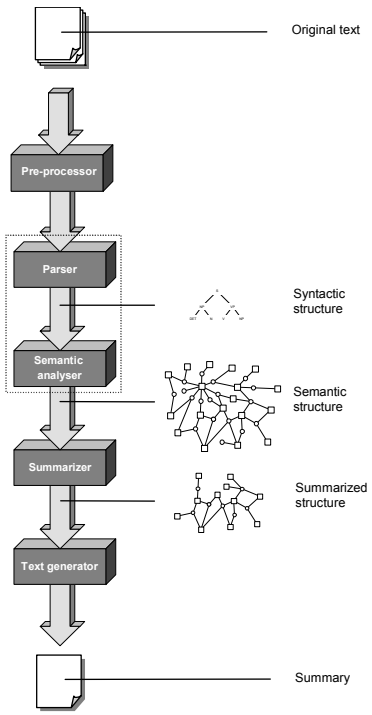



Figure 1 – Architecture of the Sumatra system

The input of the system consists of a text. The output consists of a summarized version of that text. Let's take a closer look at the different components of the system.

## Pre-processor

Before a text can be processed it has to be pre-processed. This task consists of segmenting the text into paragraphs, sentences and words, and looking up the words in a lexicon.

## Parser

The parser analyzes the input text and generates a syntactic structure. Furthermore, the parser uses a rewrite strategy during parsing to reduce the complexity of a sentence and restates it in fewer words [5].

## Semantic analyzer

This component uses the output of the parser to build a meaning representation of the text: a semantic structure. Note that in the Sumatra system, the parser and the semantic analyzer are put together in the same component.

## Summarizer

The summarizer takes a semantic structure and prunes it in such a way that the most important parts remain.

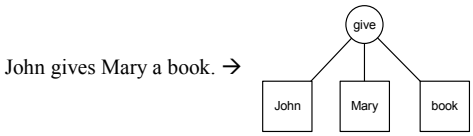
## Text generator

The text generator uses the output of the summarizer – a summarized semantic structure – to generate a new text. Furthermore, the text generator is responsible for the aggregating of sentences.

The rest of this document will only discuss the semantic analysis and the summarization step.

# 3.1 SEMANTIC ANALYSIS

A text can be viewed as a collection of relations between objects. For example, the sentence "John gives Mary a book." denotes some relation between the objects "John", "Mary", and "a book". The goal of the semantic analysis is to identify these relations, and store them in a semantic structure:




Example 1 – Conversion between sentence and relation

If the semantic analyzer identifies a relation, it is stored in a semantic structure. If a node already exists, it will not be stored twice, but shared, resulting in a semantic network of objects and relations.

# 3.2 SUMMARIZATION

Because the semantic analyzer emits a semantic structure in the form of a graph, the task of omitting information can be considered as pruning in this graph. Two pruning methods will be described: a relation oriented pruning method (3.2.1) and a node oriented pruning method (3.2.2).

## 3.2.1 Relation oriented pruning method

A relation oriented graph-pruning method selects certain relations from a semantic structure, together with their arguments, and discard the rest of the structure. Figure 2 shows the selection of relation *A* and the resulting summarized structure.

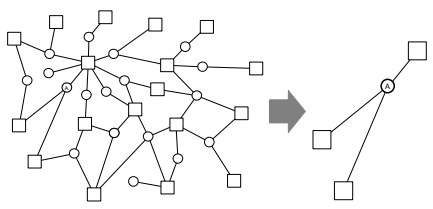


Figure 2 – The selection of a relation

Prior to selecting relations, some algorithm must be used to decide which relations to select. Such an algorithm could assign an importance value to each relation, and discard all relations with an importance value below a certain threshold, or select *n* relations with the highest importance value, and discard the rest of the structure. Either method uses an importance value.

There are several ways to define the importance value of a relation. A possible definition could use the connectivity of a relation as its importance value. The connectivity of a relation is defined as the number of connections that must be severed to cut the relation and all its arguments free from the structure. Consider, for example, the following semantic structure:

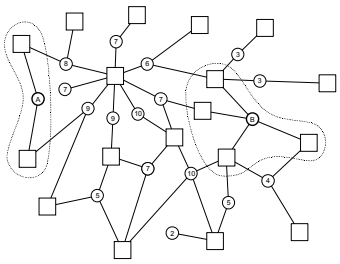


Figure 3 – The connectivity of a relation

The numbers inside the relations indicate their connectivity. Relation A has a connectivity of 2 and thus an importance value of 2, whereas relation B has a connectivity of 8. When using connectivity as a definition of importance, relation B is more important than relation A. This can be defined as follows:

*The importance value v of a relation r equals the sum of the degrees of r's arguments minus the number of r's arguments.* Definition 2

Or, more formally:

*Let r be the relation and n its number of arguments*
*importance(r)=* **Error!***– n*
*argumentDegree(r,x)=The degree of r's* $x^{th}$ *argument*

Using Definition 2, let's select the six most important relations and discard the rest. This would result in the following structure:

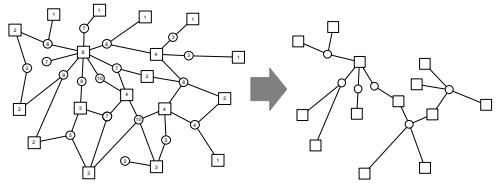


Figure 4 – Selecting the six most important relations using Definition 2.

## Heuristics

Beside connectivity, other parameters can be used to determine the importance value of a relation. The first and last sentence of a paragraph are usually more important than other sentences, and the occurrence of certain signal words in a sentence can indicate that it is important. The importance value of relations that have been derived from such sentences can be altered by multiplying it with a certain boost factor. The Sumatra system uses the following heuristics to alter the importance value if a relation conforms to certain conditions:

| Condition | Boost factor |
|---|---|
| First sentence of a paragraph | 4 |
| Last sentence of a paragraph | 4 |
| Contains an enumeration signal word* | 4 |
| Contains a quantor signal word* | 1.5 |
| Contains a signal word* | 2 |
| Contains an example signal word | 0.5 |

* Only one of these boost conditions are applied to a relation

The values for these boost factors have been obtained by extensive testing with six texts used in the final exams of the Dutch grammar school. For each text, a list of the relevant information elements was available and a script has been used to automatically determine the percentage of relevant information elements a summary contains. The values for the boost factors have been varied to maximize this percentage. The combination of values for the boost factors have been chosen manually, and because of the enormous search space of this optimizing problem, it is very likely that a better combination exists. A better combination could be found by using a neural network or genetic algorithm to find a higher (local) maximum.

### 3.2.2 Node-oriented pruning method

Instead of selecting relations, a node-oriented graph pruning method selects certain nodes from a semantic structure, together with the relations they participate in, and the other arguments of these relations. The rest of the structure will be discarded. Figure 5 shows the selection of node A and the resulting summarized structure:

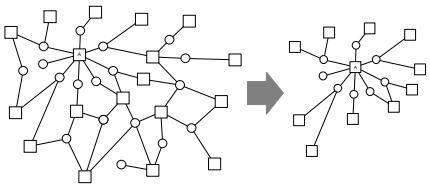


Figure 5 – The selection of a node

Just like in the relation-oriented method, an importance value must be defined to decide which nodes to select. Using the connectivity thought, one could easily define the importance value of a node as its degree:

*The importance value v of a node n equals the degree of n* Definition 3

Although this node-oriented graph pruning method seems more flexible than the relation-oriented method – because the user can easily interact with it – it does not provide much control over the size of the summarized structure. When the user selects an extra node, all relations that include this node as an argument are added to the structure, whereas the relation-oriented method provides a more granular control over the number of relations. A hybrid approach however, is possible and can give us the best of both worlds.

## 4 EVALUATION

Among other things, the Sumatra system has been evaluated by letting it summarize some final exam texts from the Dutch grammar school in summarizing. The following graph lists the results of this evaluation.

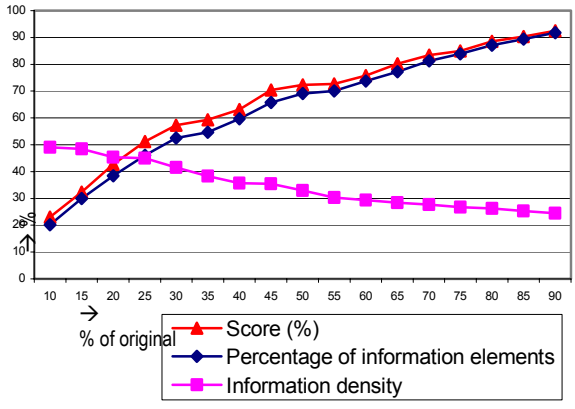



Figure 6 - Evaluation

The results have been obtained by using the official solutions for the final exams. These solutions contain all relevant information elements and their importance. Using these importance values, a score have been calculated. As can be seen in the graph above, the score is higher than the percentage of information elements. This indicates that the Sumatra system favors important information elements above less important ones.

Furthermore, the graph shows that the Sumatra system is able to extract almost 50% of the relevant information elements when the summary size is set to 25% of the original.

## 5 CONCLUSION

It is feasible to build a functional domain independent automatic summary generation system. Furthermore:

- The definition of a summary depends on its usage.
- A text can be seen as a graph: a collection of objects with relations between them. A text can be summarized by pruning this graph.
- Summarization can be done on three levels: 1. Omitting information, 2. Restating a sentence in fewer words and 3. Aggregating sentences. These three tasks are performed by three different components in the Sumatra system. The parser is able to restate a sentence during parsing by means of string rewrite rules, the summarizer decides what information to omit and finally, the text generator aggregates compatible sentences.
- The Sumatra system is adequately capable of extracting the important information elements from a text.

## REFERENCES


[1] D.H. Lie, Automatic Summary Generation: a Natural Language Processing approach (M. Sc. Thesis), *University of Twente*, 1998.

[2] J. Reid, The Process of Composition, *Prentice-Hall Inc.*, New Jersey 1982.

[3] H. Farrell, How To Write A Summary, *Learning Centre of the University of Western Sydney*, 1991.

[4] K.S. Jones, Automatic summarising: factors and directions, *Computation and Language* (cmp-lg/9805011), 1998.

[5] D.H. Lie, J. Hulstijn, R. op den Akker, A. Nijholt, A Transformational Approach to NL Understanding in Dialogue Systems, *Centre of Telematics and Information Technology (CTIT)*, *University of Twente*, 1997.

# Towards Automatic Indexing and Retrieval of Video Content: the VICAR system


Marten den Uyl
Sentient machine Research
Baarsjesweg 224
1058AA Amsterdam
The Netherlands
Denuyl@smr.nl

Ed S. Tan
Word and Image Studies
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV
Amsterdam, The Netherlands
tane@let.vu.nl

Heimo Müller
Joanneum Research[1]
Steyrergasse 17
A 8010 Graz
Austria
mue@smr.nl

Peter Uray
Joanneum Research
Steyrergasse 17
A 8010 Graz
Austria
p.uray@joanneum.ac.at



## ABSTRACT

VICAR[2] is a system for automatic Video Indexing, Cataloguing, Annotation, and Retrieval. Four functionalities are discussed: indexing, interpretation, interrogation and instruction. A brief glimpse is offered at the system architecture, and some of its components reviewed, including image processing, feature abstraction and indexing. Examples of the system's main use in the context of television archives are given. It has been concluded that VICAR's technology my be fruitfully combined with language technology in order to enhance its performance, while the system as it is may be of help in handling mixed media databases, such as ones containing conversation and speech in video.



**Keywords:** Video Analysis, Video Archiving, Video Cataloguing, Video Indexing, Video Retrieval, Language Technology, Multimedia Information Retrieval.


## 1 INTRODUCTION

### 1.1 WHAT IS VICAR?

VICAR stands for Video Indexing, Classification, Annotation and Retrieval. The VICAR system can be used in a televison archive environment as a support in cataloguing and retrieving huge amounts of video [3], [4]. VICAR automates part of the cataloguing work done by documentalists, while leaving control to them, and taking over a great many routine actions.

VICAR 's enhanced search facilities offer relief for the major bottle-neck in retrieving archive materials for reuse, the selection of target shots by viewing lots of tapes. [7] VICAR automatically creates

- semantic indices
- annotations and in addition to present retrieval practice, it
- enables query by image

VICAR is based on cutting edge technology, including the latest developments in image processing, adaptive (neural network) classification and matching techniques, massively parallel computation algorithms and an intuitive user interface.

### 1.2 WHAT DOES VICAR DO? THE FOUR I'S OF VICAR

VICAR's core system functionalities have to be distinguished from its main uses by documentalists and producers, that will be dealt with in the last section of this paper, `How to use VICAR'. Core functionalities go beyond those of present state of the art applications in the area of video indexing and retrieval [1, 8], in that VICAR's index space is huge and contains not only image features, but also semantically interpretable data.

#### 1.2.1 Indexing

After *segmentation* of the video stream, sequences, shots and frames are the source materials for feature extraction. Features serve as indices for each particular piece of video. Indices produced by VICAR are based on a wide variety of features. On the one hand, they include video-specific ones, such as frame colour, distribution of brightness over a frame, camera movement and changes in contrast over a shot. On the other, features related to content are extracted to be part of an index such as the presence of persons, textures and objects. Taken together, the features are integrated into a rich *index strucure* that is linked to time-coded segments of the original video

[1] Presently at Word and Image Studies,Vrije Universiteit Amsterdam, The Netherlands

[2] Partners in the VICAR Consortium: Sentient Machine Research (NL), Joanneum Research (A), VCPC vienna Center for Parallel Computation (A), Österreichischer Rundfunk (A), Südwestfunk Baden-Baden (D), Sveriges Television AB (S), Nederlands Audiovisueel Archief (NL), Faculteit der Letteren Vrije Universiteit (NL), Koot Management Concultancy (NL)

tape. The complete collection of an archive can be indexed and the indices, in turn, can be integrated in a huge index space.

### 1.2.2 Interpretation

The index structure is both very rich and uninterpreted. You cannot discern what is in a segment of a tape just by looking at its index structure. A second functionality in VICAR is interpretation of index structures. It delivers *classification*, a labeling of video content units in terms of some class of objects that is relevant to the user. The class may consist of objects in a narrower sense, e.g. cars, buildings, but also of certain people, events and so on.

On the basis of classification, VICAR *recognizes* objects, first as an instance of a class (‘a car’) and then as specific objects by indicating a found match of an object with some target within a class (e.g. ‘a lorry’).

VICAR's classification is followed by a report, which is an *annotation*. Annotation can follow more or less straighforwardly from classification: each classification label acts as a key term. The system can also be trained to assign extended verbal descriptions to interpreted index structures, e.g. by refining its classification or by combining various classification outputs (`a blue lorry moving left; background is a street with trees')

The standard VICAR Video Explorer is equipped to classify, recognize and annotate *VIPs* (‘Clinton’, ‘Kohl’, ‘Martina Hingis'), *settings* (‘interior’, ‘forest’, ‘park’, ‘mountain scenery') and *movement* (‘walking', ‘running', ‘turmoil', but also ‘camera moves right')

### 1.2.3 Interrogation

VICAR's index and annotation database can be accessed in ways analogous to current query and retrieval practice. The user composes queries using key terms and free text matching extended annotation. The sytem returns video segments that have matching annotations. Extensions to existing practice are the following:

1) associative query by text. Matching is not limited to annotation. The query profile is transformed from a textual format into an integrated index, and this is matched with the large index structure of all videos.
2) associative query by image. The user inputs a frame, shot or sequence, VICAR assigns an index structure to it, that can be matched in various ways with the large index structure of all videos.

Like the current facilities, VICAR enables the user to search in various cycles, that is refine a profile by repeatedly selecting and weighing returned materials into a new query set. Thus, the user can search for targets that are similar in various ways, e.g. as to colour, shape, texture, or semantically.

The advantages of the VICAR for retrieval will be clear. First, the set of unrelevant materials can be decreased. As a number of features is available in the database that are at present too time consuming to annotate, VICAR supports a more detailed search and retrieval mechanism. In principle, the length of each shot is automatically registered, as is camera movement. Other features that are relevant for editing, such as desired or to be avoided overall colour and object movement of target footage, can be accessed as well. Presence and accessibility of these features means that the producer does not have to view a larger number of tapes to select the usable shots. Second, more imagery that is in store, but cannot be retrieved in principle, because it has not been annotated can be acessed through repeated associative search, approaching what one wants by refining rough initial examples to ones that are progressively closer to an ideal.

### 1.2.4 Instruction

Indexing is based on *training*. Specialized modules for indexing or *FAM*s have been fed with numerous examples in order extract those features that are required for classification of certain object. For instance, VICAR’s *VIP finder* has been trained to detect persons and deliver the features of faces that are necessary to recognize each particular VIP.

Interpretation, likewise, is the result of *training* of VICAR’s recognition and classification modules. The index structure being rich, any interpretation will always involve a part of it. This means that any given interpretation can be extended and altered. New target objects can be added to classes by the user - however, in some cases this may mean that new features have to be extracted and integrated into the index structure. For instance, when VICAR classifies cars, new cars can be added. Also, classifications can be refined to suit the user's ends. As importantly, VICAR can be trained to deal with completely new classes of object.

Instruction of new modules involves mainly the selection of training materials and specifying the distinctions and classifications that VICAR has to make. All other work is done by the a specialized function of VICAR's, the *FAM Generator*. A user friendly environment for training is being developed. By user friendly we mean that no knowledge of neural computation is required in order to instruct a module.

# 2 How does VICAR do it? A glimpse at system architecture[2]

## 2.1 VICAR Basic Image Processing

### 2.1.1 Image processing

Procedures employed in VICAR are preform image segmentation and low level feature extraction. This is used for cutting down the vast amount of information contained in a frame into tractable pieces. An important issue is the robustness of the segmentation procedures against MPEG artefacts as well as against changes in brightness, contrast and aspect ratio which must be ensured for any image processing method in order to be useful in a real-world application scenario.

In addition to segmentation, low level image processing can also be used for extracting features which might constitute valuable information for subsequent processing by neural networks. Examples for such features are texture, average color or boundary shape. The low level image processing is the bridge between the raw data and highly sophisticated content extraction technologies like neural networks.

## 2.2 VICAR feature abstraction and indexing

Feature abstraction modules (FAMs) have low level image elements, such as pixel blobs and splashes, as their input. Each *FAM* has knowledge in the form of *traces* of the objects it has to recognize. Traces consist of large arrays of numbers, each representing some low-level image features. Traces have been obtained by training the *neural networks* of the module. Traces represent rather simple objects and properties of video content. In order to classify and annotate composite objects and events, the output of several FAMs are being combined. In a series of different focussings, the input stream is divided across the various modules. The VICAR Video Explorer has three standard FAMs for faces, setting components and body movement. FAMs are being generated by the FAM Generator or *FAMG*.

The partial indices based on the features delivered by the FAMs are being integrated into a single, rich index structure by the Index Integrator or ININ. *Index structures at this stage refer to semantically uninterpreted, abstract, numerical representations, and not to textual content indexes manually assigned by viewing in the process of cataloguing as is usual at present*.

Finally, the *VAG* or Video Annotation Generation Module classifies indexes, and here the data are being semantically interpreted. A neural network is trained to assign input indexes to one of a number of predefined classes. Each of these classes has a text label to it, which is the basis of the annotation. This is what the VICAR Video Explorer does. Future extensions of the VICAR system will allow for open text annotation, based on a natural and extensible lexicon of descriptive terms and phrases.

In conclusion, VICAR realizes feature abstraction, indexing and annotation by finding unique patterns of massive amounts of low level characteristics for objects. These patterns have been acquired through training of neural networks.

# 3 How to use VICAR. A look at the user interface

The VICAR Video Explorer has a database that contains MPEG-shots, together with their formal description and index structures. The user can add new materials to the database (indexing and annotation) and retrieve target materials from the database.

The user interface is based on an environment for indexing by viewing developed as a support for documentalists working with digitalized video. *EUROMEDIA* is an *intuitive video browser* that offers immediate access to programmes, sequences, shots and frames. Screens allow the documentalist to view and manipulate these units (e.g. change the order of shots, arrange them in similarity based groups, etc.). An *indexing editor* is constantly available either in the foreground or in the background, for formal and content indexing and annotation. The system is directly linked to various databases used in cataloguing.

## 3.1 Indexing and annotation

Depending on the exact nature of the FAMs, the VICAR Video Explorer automatically generates index structures (see above) and annotations per shot. Annotations are written out in columns next to key frames for each shot. The documentalist retains continuous control over the process. Annotations proposed by the system can be overwritten by the documentalist. Semantic interpretations exceeding the bounds of VICAR's knowledge, such as historical significance of events, are added by the documentalist. The documentalist can also instantly search for similar materials already in the database, in order to check annotations and attune the indexing process.

The match of the user interface with a particular archive's standards and formats of indexing and annotation, as well as connectivity with other information systems, can be optimized on demand.

[2] This section has been derived from [6].

## 3.2 QUERY AND RETRIEVAL

The producer or archive employee can query by text, as usual, and by visual example, using stills or video sequences. Targets also can be frames, shots or sequences. In response to a query, the system delivers a set of matching sequences, shots, or frames. These are presented as thumb nails or key frames, while their annotations are also one click away at most. Depending on the storage capacity available to the user, videos can either be played instantly in the browser, or be retrieved from a tape store. In any query low-level image features like these, can be combined with semantic objects that the VICAR Video Explorer can classify. In this respect, the use of VICAR is similar to well-known image query systems like QBIC [2], see also [5]. Text queries are translated into abstract index structures, using the annotation database, queries by image require image indexing, unless the image is already part of the database. In addition, the interface guides the user in a type of query that goes completely beyond current limits of querying. Each returned set of materials that match with a given query, can be refined by selecting the best examples, without having to explicitly specify in what respects it is better. For instance, a user can search for Bill Clinton in an exterior shot and then select those shots that show him against a background of sky, and in a few cycles find the shot in which only the president and the sky are visible.

Query by example requires additonal control of aspects and criteria of matching. The user can indicate temporal segments and regions of interest, and also select values for image parameters that are represented in the abstract rich index structure, such as colour distribution and contrast level and camera movement. This functionality will considerably diminish the number of false positives, that is the number of tapes the content of which match the query profile, but in viewing turn out to be of no use due to undesired image characteristics (movement, colour match, etc.)

# 4 CONCLUSION: VICAR AND LANGUAGE TECHNOLOGY

VICAR combines vanguard technologies in the areas of image processing, adaptive learning and user interfaces in order to lift main obstacles in current television archive practice.

- Like all video browsers its user interface decreases tape handling not resulting in reuse.
- Unlike others, it allows for query that is both more flexible and precise than any existing facility.
- Finally, an unusually rich description of materials, based on a more complete indexing of content features, allows for retrieval of materials that are at present unaccessible in principle, opening up whole new and as yet unforeseen contexts of reuse

VICAR can be a most useful tool in any application dealing with large quantities of spoken or written languages accompanied by video images, such as videotaped speech corpora. When combined with speech or graphic language recognition, it affords bimedial indexing of content, allowing queries combining speech and image content. For instance, it seems feasible to retrieve segments of speech by the person who utters it, and possibly even by dynamic non-verbal expressions, gestures and body positions that can be recognized from the video. Future research may focus on the possibility to use VICAR's basic machinery in recognizing lip and jaw movements.

# REFERENCES


[1] Aigrain, P., Zhang, H., & Petkovic, D. (1996). Content-based representation and retrieval of visual media: A state of the art review. *Multimedia Tools and Applications*, *11*, Special issue on Representation and Retrieval of Visual Media.

[2] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., & Yanker, P. (1996). Query by image and video content: the QBIC system. *IEEE Computer*, Sept 1995, 23-31.

[3] Fournial, C. (1986). Documentary analysis of moving images. In IFTA, *Panorama of audiovisual archives*, 167-173.

[4] Green, E.-L. (1993). Indexing and information retrieval of moving images - experiences from a large television information database. Proceedings of the Conference on Online Information.

[5] Hibino, S., & Rundensteiner, E.A. (1996). MMVIS: Design and implementation of a Multimedia Visual Information Seeking environment. Proceedings of Multimedia 96, 75-86. New York

[6] H. Müller, M. den Uyl (Eds.) (1998). *System architecture*. VICAR-T1.1-JRS-001.03-250298.

[7] E. Tan (Ed.) (1997). *End user requirements*. VICAR-T1.2-VUA-002.04-301297

[8] Zhang, H.J., Low, C.Y., Smoliar, S.W, & Wu, J.H. (1997). Video parsing, retrieval and browsing: an integrated and content-based solution. In M. Maybury (Ed.), *Intelligent multimedia information retrieval*, 139-158. AAAI Press/MIT Press.

# Access, Exploration and Visualization of Interest Communities: The VMC Case Study (in Progress)

Anton Nijholt
Centre of Telematics and Information Technology
University of Twente
PO Box 217, 7500 AE Enschede, the Netherlands
anijholt@cs.utwente.nl

## ABSTRACT

This paper discusses a virtual world for representing information and natural interactions about performances in an existing theatre. Apart from mouse and keyboard input, interactions take place using speech and language. It is shown how this virtual environment can be considered as an interest community and it is shown what further research and development is required to obtain an environment where visitors can retrieve information about artists, authors and performances, can discuss performances with others and can be provided with information and contacts in accordance with their preferences.

**Keywords:** Virtual Reality, Speech and Language Interactions, Information Filtering, Agent Technology

## 1 INTRODUCTION

World Wide Web allows interactions and transactions through WebPages using speech and language, either by inanimate or live agents, image interpretation and generation, and, of course the more traditional ways of presenting explicitly pre-defined information by allowing users access to text, tables, figures, pictures, audio, animation and video. In a task- or domain-restricted way of interaction current technology allows the recognition and interpretation of rather natural speech and language in dialogues. However, rather than the current two-dimensional web-pages, the interesting parts of the Web will become three-dimensional, allowing the building of virtual worlds inhabited by interacting user and task agents and with which the user can interact using different types of modalities, including speech and language interpretation and generation. Agents can work on behalf of users, hence, human computer interaction will make use of 'indirect management', rather than interacting through direct manipulation of data by users.

In this paper we discuss a virtual world for representing information and natural interactions about performances in an existing theatre. Apart from mouse and keyboard input interactions take place using speech and language. It is shown how this virtual environment can be considered as an interest community and it is shown what further research and development is required to obtain an environment where visitors can retrieve information about artists, authors and performances, can discuss performances with others and can be provided with information and contacts in accordance with their preferences.

## 2 INFORMATION SPACES FOR INTEREST COMMUNITIES

Web-based digital cities have been around for some years. Like many computer games they have evolved from text environments to 2-dimensional graphical and 3D virtual environments with sounds, animation and video. Visitors, or maybe we should call them residents, of these cities visit libraries, museums, pubs, squares, etc., where they can get information, chat with others, etc. In these environments people get the feeling of being together, they are listening to each other and, in

general, they take responsibility for the environment and theirs and others behavior in such environments.

Today there are examples of virtual spaces that are visited and inhabited by people sharing common interests. With virtual spaces or environments we want to refer to computer accessible environments where users (visitors, passers-by) can enter 3D environments, browse (visual representations of) information and can communicate with objects or agents (maybe other visitors in the same environment). These spaces can for example, represent offices, shared workspaces, shops, class rooms, companies or cities. However, it is also possible to design virtual spaces that are devoted to certain themes and are tuned to users (visitors) interested in that theme or to users (visitors) that not necessarily share common (professional or educational) interests, but share some common conditions (driving a car, being in hospital for some period, have the same therapy, belonging to the same political party, etc.).

As an example we mention a virtual world developed at a cancer research institute in Seattle. This world enables people struggling with cancer to obtain information and interact with others facing similar challenges. Patients, families and friends can enter the three-dimensional world (a rendering of the actual outpatient lobby), get information at a reception desk, visit a virtual gift shop, etc. Each participant obtains an avatar representation. Participants can engage in public chat discussions or invitation-only meetings. A library can be visited, its resources can be used and participants can enter an auditorium to view presentations. Part of the project consists of developing tools to create other applications.

## 3 A VIRTUAL THEATRE COMMUNITY: A CASE STUDY IN PROGRESS

We present research on visualization and interaction in a realistic model of an existing theatre. This existing 'Muziekcentrum' offers its visitors information about performances by means of a yearly brochure. In addition, it is possible to get information at an information desk in the theatre (during office hours), to get information by phone (by talking to a human or by using IVR). The database of the theater holds the information that is available at the beginning of the 'theatre season'. Our aim is to make this information more accessible by using multi-modal accessible multi-media web pages. A more general aim is to do research in the area of web-based services, in particular interactions in virtual environments using speech and language.

Our virtual theatre has been built according to design drawings of the architects of the building. Part has been realized by converting AutoCAD drawings to VRML97. Video recordings have been used to add 'textures' to walls, floors, etc. Sensor nodes in the environment activate animations or start events (entering a dialogue mode, playing music, moving spotlights, etc.). Visitors can explore the environment of the building, hear the carillon of a nearby church, look at a neighboring pub and movie theatre, etc. They can enter the theatre and walk around, admire the paintings on the walls, enter the main performance hall, go to the balconies and, take a seat in order to get a view of the stage from that particular location. Information about today's performances is available on a screen that is automatically updated. Visitors may go to the information desk in the theatre, see previews and start a dialogue with an agent called 'Karen'. Karen has her looks against her. We are working on a more attractive – not necessarily human-like - appearance for Karen. In 1999 we aim at user evaluation studies that will concentrate on questions about, among others, the need of reasonable realistic representations of the theatre information and transaction service interactions that are offered. Another aim that will be explored is the embedding of this particular virtual environment in a virtual cultural arena where people can ask, retrieve and explore information about theatre and music performances in general. Agents, like Karen, will help the users with these tasks.

## 4 AGENTS AND INTERACTIONS

### 4.1 A NAVIGATIONAL AGENT

Clearly, the WWW-based virtual environment we are developing allows navigation input through keyboard and mouse. Such input allows

the user to move and to rotate, to jump from one location to an other, to interact with objects and to trigger them. In addition, a navigation agent has been developed that allows the user to explore the environment and to interact with objects in this environment by means of speech commands. Obviously, we do not want completely separated modalities. It should be left to the user to choose between the interacting means or to use both, sequentially or simultaneously. A smooth integration of the pointing devices and speech in a virtual environment requires means to resolve deictic references that occur in the interaction, and the navigation agent should be able to reason (in a modest way) about the geometry of the world in which it moves. We slowly extend and improve the interaction and navigation intelligence of our present navigation agent. At this moment we are exploring the possibility of speech recognition for several clients on a central server and the advantages of making the navigation agent visible for the user. One of our conclusions is that current web technology hardly allows smooth integration of speech recognition and browsing a virtual world.

### 4.2 AN INFORMATION AND TRANSACTION AGENT

As mentioned before, a second agent called Karen allows a natural language dialogue with the system about performances, artists, dates, prices, etc. Karen wants to sell tickets. She is fed from a database that contains information about performances in some of our local theatres. Developing intelligence for Karen, in this particular environment, is a main aim of our project. Presently the input to Karen is keyboard-driven natural language and the output in our for the general audience WWW accessible virtual world is screen and menu based ([3]). In a prototype system we allow Karen to use a combination of speech synthesis and information presentation on the screen ([4]). Based on the user utterance, the context and the database, the system has to decide on a response action, consisting of database manipulation and dialogue acts. In our experimental system Karen's spoken dialogue contribution is presented by visual speech, that is, a 'talking face' on the screen, embedded in the virtual world, mouths the questions and part of the responses. If necessary, information is given in a window on the screen and the user can click items to get more information. The virtual face that has been designed allows animation of lip and face movements and animation of some simple face expressions ([1]). A Dutch text-to-speech synthesis system has been used to give Karen a voice ([2]). For speech-image synchronization 3D images of visemes are called when corresponding phonemes are spoken. Since our application is web-based, it requires the solution of technical problems dealing with sending and compressing sound files, commands and synchronizing sounds and animations.

### 4.3 FUTURE AGENTS: CONVERSATION, RETRIEVAL, FILTERING

As may have become clear from the previous sections, our approach to designing a virtual environment for an interest community is bottom-up. At this moment the system has two agents with different tasks and with no interactions between them. Moreover, the agents do not employ a model of a user or of user groups. In general, when we talk about interface agents we mean software agents with a user model, that is, a user model programmed in the agent by the user, provided as a knowledge base by a knowledge engineer or obtained and maintained by a learning procedure from the user and customized according to his preferences and habits and to the history of interaction with the system. In this way we have agents that make personalized suggestions (e.g. about articles, performances, etc.) based on social filtering (look at others who seem to have similar preferences) or content filtering (detect patterns, e.g. keywords) of the items that turn out to be of interest to the user. These agents can be passive that wait until they are put into action or they sense changes, take initiative and perform actions, e.g. to inform the user without being asked about new information.

Our first concern in the near future will be the introduction of a conversational agent (which has some general knowledge about well known artists and some well known performances). In this way we have obtained three kinds of dialogues (information & transaction dialogues, command-like dialogues and conversational dialogues). A slight sharing

of knowledge (in particular, preferences of the user) between agents will become possible. In a next step this knowledge of preferences should be exploited, not only in the interactions with the user, but also in designing an agent that retrieves information that matches with the users profile. Visualization of the domains accessible to users may help to guide the interpretation of questions and requests for retrieval (we give a different interpretation to a question about artists when we are in an opera building then when we are in a music hall).

## REFERENCES


[1] M. van den Berk. Visuele Spraaksynthese: Een tot de verbeelding sprekend gezicht. M. Sc. thesis, October 1998.

[2] J. Hulstijn & A. van Hessen. Utterance Generation for Transaction Dialogues. In: Proceedings 5th *International Conf. Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998, to appear.

[3] D. Lie, J. Hulstijn, A. Nijholt, R. op den Akker. A Transformational Approach to NL Understanding in Dialogue Systems. Proceedings *NLP and Industrial Applications*, Moncton, New Brunswick, August 1998, 163-168.

[4] A. Nijholt, A. van Hessen & J. Hulstijn. Speech and Language Interaction in a (Virtual) Cultural Theatre. Proceedings *NLP and Industrial Applications*, Moncton, New Brunswick, August 1998, 176-182.

[5] A. Nijholt, J. Hulstijn & A. van Hessen. Interacties in Virtuele Web-gebaseerde Omgevingen. Proceedings Conference *Informatiewetenschap '98*, (in Dutch) Antwerpen, December 1998, to appear.

# MULINEX: Multilingual Web Search and Navigation

Joanne Capstick, Abdel Kader Diagne, Gregor Erbach, Hans Uszkoreit
German Research Center for Artificial Intelligence - Language Technology Lab
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
http://mulinex.dfki.de/

## ABSTRACT

MULINEX is a fully implemented multilingual search and navigation system for the WWW. The system allows users to search and navigate multilingual document collections using only their native language to formulate, expand and disambiguate queries, navigate the result set and read the retrieved documents. This multilingual functionality is achieved by the use of dictionary-based query translation, multilingual document categorisation and automatic translation of summaries and documents. The system has been installed in the online services of two Internet content and service provider companies.

**Keywords:** translingual information retrieval, categorisation, summarisation, language identification, query translation, machine translation

## 1 INTRODUCTION

The Internet is rapidly changing from an English dominated medium to a multilingual information and communication service. Navigation in this multilingual information space is still far from the ideal scenario – the ability to access, in one's own mother tongue, the mass of multilingual documents over the Internet in a seamless and transparent fashion. The MULINEX consortium aims to address these shortcomings by developing advanced multilingual search and navigation facilities. The consortium consists of users (Bertelsmann Telemedia and Grolier Interactive Europe), and of technology providers in the areas of information retrieval, translation tools and language technology (DATAMAT, TRADOS and DFKI).

## 2 FUNCTIONALITY

MULINEX is a multilingual Internet search engine that supports selective information access, navigation and browsing in a multilingual environment. During the phase of document gathering by the web spider, documents are analysed in order to obtain useful information in addition to the traditional keyword-based indices. The project emphasises a user-friendly interface, which supports the user by presenting search results along with information about language, thematic categories, automatically generated summaries, and allows the user to sort results by multiple criteria. Translingual search is supported by interactive translation of user queries. Commercial machine translation technology (LOGOS) is used to provide translations of foreign-language documents on demand.

The demonstrator offers the following functionalities:

- A document gatherer (web spider) which performs language identification, thematic document classification and document summarisation
- Translation of the user's query
- Simultaneous search in English, German and French document collections
- Automatically generated document summaries
- Restriction of the search by language and categories
- On-demand translation of summaries and search results.
- Registration of users and user preferences for a personalised search environment

## 3 TECHNOLOGIES AND RESOURCES

In this section, we describe the technologies and resources that are used in the components of the MULINEX system.

**Document Acquisition:** The gathering of documents is performed by a modified version of the Harvest gatherer. Harvest has been augmented to call routines for language identification, document classification, and automatic summarisation, and to work in conjunction with the Fulcrum SearchServer, which is used as the information retrieval core engine in our system.

**Language Identification:** Language identification is performed by making use of an algorithm which

compares the relative frequencies of the most frequent n-grams (from 1 to 5 characters) in a document to 40 stored language profiles.

**Categorisation:** Document classification is performed by the k-nearest-neighbour algorithm, a statistical algorithm which classifies a new document by combining the category assignments of the k most similar training documents, weighted by the statistical distance between the new document and each of the k best matching training document. The categoriser has been trained with documents from newsgroups in French, German and English.

In addition, there is a regular-expression-based categorisation algorithm for narrow, specialised categories.

**Summarisation:** Document summarisation is performed by selecting the sentences which best characterise a document. During document gathering, it operates in query-independent mode by selecting sentences on the basis of structural and layout HTML markup, and by position in the document or paragraph.

**Query Formulation / Translation:** The MULINEX system translates and expands the users' queries. Since the retrieval performance of automatically translated queries is inferior to monolingual information retrieval, there is an (optional) step of user interaction, where the user can select terms from the translated query and add his own translation. Queries are morphologically analysed by making use of Morphix and MMORPH, and then translated by making use of multilingual dictionaries. The translated queries are the input to the search in the document collection.

**Information Retrieval:** The search is being performed by the Fulcrum SearchServer, a state-of-the-art information retrieval system, which incorporates linguistic technologies for morphological normalisation of documents and queries. The results are presented along with information about their language, thematic categories and automatically generated summaries. If the result pages are accessible to the system without large delays (e.g., if they reside on the same intranet), a summary which is tailored to the user's query can be produced. Results can be ordered by relevance or by thematic categories.

**Databases:** Two SQL-based database management systems are used in the MULINEX system: Fulcrum Search Server for all information retrieval tasks and for storing category profiles, and a standard SQL database (MSQL) for storing user profiles and the multilingual lexicon.

**Multilingual Lexical Resources:** The MULINEX system uses six bilingual lexicon databases with 100.000 to 200.000 entries each for all six language pairs supported by the system (German-English, German-French, French-English and the converse pairs).

**Machine Translation:** Summaries and result documents can be translated on demand by making use of the LOGOS machine translation system.

## 4 SYSTEM VALIDATION

The system will be made publicly available by the user partners in the consortium, who will obtain feedback from the end users of the system in order to evaluate the appeal and usability of the system.

**Validation Sites:** In the summer of 1998, the system has been installed in the online services of Grolier Interactive Europe and Bertelsmann Telemedia, two large internet service and content providers in France and Germany. They will use it to provide multilingual search facilities for their sites, and to enhance the functionality of their existing search engines. These services will become publicly available in the 4th quarter of 1998.

**Validation Methodology:** The end users of the system will be invited to provide feedback on the usability of the system via questionnaires, in which they evaluate the system, suggest improvements and can provide personal details. Users can also use a mailto-link to give feedback in free form. In addition, there will be in-depth interviews with a selected group of end-users.

## ACKNOWLEDGMENTS

The work reported here was financially supported by the EU's Telematics Application Programme, contract LE-4203 in the sector Language Engineering.

## REFERENCES

[1] Joanne Capstick, Gregor Erbach and Hans Uszkoreit. *Design and Evaluation of a Psychological Experiment on the Effectiveness of Document Summarisation for the Retrieval of Multilingual WWW Documents*. AAAI Spring Symposium on Intelligent Text Summarisation, Stanford, 1998.

[2] Joanne Capstick, Abdel Kader Diagne, Gregor Erbach, Hans Uszkoreit, Francesco Cagno, Giovanni Gadaleta, Juan A. Hernandez, RenéKorte, Anne Leisenberg, Manfred Leisenberg and Oliver Christ: *MULINEX: Multilingual Web Search and Navigation.* Proceedings of Industrial Applications of Natural Language Processing, Moncton, Canada, 1998

# OLIVE: speech based video retrieval

*Klaus Netter*
Language Technology Lab
German Research Center for Artificial Intelligence – DFKI GmbH
Stuhlsatzenhausweg 3, D-66123
Saarbrücken, Germany.
E-mail: netter@dfki.de

*Franciska de Jong*
TNO/University of Twente
address: Department of Computer Science
P.O. Box 217, 7500 AE Enschede
The Netherlands.
E-mail: fdejong@cs.utwente.nl

## ABSTRACT

This paper describes the Olive project which aims to contribute to the needs of video archives by supporting the automated indexing of video material on the basis of human language processing. Olive develops speech recognition to automatically derive transcripts for the sound track, thus generating time coded linguistic elements which are the basis for text-based retrieval functionality.

**Keywords:** language technology, content-based video retrieval, speech recognition

## 1 INTRODUCTION

In archives detailed documentation and profiling of the archived material is a prerequisite for an efficient and precise access to the data. While in the domain of textual digital libraries advanced methods of information retrieval can support such processes, there are so far no effective methods for automatically profiling, indexing, and retrieving image and video material on the basis of a direct analysis of its visual content. Although there have been some advances in the automatic recognition of images, these are still so limited that they will not provide a sufficiently robust basis for effectively profiling large amounts of visual data. Instead Olive uses natural language as the media interlingua, focusing on technology for processing the sound track. It is a follow-up of the Pop-Eye project (http://pop-eye.tros.com/) which takes subtitles as starting point.

## 2 USER NEEDS

The primary users of Olive are two broadcast organisations (ARTE and TROS), as a national audio-video archive (INA) and a large service provider for broadcasting and TV productions (NOB). For all of these institutions archiving of video productions plays an important role, be it for the purpose of re-broadcasting or reselling existing productions, for reusing part of the material in new productions or for generally supporting research in video data bases. In particular, the latter two functions make it very important that the customers of the archives have maximally detailed access to the content of the video material.

Reusing parts of existing material can reduce the production costs considerably and therefore makes it highly desirable that the full and detailed content of a video be documented and accessible without having to view the entire video. This implies that indexes to video's would have to disclose not just the video production as a whole, but also fragments of the material via their timecodes. As generating the necessary content descriptions for large numbers of video shots per production is very costly and labor-intensive, automated indexing is a way to meet the demands of present day multimedia archives.

Olive aims to develop a system which automatically produces indexes from a transcription of the sound track of a programme. In addition the Olive system will provide access to the digitised video material through some intranet or even the internet. As a result user should be able to query a digital video library, browse through the returned descriptions and then download and pre-view the relevant sequences.

## 3 BASELINE TECHNOLOGY

To answer such problems and demands as just described Olive attempts to provide online access to video material on the basis of linguistic material associated with the visual data. The linguistic data associated with a video basically come in two classes. They are either linked to the video time code or not. Among the former are subtitles and of course the spoken word itself. In addition to disclosure technology for the tasks to be performed by any retrieval system, Olive will develop speech recognition for German and French for the automatic generation of timecoded sound track transcriptions. It will also apply translation technology.

### 3.1 SPEECH RECOGNITION

Currently, speech technology is still somewhat limited and does not guarantee completely domain- and speaker-independent reliable recognition. However, it has to be kept in mind, that for the purpose of indexing and retrieval a 100% recognition rate is not absolutely necessary, since not every word will have to make it into the index, and not every expression in the index is likely to be queried. In addition, speech recognition can also be used as a secondary means to support automatic time coding of the second class of data, as for example manual transcriptions. The cleaner and more reliable transcriptions can be used as the basis for indexing. The necessary time-coding can then be derived by automatically aligning the result of speech recognition with such a transcription. Basically the same method can be used if there are production scripts or other types of descriptions reflecting the time line and the spoken word.

### 3.2 TRANSLATION

Following the approach developed within Twenty-One (http://twentyone.tpd.tno.nl/) functionality will be added to support cross-language information retrieval. For example, video's with a German soundtrack will be accessible via queries in any of the languages French, English, Dutch and German.

### 3.3 INHERENT LIMITATIONS

It should be clear, of course, that the discourse and linguistic data associated with a video will not always be a direct reflection of the images and the visual content of the video. In particular, there will be a broad range of variation between more descriptive texts, like documentaries, where the commentary refers to and explains the visual content, and programmes of the drama type, where the dialogue and discourse complements the visual content. Thus, the approach taken in the projects will have some clear limitations, and future experience and evaluation will have to show for what type of programmes the approach is most suitable.

## 4 OTHER PROJECT INFORMATION

The users in the Olive consortium are two television stations, comprising ARTE (Strasbourg, France) and TROS (Hilversum, Netherlands), as well as the French national audio-video archive, INA/Inatheque in Paris, France, and a large service provider for broadcasting and TV productions, viz., NOB in Hilversum, Netherlands.

The system will be implemented through the co-operation of several organisations: TNO-TPD Delft, the project co-ordinator which brings in the core indexing and retrieval functionality, VDA BV Hilversum building the video capturing software, the University of Twente and the LT Lab of DFKI GmbH Saarbrücken, responsible among others for the language technology, the University of Tübingen, carrying out the evaluation in Pop-Eye, CNRS LIMSI and Vecsys SA Paris which are developing and integrating the speech recognition modules, respectively.

Olive (LE4-8364) is funded by the European Commission under the Telematics Application Programme, sector Language Engineering. The project started in 1998 and will last until 2000, More information about Olive can be found under http://twentyone.tpd.tno.nl/olive.

# Twenty-One: a baseline for multilingual multimedia retrieval

*Franciska de Jong*
University of Twente
Department of Computer Science /CTIT
P.O. Box 217, 7500 AE Enschede, The Netherlands.
E-mail: fdejong@cs.utwente.nl

## ABSTRACT

In this paper we will give a short overview of the ideas underpinning the demonstrator developed within the EU-funded project Twenty-One; this system provides for the disclosure of information in a heterogeneous document environment that includes documents of different types and languages. As part of the off-line document processing that has been integrated in the system noun phrases are extracted to build a phrase-based index. They are the starting point for the generation of both a fuzzy phrase index and a translation step that is needed for the realisation of cross-language retrieval functionality.

**Keywords:** language technology, multimedia information retrieval, cross-language information retrieval

## 1 INTRODUCTION

In many environments, such as the World Wide Web, full text retrieval tools appear to be attractive for the searching and accessing of unstructured information content. Twenty-One intends to contribute to the need for more powerful approaches to content disclosure in a number of ways. The project aims to develop a demonstrator system supporting the disclosure of information in a heterogeneous document environment that includes documents of different types and languages.

The technology that has been developed within the project and integrated with useful background components was evaluated within two tasks of the international IR evaluation conference TREC-7. Both in the main task and in the cross-language task, the Twenty-One system performed at the level of today's world leading experimental IR systems. Cf. [8].

The project has resulted in the first on-line text retrieval system in Europe supporting cross-language retrieval (accessible since 1996). It has set a baseline for a series of other EU-funded projects, and has led already to some spin-off applications, such as the retrieval engine supporting the web-site of the Dutch national Millenium Platform[1].

Section will 2 will describe the user perspective, and the system design, section 3 analyses the role of natural language in disclosure of multimedia collections, section 4 addresses the relationship with some other projects, and in section 5 an overview of the current functionality of the demonstrator will be given.

## 2 THE TWENTY-ONE PROJECT

The full name of the project is 'Twenty-One: development of a multimedia dissemination and transaction tool', and the main objective of Twenty-One is to develop domain-independent technology to improve the dissemination level of digitised and non-digitised multimedia information, and to make it more readily and cheaply accessible to a larger group of people. The system can be inspected through the project homepage: `http://twentyone.tpd.tno.nl/` Cf. also [2].

### 2.1 USER-ORIENTATION

The project focus is on the information need in the field of ecology and sustainable development. The project's user group consists o five environmental organisations involved in the publishing of information in this field. Because of the generic characteristics of the distinct software modules they can also be applied outside the domain of environmental information. Information will be disseminated either via Internet or via a periodically distributed CD-ROM (suited for rapid access to static document bases). The Twenty-One information transaction model, also called the Galilei model, forms an important prerequisite for employing the technology developed within the project. This information transaction model triggers different environmental organisations to exchange information.

The nature of the information to be handled by the system varies enormously, both in format, source,

[1] `Cf. http://www.mp2000.nl/`

and content. A lot of material for which a publication need exists can be characterised as 'grey literature', for which few bibliographic details and no electronic source are available. Considerable emphasis in Twenty-One is put on the development of preprocessing modules for the digital conversion of paper documents. Another example of a domain or application specific focus is the indexing of documents on web sites maintained by organisations outside the Twenty-One user group via a web crawler.

## 2.2 SYSTEM DESIGN

Though the primary focus may vary with the kind of users for which the technology is put to use, it can in principle deal with information objects in various different media: paper, word processor texts, pictures, video, and audio. Also information in database-format, either from local or from remote sources, falls within its scope. The language elements in the documents to be disclosed are the basis for the automatic generation of a text based index that enables the kind of functionality commonly known as full text retrieval. This provides users access to information not via a controlled set of search terms, but via any word in the document. It also allows users not only to look for entire documents, but also for information within the documents. This functionality is particularly suited for large collections of *unstructured* data. Two crucial sets of software can be distinguished:

- Software to disclose multimedia information
- Software to retrieve multimedia information (with state-of-the-art browsing applications) from remote or local servers, or from a local CD-ROM.

The core of the retrieval software is based upon proprietary software from TNO-TPD and consists of a search kernel supporting several query modes and interface languages. In Figure 1 it is depicted how the various document types are submitted to a three-stage off-line disclosure process.

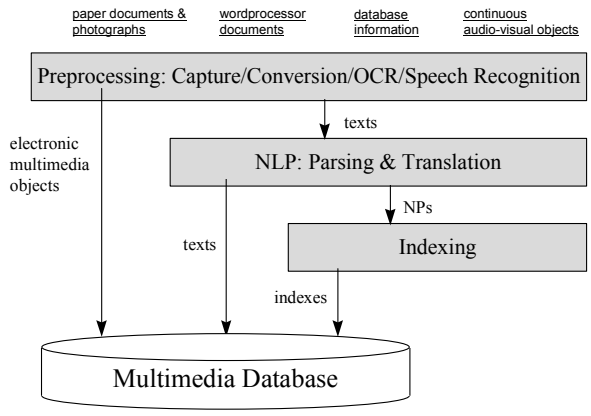



*Figure 1 Multimedia disclosure (off-line)*

### 2.2.1 Disclosure

First there is a preprocessing step in which all objects are converted to a format suitable for storage into a database. This includes the isolation of language material, recognition of characters (in the case of paper input) and formatting (SGML/HTML) of language elements. The information objects to be indexed may be stored in multiple representations. For example, a paper document is stored both in HTML-format and as bitmap. In a second step the language elements are submitted to a series of natural language processing modules. This stage includes morphological analysis and part-of-speech tagging[2], parsing (noun phrase extraction)[3] and translation. The parser output consists of a version of the original document in which the noun phrases (NPs) -which are considered to be potential index terms- have been marked. The role of the translation modules is to facilitate cross-language retrieval. This aspect of the functionality is described in detail in [3] and [4]. The third stage is the building of an index on the basis of the output generated by the parser. The results of all stages are stored into one database. A series of modules can be distinguished for each of the disclosure stages. Together the modules can be viewed as a multimedia indexing and retrieval workbench. Each new application may focus on a specific type of document, and can make a different selection out of the modules from the workbench accordingly. Because of this modular design, the system appeared to be a very useful baseline for a series of applications, covering tools for disclosing both static and continuous data types, such as video and audio.

### 2.2.2 Retrieval

Retrieval relies on language as the medium for indexing and querying and it also exploits language as a means to filter and narrow down in several steps the space of potentially relevant target objects. One of the obvious advantages of this stepwise process is that the downloading of condense data objects such as images and pictures can be postponed until there is confirmed evidence that there is a match with the actual information need.

Searching the index on the basis of a query will support the retrieval of the stored textual representations and (fragments of) the objects linked to the index terms. The automatically acquired text based index is the link between the disclosure and retrieval modules. The index is, unlike most ordinary

[2] The modules for morphological analysis and POS-tagging make use of the Xelda-toolkit developed by Xerox parsing

[3] The parsing modules make use of the NLP-toolkit from TNO-TPD (the Netherlands). Grammars have been developed by DFKI (Germany), University of Twente (the Netherlands) and Xerox (Grenoble, France)

retrieval systems, not limited to an index based on single words or lemmata. In fact it is a combination of several indexes, among which a fuzzy phrase based index, a lemmatized vector space index and a bibliographic index. Using a phrase based index, users are allowed to query the system by using not only simple keywords, but also complete phrases, such as: 'effects of acid rain on forests in the Netherlands'. The matching between query text and index can be done via a one-run fuzzy match that ranks documents on the basis of similarity and number of matching phrases. The incorporation of a vector space index allows a user to improve the initial retrieval results by feeding the most relevant pages back into the retrieval system to get similar documents returned. This mixed approach taken has been proven to yield a considerable improvement in retrieval performance. Recall profits from the morphological analysis (compound splitting) and fuzzy matching, step-wise retrieval with user interaction and relevance feedback improves precision. Cf. also [5, 6, 7].

## 3 MULTIMEDIA & NATURAL LANGUAGE

Though the focus in Twenty-One is on the disclosure of paper documents, from a more programmatical perspective it is meant to set a framework for a wider range of formats. Ideally indexing and retrieval of multimedia objects should be based on technologies for the automated processing and analysis of information content, for example, image feature extraction. However, though bit-wise and pattern-based recognition of sound, pictures, still images, film sequences etc., already is or may soon become feasible, the state-of-the-art in the relevant technological domains allows only very limited automatic interpretation of the objects involved. (For an overview of problems and approaches, see for example the introductory chapter to [1]). Progress in the development of applications not restricted to specific domains is not to be expected in the short run. More advanced multimedia retrieval could only be achieved if long term research efforts are put in the improvement of content analysis. And even then it will be questionable what the appropriate medium for representing and querying such content could be and whether the human language will not remain *the* access and search medium after all. In any case, as indicated above, some of the needs for multimedia information access can already be solved by applying human language technologies in combination with state-of-the art retrieval technology. Evidently, automated indexing of textual objects is supported by a relatively matured technology and fortunately, natural language is often part of the various media. The disclosure of objects that are not purely or primarily textual can therefore benefit from the advances in indexing based on natural language processing. And as projects such as Pop-Eye and Olive show, Twenty-One has offered a very useful baseline for the proof of concept.

## 4 RELATION WITH OTHER PROJECTS

In two other EU-projects a similar approach towards indexing, retrieval and translation is applied, but there the focus is on the disclosure of video material and the preprocessing modules needed to capture the language elements: subtitle capturing (Pop-Eye[4]) and speech recognition (Olive[5]). In DRUID, a project carried within the Telematics Institute (Netherlands), the results from Twenty-One will be the starting point for the development of information filtering techniques and for the application of speech recognition in disclosing digital archives in the Netherlands.

## 5 THE DEMONSTRATOR

The retrieval functionality of the Twenty-One demonstrator can best be described in three steps, that each will be illustrated with a screen shot of a specific part of the interface. We will distinguish:

- querying
- browsing and selection
- intermediate presentation
- presentation of original

We will ignore here the possibility to query the document base with bibliographic keys, but only discuss the so-called 'Normal Query' mode.

As Twenty-One discloses documents and supports querying in four languages (English, German, French, Dutch), the user can select the query language of his preference in the left hand bar of the interface. In Figure 2 the selected query language is German and the query is '*Kompostierung von Haushaltsabfall*'. In the result screen to the right of

[4] Full project name: "Pop-Eye: a multilingual continuous video disclosing tool based on subtitle indexing and partial translation". Pop-Eye is a EU-funded project within the Telematics Application Programme, sector Language Engineering (LE1-4234). Duration: 1997-1998. For further information, cf. the project homepage `http://pop-eye.tros.com/`. Cf. also the contribution by Wim van Bruxvoort in this volume.

[5] Full project name: "Olive: a multilingual indexing tool for broadcast material based on speech recognition". Olive is a EU-funded project within the Telematics Application Programme, sector Language Engineering (LE4-8364). Duration: Spring 1998- Spring 2000. For further information, cf. the project homepage `http://twentyone.tpd.tno.nl/~olive/`.

the query bar a table is presenting the documents that contain this query phrase or a phrase that according to some similarity measure is related to it. The table gives the relevant document, its source (which can be either the Twenty-One database containing electronic versions of paper documents, or documents from remote sites that have been marked as relevant to the application domain), the page in the document containing the matching phrase, the matching phrase itself, and an icon indication the original language of the document.

By default the results are ordered on document relevance (Doc-Score). Optionally a user may ask for ordering on the basis of phrase ranking (Phrase Score).

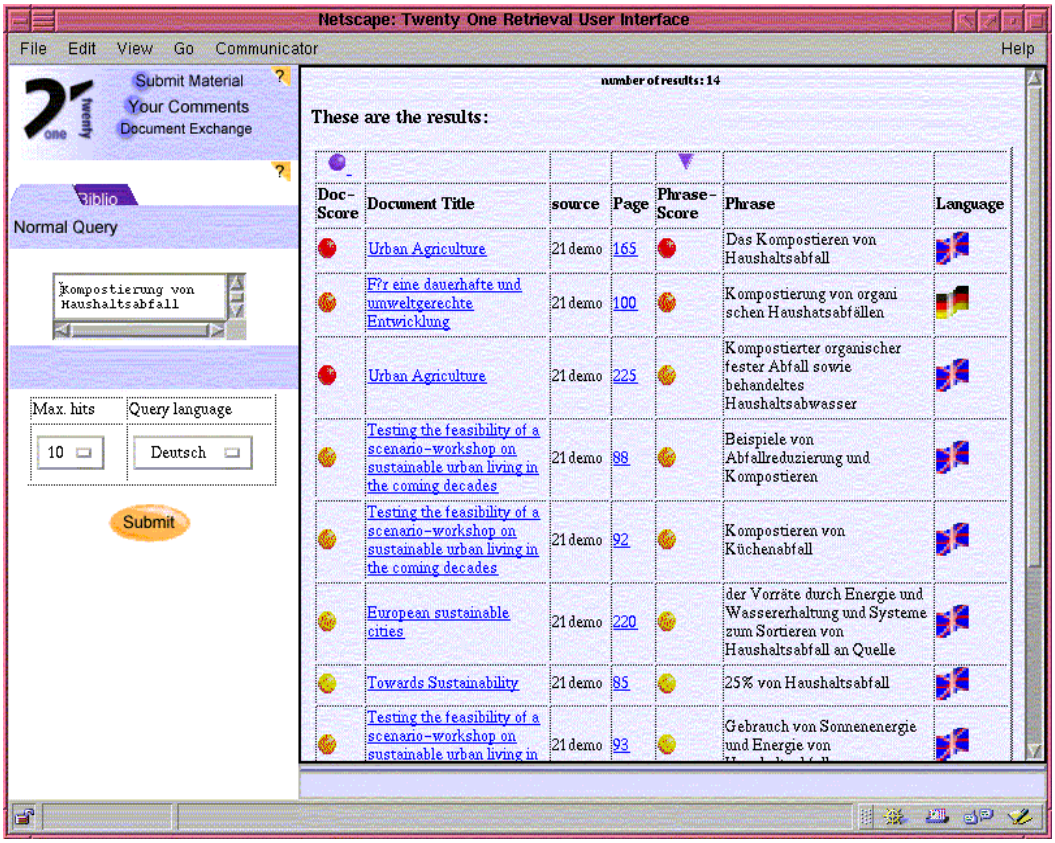



*Figure 2, Query screen*

The results in Figure 2 illustrate that the system offers the user contextual information on the basis of which he can select the phrase that is most likely to match with his information need.

After clicking the Page cell number for the selected document/page, the screen in Figure 3 pops up. It gives the HTML-version of the document at the point where the selected phrase occurs. This HTML-version can be either in the language in which the document was originally submitted to the disclosure modules, or it can be the result of translation. Figure 3 shows the HTML-version of an originally English document. The original text has been translated off-line into German with LOGOS (commercial MT software) and the translation has been stored in the database, together with the source text. The link with the German query could be established because the German translation has been indexed off-line, in the same way as source language documents.

Via the buttons at the bottom of the screen, the user can ask for bibliographic information, for other pages from the same document, for other documents with similar content (via button 'search similar'), for a version in another language (if available), and (via the button in the lower left corner) the user can also ask for a presentation of the document in its original lay-out. See Figure 4.

The latter option allows the user to view the bitmap of the original page, which is of course always in the original language. This part of the functionality is especially useful in cases where the relevant page contains tables or figures than can not be captured properly by the OCR-module. The initial presentation in HTML-format prevents the unnecessary downloading of irrelevant massive objects. It is one of the aspects of the Twenty-One concept that makes the approach suitable for application in multimedia retrieval.

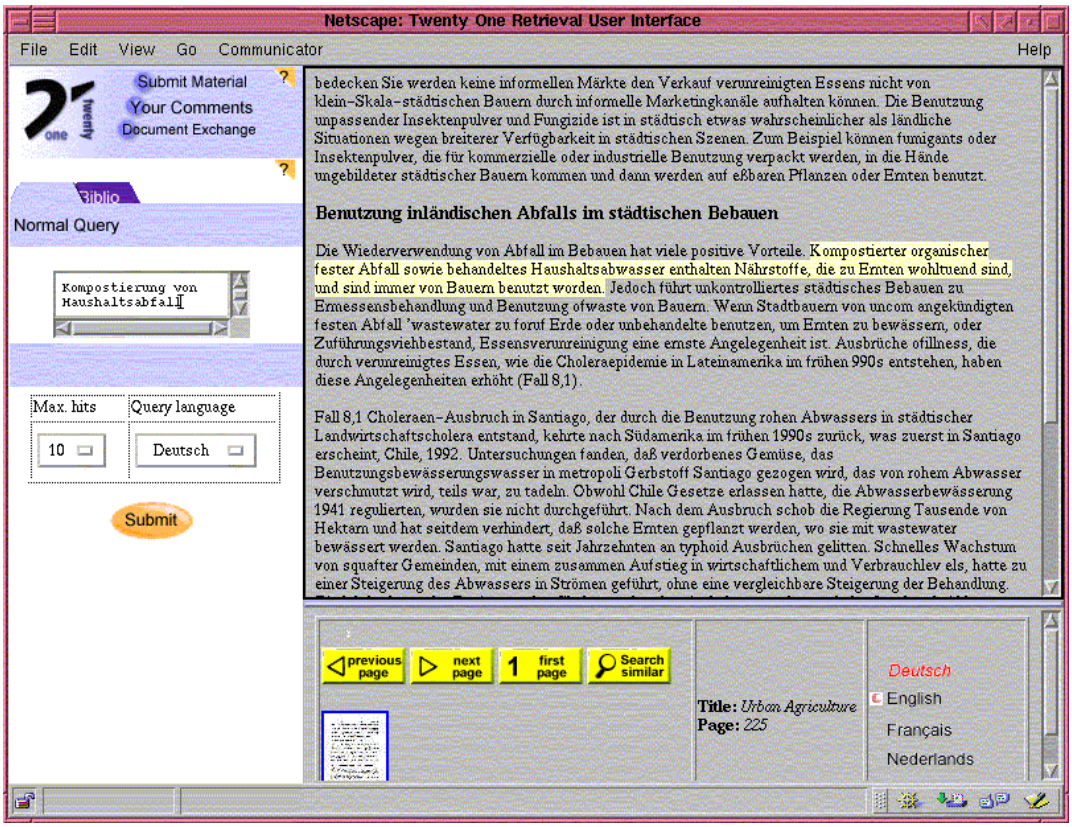



*Figure 3, HTML-presentation*

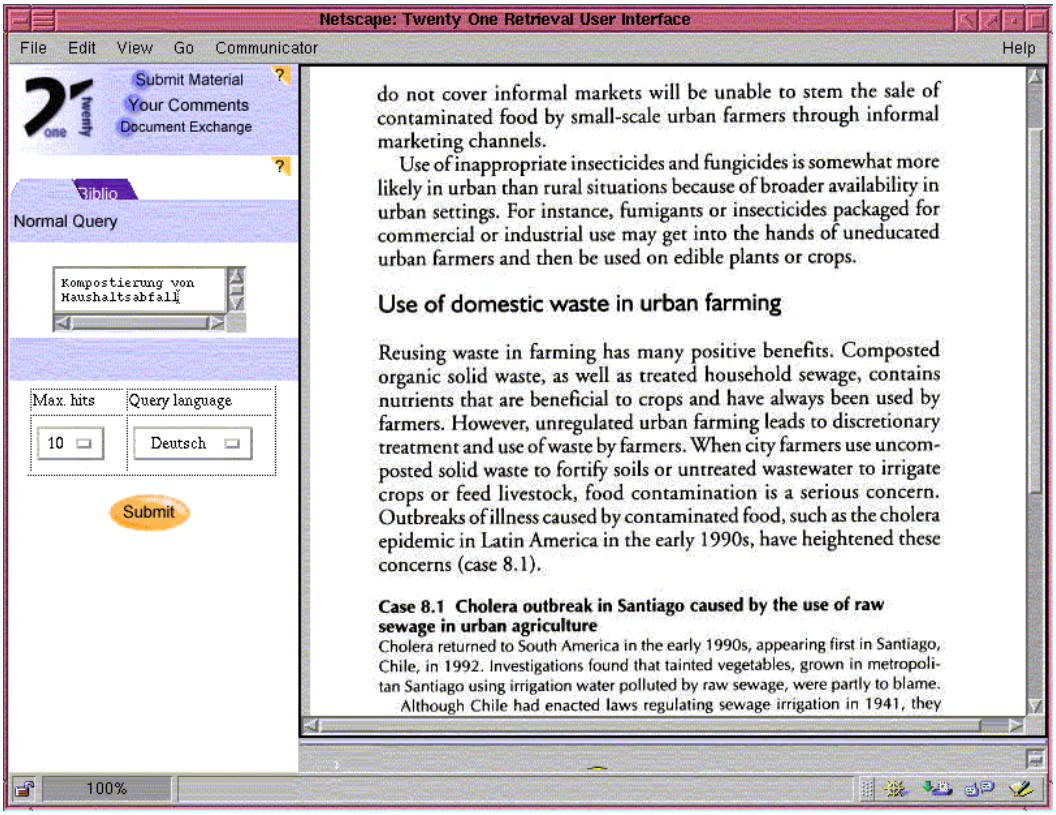



*Figure 4, Bitmap presentation*

# REFERENCES


[1] M. Maybury (ed.), "Intelligent Multimedia Information Retrieval", MIT Press, Cambridge, 1997.

[2] W.G. ter Stal. J-H Beijert, G. de Bruin, J. van Gent, F.M.G. de Jong, W. Kraaij, K. Netter and G. Smart, "Twenty-One: Cross-language disclosure and retrieval of multimedia documents on sustainable development", *Journal of Computer Networks and ISDN Systems* Vol. 30, Elsevier, pp. 1237-1248, 1998.

[3] D. Hiemstra, F.M.G. de Jong and W. Kraaij, "A Domain Specific Lexicon Acquisition Tool for Cross-Language Information Retrieval", *Proceedings of RIAO'97 Montreal*, L. Devroye and C. Chrisment (eds.), pp. 217-232, 1997.

[4] W. Kraaij and D. Hiemstra, "Cross Language Retrieval with the Twenty-One system", *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*. In: Ellen Voorhees and Donna Harman (eds.), Proceedings of the sixth Text Retrieval Conference TREC-6, NIST, Special Publication 500-240, pages 753-761, 1998.

[5] W. Kraaij and R. Pohlmann, "Viewing Stemming as Recall Enhancement", *Proceedings of the 19th ACM-SIGIR Conference on Research and Development in Information Retrieval* (SIGIR96), H.P. Frei, D. Harman, P. Schauble and R. Wilkinson eds., Zürich, pp. 40-48, 1996.

[6] R. Pohlmann and W. Kraaij, "The Effect of Syntactic Phrase Indexing on Retrieval Performance for Dutch Texts", *Proceedings of RIAO'97*, L. Devroye and C. Chrisment (eds.), pp. 176-187, 1997.

[7] W. Kraaij and R. Pohlmann, Comparing the Effect of Syntactic vs. Statistical Phrase Indexing Strategies for Dutch. In: Christos Nicolaou and Constantine Stephanidis, editors, Proceedings of the second European Conference on Research and Advanced Technology for Digital Libraries: ECDL'98, Springer-Verlag, pages 605-614, 1998.

[8] Ellen Voorhees and Donna Harman (eds.), Proceedings of the sixth Text Retrieval Conference TREC-7, NIST, Special Publication, to appear.

# Twente Workshops on Language Technology

The TWLT workshops are organised by the PARLEVINK project of the University of Twente. The first workshop was held in Enschede, the Netherlands on March 22, 1991. The workshop was attended by about 40 participants. The contents of the proceedings are given below.

**Proceedings Twente Workshop on Language Technology 1 (TWLT 1)**
*Tomita's Algorithm: Extensions and Applications*
Eds. R. Heemels, A. Nijholt & K. Sikkel, 103 pages.

Preface and Contents

**A. Nijholt** *(University of Twente, Enschede.)* (Generalised) LR Parsing: From Knuth to Tomita.

**R. Leermakers** *(Philips Research Labs, Eindhoven.)* Recursive Ascent Parsing.

**H. Harkema & M. Tomita** *(University of Twente, Enschede & Carnegie Mellon University, Pittsburgh)* A Parsing Algorithm for Non-Deterministic Context-Sensitive Languages.

**G.J. van der Steen** *(Vleermuis Software Research, Utrecht)* Unrestricted On-Line Parsing and Transduction with Graph Structured Stacks.

**J. Rekers & W. Koorn** *(CWI, Amsterdam & University of Amsterdam, Amsterdam)* Substring Parsing for Arbitrary Context-Free Grammars.

**T. Vosse** *(NICI, Nijmegen.)* Detection and Correction of Morpho-Syntactic Errors in Shift-Reduce Parsing.

**R. Heemels** *(Océ Nederland, Venlo)* Tomita's Algorithm in Practical Applications.

**M. Lankhorst** *(University of Twente, Enschede)* An Empirical Comparison of Generalised LR Tables.

**K. Sikkel** *(University of Twente, Enschede)* Bottom-Up Parallellization of Tomita's Algorithm.

The second workshop in the series (TWLT 2) has been held on November 20, 1991. The workshop was attended by more than 70 researchers from industry and university. The contents of the proceedings are given below.

**Proceedings Twente Workshop on Language Technology 2 (TWLT 2)**
*Linguistic Engineering: Tools and Products.*
Eds. H.J. op den Akker, A. Nijholt & W. ter Stal, 115 pages.

Preface and Contents

**A. Nijholt** *(University of Twente, Enschede)* Linguistic Engineering: A Survey.

**B. van Bakel** *(University of Nijmegen, Nijmegen)* Semantic Analysis of Chemical Texts.

**G.J. van der Steen & A.J. Dijenborgh** *(Vleermuis Software Research, Utrecht)* Lingware: The Translation Tools of the Future.

**T. Vosse** *(NICI, Nijmegen)* Detecting and Correcting Morpho-syntactic Errors in Real Texts.

**C. Barkey** *(TNO/ITI, Delft)* Indexing Large Quantities of Documents Using Computational Linguistics.

**A. van Rijn** *(CIAD/Delft University of Technology, Delft)* A Natural Language Interface for a Flexible Assembly Cell.

**J. Honig** *(Delft University of Technology, Delft)* Using Deltra in Natural Language Front-ends.
**J. Odijk** *(Philips Research Labs, Eindhoven)* The Automatic Translation System ROSETTA3.
**D. van den Akker** *(IBM Research, Amsterdam)* Language Technology at IBM Nederland.
**M.-J. Nederhof, C.H.A. Koster, C. Dekkers & A. van Zwol** *(University of Nijmegen, Nijmegen)* The Grammar Workbench: A First Step Toward Lingware Engineering.

---

The third workshop in the series (TWLT 3) was held on May 12 and 13, 1992. Contrary to the previous workshops it had an international character with eighty participants from the U.S.A., India, Great Britain, Ireland, Italy, Germany, France, Belgium and the Netherlands. The proceedings were available at the workshop. The contents of the proceedings are given below.

---

**Proceedings Twente Workshop on Language Technology 3 (TWLT 3)**
*Connectionism and Natural Language Processing.*
Eds. M.F.J. Drossaers & A. Nijholt, 142 pages.

Preface and Contents
**L.P.J. Veelenturf** *(University of Twente, Enschede)* Representation of Spoken Words in a Self-Organising Neural Net.
**P. Wittenburg & U. H. Frauenfelder** *(Max-Planck Institute, Nijmegen)* Modelling the Human Mental Lexicon with Self-Organising Feature Maps.
**A.J.M.M. Weijters & J. Thole** *(University of Limburg, Maastricht)* Speech Synthesis with Artificial Neural Networks.
**W. Daelemans & A. van den Bosch** *(Tilburg University, Tilburg)* Generalisation Performance of Back Propagation Learning on a Syllabification Task.
**E.-J. van der Linden & W. Kraaij** *(Tilburg University, Tilburg)* Representation of Idioms in Connectionist Models.
**J.C. Scholtes** *(University of Amsterdam, Amsterdam)* Neural Data Oriented Parsing.
**E.F. Tjong Kim Sang** *(University of Groningen, Groningen)* A connectionist Representation for Phrase Structures.
**M.F.J. Drossaers** *(University of Twente, Enschede)* Hopfield Models as Neural-Network Acceptors.
**P. Wyard** *(British Telecom, Ipswich)* A Single Layer Higher Order Neural Net and its Application to Grammar Recognition.
**N.E. Sharkey & A.J.C. Sharkey** *(University of Exeter, Exeter)* A Modular Design for Connectionist Parsing.
**R. Reilly** *(University College, Dublin)* An Exploration of Clause Boundary Effects in SRN Representations.
**S.M. Lucas** *(University of Essex, Colchester)* Syntactic Neural Networks for Natural Language Processing.
**R. Miikkulainen** *(University of Texas, Austin)* DISCERN: A Distributed Neural Network Model of Script Processing and Memory.

---

The fourth workshop in the series has been held on September 23, 1992. The theme of this workshop was ”Pragmatics in Language Technology”. Its aim was to bring together the several approaches to this subject: philosophical, linguistic and logic. The workshop was visited by more

than 50 researchers in these fields, together with several computer scientists. The contents of the proceedings are given below.

---

**Proceedings Twente Workshop on Language Technology 4 (TWLT 4)**

*Pragmatics in Language Technology*

Eds. D. Nauta, A. Nijholt & J. Schaake, 114 pages.



---

The fifth workshop in the series took place on 3 and 4 June 1993. It was devoted to the topic "Natural Language Interfaces". The aim was to provide an international platform for commerce, technology and science to present the advances and current state of the art in this area of research.

---

**Proceedings Twente Workshop on Language Technology 5 (TWLT 5)**

*Natural Language Interfaces*

Eds. F.M.G. de Jong & A. Nijholt, 124 pages.

**D. Speelman** *(University of Leuven)* A Natural Language Interface that Uses Generalised Quantifiers.
**R.-J. Beun** *(IPO, Eindhoven)* The DENK Program: Modeling Pragmatics in Natural Language Interfaces.
**W. Menzel** *(University of Hamburg)* ASL: Architectures for Speech and Language Processing
**C. Huls & E. Bos** *(NICI, Nijmegen)* EDWARD: A Multimodal Interface.
**G. Neumann** *(University of Saarbrücken)* Design Principles of the DISCO system.
**O. Stock & C. Strapparava** *(IRST, Trento)* NL-Based Interaction in a Multimodal Environment.

---

The sixth workshop in the series took place on 16 and 17 December 1993. It was devoted to the topic "Natural Language Parsing". The aim was to provide an international platform for technology and science to present the advances and current state of the art in this area of research, in particular research that aims at analysing real-world text and real-world speech and keyboard input.

---

**Proceedings Twente Workshop on Language Technology 6 (TWLT 6)**
*Natural Language Parsing: Methods and Formalisms*
Eds. K. Sikkel & A. Nijholt, 190 pages.

Preface and Contents
**A. Nijholt** *(University of Twente)* Natural Language Parsing: An Introduction.
**V. Manca** *(University of Pisa)* Typology and Logical Structure of Natural Languages.
**R. Bod** *(University of Amsterdam)* Data Oriented Parsing as a General Framework for Stochastic Language Processing.
**M. Stefanova & W. ter Stal** *(University of Sofia / University of Twente)* A Comparison of ALE and PATR: Practical Experiences.
**J.P.M. de Vreught** *(University of Delft)* A Practical Comparison between Parallel Tabular Recognizers.
**M. Verlinden** *(University of Twente)* Head-Corner Parsing of Unification Grammars: A Case Study.
**M.-J. Nederhof** *(University of Nijmegen)* A Multi-Disciplinary Approach to a Parsing Algorithm.
**Th. Stürmer** *(University of Saarbrücken)* Semantic-Oriented Chart Parsing with Defaults.
**G. Satta** *(University of Venice)* The Parsing Problem for Tree-Adjoining Grammars.
**F. Barthélemy** *(University of Lisbon)* A Single Formalism for a Wide Range of Parsers for DCGs.
**E. Csuhaj-Varjú & R. Abo-Alez** *(Hungarian Academy of Sciences, Budapest)* Multi-Agent Systems in Natural Language Processing.
**C. Cremers** *(University of Leiden)* Coordination as a Parsing Problem.
**M. Wirén** *(University of Saarbrücken)* Bounded Incremental Parsing.
**V. Kubon & M. Platek** *(Charles University, Prague)* Robust Parsing and Grammar Checking of Free Word Order Languages.
**V. Srinivasan** *(University of Mainz)* Punctuation and Parsing of Real-World Texts.
**T.G. Vosse** *(University of Leiden)* Robust GLR Parsing for Grammar-Based Spelling Correction.

---

The seventh workshop in the series took place on 15 and 16 june 1994. It was devoted to the topic "Computer-Assisted Language Learning" (CALL). The aim was to present both the state of the art in CALL and the new perspectives in the research and development of software that is meant to be used in a language curriculum. By the mix of themes addressed in the papers

and demonstrations, we hoped to bring about the exchange of ideas between people of various backgrounds.

---

**Proceedings Twente Workshop on Language Technology 7 (TWLT 7)**
*Computer-Assisted Language Learning*
Eds. L. Appelo, F.M.G. de Jong, 133 pages.

Preface and Contents

**L.Appelo, F.M.G. de Jong** *(IPO / University of Twente)* Computer-Assisted Language Learning: Prolegomena

**M. van Bodegom** *(Eurolinguist Language House, Nijmegen, The Netherlands)* Eurolinguist test: An adaptive testing system.

**B. Cartigny** *(Escape, Tilburg, The Netherlands)* Discatex CD-ROM XA.

**H.Altay Guvenir, K. Oflazer** *(Bilkent University, Ankara)* Using a Corpus for Teaching Turkish Morphology.

**H. Hamburger** *(GMU, Washington, USA)* Viewpoint Abstraction: a Key to Conversational Learning.

**J. Jaspers, G. Kanselaar, W. Kok** *(University of Utrecht, The Netherlands)* Learning English with It's English.

**G. Kempen, A. Dijkstra** *(University of Leiden, The Netherlands)* Towards an integrated system for spelling, grammar and writing instruction.

**F. Kronenberg, A. Krueger, P. Ludewig** *(University of Osnabruek, Germany)* Contextual vocabulary learning with CAVOL.

**S. Lobbe** *(Rotterdam Polytechnic Informatica Centrum, The Netherlands)* Teachers, Students and IT: how to get teachers to integrate IT into the (language) curriculum.

**J. Rous, L. Appelo** *(Institute for Perception Research, Eindhoven, The Netherlands)* APPEAL: Interactive language learning in a multimedia environment.

**B. Salverda** *(SLO, Enschede, The Netherlands)* Developing a Multimedia Course for Learning Dutch as a Second Language.

**C. Schwind** *(Universite de Marseille, France)* Error analysis and explanation in knowledge based language tutoring.

**J. Thompson** *(CTI, Hull, United Kingdom/EUROCALL)* TELL into the mainstream curriculum.

**M. Zock** *(Limsi, Paris, France)* Language in action, or learning a language by whatching how it works.

---

The eighth workshop in the series took place on 1 and 2 December 1994. It was devoted to speech, the integration of speech and natural language processing, and the application of this integration in natural language interfaces. The program emphasized research of interest for the themes in the framework of the Dutch NWO programme on Speech and Natural Language that started in 1994.

---

**Proceedings Twente Workshop on Language Technology 8 (TWLT 8)**
*Speech and Language Engineering*
Eds. L. Boves & A. Nijholt, 176 pages.

Preface and Contents

---

The ninth workshop in the series took place on 9 June 1995. It was devoted to empirical methods in the analysis of dialogues, and the use of corpora of dialogues in building dialogue systems. The aim was to discuss the methods of corpus analysis, as well as results of corpus analysis and the application of such results.

---

**Proceedings Twente Workshop on Language Technology 9 (TWLT 9)**

*Corpus-based Approaches to Dialogue Modelling*

Eds. J.A. Andernach, S.P. van de Burgt & G.F. van der Hoeven, 124 pages.

**J.H. Connolly, A.A. Clarke, S.W. Garner & H.K. Palmén** *(Loughborough University of Technology, UK)* Clause-internal structure in spoken dialogue.

**J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon & A. Anderson** *(HCRC, Edinburgh, UK)* The coding of dialogue structure in a corpus.

**J. Alexandersson & N. Reithinger** *(DFKI, Saarbrücken, Germany)* Designing the dialogue component in a speech translation system - a corpus-based approach.

**H. Aust & M. Oerder** *(Philips, Aachen, Germany)* Dialogue control in automatic inquiry systems.

**M. Rats** *(ITK, Tilburg, the Netherlands)* Referring to topics - a corpus-based study.

**H. Dybkjær, L. Dybkjær & N.O. Bernsen** *(Centre for Cognitive Science, Roskilde, Denmark)* Design, formalization and evaluation of spoken language dialogue.

**D.G. Novick & B. Hansen** *(Oregon Graduate Institute of Science and Technology, Portland, USA)* Mutuality strategies for reference in task-oriented dialogue.

**N. Fraser** *(Vocalis Ltd, Cambridge, UK)* Messy data, what can we learn from it?

**J.A. Andernach** *(University of Twente, Enschede, the Netherlands)* Predicting and interpreting speech acts in a theatre information and booking system.

---

The tenth workshop in the series took place on 6-8 December 1995. This workshop was organized in the framework provided by the Algebraic Methodology and Software Technology movement (AMAST). It focussed on algebraic methods in formal languages, programming langauges and natural languages. Its aim was to bring together those researchers on formal language theory, programming language theory and natural language description theory, that have a common interest in the use of algebraic methods to describe syntactic, semantic and pragmatic properties of language.

---

**Proceedings Twente Workshop on Language Technology 10 (TWLT 10)**

*Algebraic Methods in Language Processing*

Eds. A. Nijholt, G. Scollo & R. Steetskamp, 263 pages.

**Vladimir A. Fomichov** *(Moscow, R)* A Variant of a Universal Metagrammar of Conceptual Structures. Algebraic Systems of Conceptual Syntax.

**Theo M.V. Jansen** *(Amsterdam, NL)* The Method of ROSETTA, Natural Language Translation Using Algebras.

**C.M. Martín-Vide, J. Miquel-Verges & Gh. Paun** *(Tarragona, E)* Contextual Grammars with Depth-First Derivation.

**Pál Dömösi & Jürgen Duske** *(Kussuth University H, University of Hannover, G)* Subword Membership Problem for Linear Indexed Languages.

**C. Rico Perez & J.S. Granda** *(Madrid, E)* Algebraic Methods for Anaphora Resolution.

**Vincenzo Manca** *(Pisa, I)* A Logical Formalism for Intergrammatical Representation.

---

The eleventh workshop in the series took place on 19-21 June 1996. It focussed on the task of dialogue management in natural-language processing systems. The aim was to discuss advances in dialogue management strategies and design methods. During the workshop, there was a separate session concerned with evalutation methods.

---

**Proceedings Twente Workshop on Language Technology 11 (TWLT 11)**

*Dialogue Management in Natural Language Systems*

Eds. S. LuperFoy, A. Nijholt and G. Veldhuijzen van Zanten, 228 pages.

Preface and Contents

**David R. Traum** *(Université de Genève, CH)* Conversational Agency: The TRAINS-93 Dialogue Manager.

**Scott McGlashan** *(SICS, SW)* Towards Multimodal Dialogue Management.

**Pierre Nugues, Christophe Godéreaux, Pierre-Olivier and Frédéric Revolta** *(GREYC, F)* A Conversational Agent to Navigate in Virtual Worlds.

**Anne Vilnat** *(LIMSI-CNRS, F)* Which Processes to Manage Human-Machine Dialogue?

**Susann LuperFoy** *(MITRE, USA)* Tutoring versus Training: A Mediating Dialogue Manager for Spoken Language Systems.

**David G. Novick & Stephen Sutton** *(Portland, USA)* Building on Experience: Managing Spoken Interaction through Library Subdialogues.

**Latifa Taleb** *(INRIA, F)* Communicational Deviation in Finalized InformativeDialogue Management.

**Robbert-Jan Beun** *(IPO, NL)* Speech Act Generation in Cooperative Dialogue.

**Gert Veldhuijzen van Zanten** *(IPO, NL)* Pragmatic Interpretation and Dialogue Management in Spoken-Language Systems.

**Joris Hulstijn, René Steetskamp, Hugo ter Doest, Anton Nijholt & Stan van de Burgt** *(University of Twente, NL & KPN Research, NL)* Topics in SCHISMA Dialogues.

**Gavin Churcher, Clive Souter & Eric S. Atwell** *(Leeds University, UK)* Dialogues in Air Traffic Control

**Elisabeth Maier** *(DFKI, D)* Context Construction as Subtask of Dialogue Processing – the VERBMOBIL Case.

**Anders Baekgaard** *(CPK, DK)* Dialogue Management in a Generic Dialogue System.

**Wayne Ward** *(Carnegie Mellon University, USA)* Dialog Management in the CMU Spoken Language Systems Toolkit.

**Wieland Eckert** *(University of Erlangen, D)* Understanding of Spontaneous Utterances in Human-Machine-Dialog.

**Jan Alexandersson** *(DFKI, D)* Some Ideas for the Automatic Acquisition of Dialogue Structure.

**Kristiina Jokinen** *(Nara Institute of Science and Technology, JP)* Cooperative Response Planning in CDM: Reasoning about Communicative Strategies.

**Elizabeth Hinkelman** *(Kurzweil Applied Science, USA)* Dialogue Grounding for Speech Recognition Systems.

**Jennifer Chu-Carroll** *(University of Delaware, USA)* Response Generation in Collaborative Dialogue Interactions.

**Harry Bunt** *(Tilburg University, NL)* Interaction Management Functions and Context Representation Requirements.

**Peter Wyard & Sandra Williams** *(BT, GB)* Dialogue Management in a Mixed-Initiative, Cooperative, Spoken Language System.

**Rolf Carlson** *(KTH, SW)* The Dialog Component in the Waxholm System.

**Laila Dybkjær, Niels Ole Bernsen & Hans Dybkjaer** *(Roskilde University, DK)* Evaluation of Spoken Dialogue Systems.

**Vincenzo Manca** *(Pisa, I)* A Logical Formalism for Intergrammatical Representation.

---

TWLT 12 took place on 11-14 September 1996. It focussed on 'computational humor' and in particular on verbal humor. TWLT12 consisted of a symposium (Marvin Minsky, Douglas Hofstadter, John Allen Paulos, Hugo Brandt Corstius, Oliviero Stock and Gerrit Krol as main speakers), an essay contest for computer science students, two panels, a seminar organized by Salvatore Attardo and Wladyslaw Chlopicki and a two-day workshop (Automatic interpretation and Generation of Verbal Humor) with a mix of invited papers and papers obtained from a Call for Papers.

---

**Proceedings Twente Workshop on Language Technology 12 (TWLT 12)**

*Computational Humor: Automatic Interpretation and Generation of Verbal Humor*

Eds. J. Hulstijn and A. Nijholt, 208 pages.

Preface and Contents

**Oliviero Stock** 'Password Swordfish': Verbal Humor in the Interface.

**Victor Raskin** Computer Implementation of the General Theory of Verbal Humor.

**Akira Ito & Osamu Takizawa** Why do People use Irony? - The Pragmatics of Irony Usage.

**Akira Utsumi.** Implicit Display Theory of Verbal Irony: Towards a Computational Model of Irony.

**Osamu Takizawa, Masuzo Yanagida, Akira Ito & Hitoshi Isahara** On Computational Processing of Rhetorical Expressions - Puns, Ironies and Tautologies.

**Carmen Curcó** Relevance Theory and Humorous Interpretations.

**Ephraim Nissan** From ALIBI to COLOMBUS. The Long March to Self-aware Computational Models of Humor.

**Salvatore Attardo** Humor Theory beyond Jokes: The Treatment of Humorous Texts at Large.

**Bruce Katz** A Neural Invariant of Humour.

**E. Judith Weiner** Why is a Riddle not Like a Metaphor?

**Tone Veale** No Laughing Matter: The Cognitive Structure of Humour, Metaphor and Creativity.

---

TWLT 13 took place on 13-15 May 1998. It was the follow-up of the Mundial workshop, that took place in München in 1997. Both the Mundial workshop as TWLT13 focussed on the formal semantics and pragmatics of dialogues. In addition to the three-day workshop in Twente, with invited and accepted papers, on 18 May a workshop titled 'Communication and Attitudes' was organized at ILLC/University of Amsterdam.

---

**Proceedings Twente Workshop on Language Technology 13 (TWLT 13)**
*Formal Semantics and Pragmatics of Dialogue (Twendial'98)*
Eds. J. Hulstijn and A. Nijholt, 274 pages.