Bärbel Ripplinger

Linguistic Knowledge in Cross-Language Information Retrieval

Herbert Utz Verlag · Wissenschaft
München

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Ein Titeldatensatz für diese Publikation ist
bei Der Deutschen Bibliothek erhältlich

Zugleich: Dissertation, Saarbrücken, Univ., 2002

ISBN 3-8316-0181-X

Printed in Germany

Herbert Utz Verlag GmbH, München
Tel.: 089/277791-00 – Fax: 089/277791-01

Contents

Acknowledgements **v**

Zusammenfassung **vi**

Abstract **xi**

1 Introduction **1**
1.1 Information Systems 1
1.2 Content Overview 3

2 Basic Concepts of Information Retrieval **7**
2.1 Content Analysis and Presentation 7
2.2 Retrieval Methods 9
2.3 Performance Measurements 11

3 NLP in IR **15**
3.1 Linguistic Phenomena in Text Retrieval 15
3.1.1 Inflectional and Derivational Morphology – Word Formation 16
3.1.2 Compounding 17
3.1.3 Syntactic Processing 18
3.1.4 Semantics 19
3.1.4.1 Synonymy 20
3.1.4.2 Lexical Ambiguity 21
3.2 NLP Techniques Applied to IR Tasks 23
3.2.1 Stemming and Compounding 23
3.2.2 Phrase Indexing 27
3.2.3 Concept-Based Indexing 28
3.2.4 Query Expansion 30

4 Cross-Language Information Retrieval **33**
4.1 The Translation Problem 34
4.2 Current Approaches to CLIR 38
4.2.1 Knowledge-Based Systems 39

4.2.2 Approaches Using MT Systems . . . 41
4.2.3 Corpus-Based Systems . . . 42
4.2.4 Hybrid Systems . . . 43

5 The CLIR Component Mpro-IR 45
5.1 Motivation . . . 45
5.2 Linguistic Processing . . . 47
5.2.1 Morpho-Syntactic Analysis . . . 47
5.2.1.1 Word Form Identification . . . 48
5.2.1.2 Lemmatisation and Tagging . . . 50
5.2.2 Homograph Disambiguation . . . 58
5.2.3 Shallow Parsing . . . 60
5.3 Indexing . . . 63
5.4 Query Translation . . . 67
5.5 Query Expansion . . . 71
5.6 The Retrieval Algorithm . . . 73
5.6.1 German Retrieval . . . 74
5.6.2 English and French Retrieval . . . 87
5.6.3 Phrase Search . . . 93
5.6.4 Ranking and Merging . . . 99

6 Mpro-IR in a Legal Domain 101
6.1 Related Work . . . 102
6.2 The Multilingual Information System EMIS . . . 103
6.2.1 Document Collection . . . 104
6.2.2 Pre-defined Access Facilities . . . 105
6.2.2.1 The Systematic Structure . . . 105
6.2.2.2 The Thematic Structure . . . 106
6.2.2.3 The Keyword Search . . . 107
6.2.3 Free Text Retrieval . . . 108
6.2.3.1 Mpro-IR in EMIS . . . 108
6.2.3.2 Boolean Search . . . 112
6.2.4 Display of a Retrieved Document . . . 115
6.2.5 Technical Data . . . 116

7 Evaluation and Discussion 119
7.1 Mpro-IR in a Legal Domain . . . 120
7.1.1 Multilingual Run . . . 121
7.1.2 Monolingual Runs . . . 124
7.1.2.1 German Run . . . 125
7.1.2.2 English Run . . . 129
7.1.2.3 French Run . . . 130
7.2 Mpro-IR in clef 2000 . . . 132

7.3 Conclusion 136

8 Future Work **139**

9 Summary **145**

Annex **147**
Annex A: Mpro-IR Run 147
Annex B: CLEF 2000 Topics 150

Bibliography **152**

Chapter 1

Introduction

1.1 Information Systems

With the growing use of the Internet in business and private, there is an increasing need for large scale data processing to make sense of the information we receive. In such an environment, information systems (IS) take over a special importance. In the information age, information systems do not change their traditional purpose, i.e. to allow the users to retrieve information about certain subjects such as bibliographic information, patents, economic data (stock information), news etc. but they are changing in the size of the data archive and the time to prepare these data for the users. The World Wide Web (WWW) provides an enormous amount of data which are changing very fast. For an information provider this means all these data have to be prepared on a certain level of actuality to be of an economical benefit. This does not allow to process the data manually. The more actuality is important the more important become the use of online information systems maintained in short time periods.

Speaking of an information system, at a first place the organisational structures and structural relations of the professional information gathering are meant. These depend on sources and contents as well as on the application scope of the information to be deployed, and its function. In general, information systems can be classified into *bibliographic information systems* which are document related, and *factual information systems* which can be further classified according to their application area into *management information systems, database information systems, decision support systems*, and *question and answer (Q&A) systems*[1]. The task of bibliographic systems is to evaluate documents by indexing, to store as well to provide them on request. Such systems are frequently used in libraries, by publishers to promote their products and by bibliographic database producers who provide abstracts of scientific and technical articles to their customers.

[1] For a more detailed introduction, see [Buder et al. 90].

The performance of these systems mostly depends on the correctness and completeness of the related tasks. This comprises a consistent document processing (thesaurus or term list for indexing) as well as a fast and reliable access facility for the users by a fixed input form where the user can provide various data types such as author, title, keywords, subject field, etc. to find relevant documents. Information systems related to factual data stored in databases process for instance statistics of public services, or the accounting data of enterprises. The task of a factual system is to provide the user directly with the requested data. This type often uses a database system which allows retrieval of only a certain type of facts.

The two types of information systems can be differentiated by the following characteristics: Bibliographic systems have to process unformatted data of a variable length (texts), in factual systems most data are of a fixed length and format. Thus, to describe the set of data, only one unit is required, but various units (or database entities) for unformatted data. Additionally, access to bibliographic data is static, whereas retrieval of certain facts can be dynamic. In the first case, all elements stored in the systems could be queried, in the second, only a certain set of search terms are allowed. In factual databases the information itself is changing, whereas in bibliographic systems the information stored continually increases. To guarantee an efficient use of information, which is difficult to reach due to the fact that the volume of information expands unevenly for different topics, one major requirement is the physical and logical organisation of the information (should be multidimensional). On the other hand, to get the right information presupposes a precise query, i.e. to get only items that match the information request exactly. While former users of information systems were almost experts and thus capable of formulating a precise query using a Boolean expression, the growing online data sources and publicly available information systems demand new access models. The best utility for users to formulate a query is their own language. However, to access information in a natural language setting is more difficult because the system has first to analyse the user's query, extract the correct items to query the stored knowledge and then assemble a suitable answer. Nevertheless, in Q&A systems the use of an interface which allows natural language queries is most common. In the other types of information systems (see above) only first attempts have been made to facilitate the access by allowing complete sentences as queries instead of just single terms, or terms combined by Boolean operators.

Systems allowing access to textual data are called **information retrieval systems**. Even if they are not limited to textual data, this is their most prominent application area. The aim of such systems is to retrieve documents relevant to a query which can be a single word, a phrase, a whole sentence, or a narrative. To do this, the query is matched against an index built over the document collection. Because the author of the documents and of the query are not the same,

there is usually a difference between the terms used in the query and those in the documents, the so-called *paraphrasing problem*. Thus, the major task of an IR system is to overcome this diversion.

With growing globalisation, the need for multilinguality also in information systems increases, due to the expense of translation in both time and cost. In an environment where actual information is absolutely essential, only on-the-fly translations are applicable provided by automatic machine translation (MT) systems (human translations are often time consuming and expensive). However, the quality of such systems is a rather drawback. Where only applied to special domains, the translations are of an acceptable quality but where unrestricted text has to be processed, an incorrect translation can negatively impact the performance by retrieving the wrong documents. To overcome such deficiencies, recent developments of information systems incorporate not only information retrieval components for different languages, so-called multilingual information retrieval but a **cross-language information retrieval** module which allows the users to formulate their queries in their preferred language and to retrieve documents written also in foreign language. Nevertheless this presupposes at least a passive knowledge of the foreign languages by the users. On the other hand, the document retrieved can be translated by an MT system with an acceptable quality by providing the system with information about the domain, i.e. whether the document is about economics, medicine, or telecommunications, etc. realised for instance via an interactive user interface.
The development of such a cross-language information retrieval component is the subject of this work.

1.2 Content Overview

The focus of this work is on the description of the cross-language component MPRO-IR which takes advantage of linguistic processing in all tasks of the retrieval process, and makes this work unique. Even when retrieval is, in the first instance, related to text documents, the use of natural language processing techniques are not very common. Current retrieval systems exploit methods developed in Computational Linguistics in most cases for indexing purposes, i.e. morphological analysis for stemming, and syntactic parsers for phrase (in most cases nominal phrase) identification. Nevertheless, the usefulness of these methods is still to be proven. In cross-language retrieval, methods developed within Machine Translation, are used for the translation task, and thus linguistic algorithms are more common.

Chapter 2 gives an overview of the main concepts of information retrieval such as content analysis including indexing, and methods for the retrieval process.

Performance is essential for information retrieval, so a brief overview about the measurements and the methods to improve the performance are presented.

Chapter 3 shows the relations between text retrieval and natural language processing. First, those linguistic phenomena which may effect information retrieval are briefly described. Afterwards, the various attempts to take advantage of natural language processing (NLP) techniques developed in the field of Computational Linguistics as well as in Artificial Intelligence are summarised.

Chapter 4 gives an overview of cross-language information retrieval focusing on the problems multilinguality introduces to IR. Current concepts dealing with this additional aspect are presented, ranging from knowledge-based systems which take advantage of bilingual lexicons and/or thesauri, systems using full-fledged machine translation systems up to corpus-based approaches using statistics over word co-occurrences. Their main aim can be described as achieving an efficiency similar to that monolingual retrieval systems show.

Chapter 5 is dedicated to the detailed presentation of the cross-language component MPRO-IR. The approach introduced within this work is based on extended linguistic knowledge provided by a morpho-syntactic analysis which is described in Section 5.2, after a short presentation of the motivations which have guided the development (Section 5.1). Within this version of the system, the investigations have been limited on the following three types of linguistic knowledge: The lexical base form, the morphological derivation and, for German only, information about decomposition. In contrast to other systems using linguistic information, in MPRO-IR this information is exploited not only for indexing (Section 5.3), query translation (Section 5.4) and expansion (Section 5.5) but is also used within the search (Section 5.6.1) which allows furthermore a new approach to rank and to merge the retrieved documents (Section 5.6.2).

Chapter 6 describes the application of MPRO-IR as a core component of an information system in a legal domain – EMIS. The information system is presented with a focus on the adaptations which were necessary in order to make the retrieval component operational.

Chapter 7 focuses on performance considerations usually measured in *recall* and *precision*. The main aim of MPRO-IR has been to increase performance by a better recall without completely neglecting the precision. To get a sufficient overview on the effectiveness of MPRO-IR, the system has been evaluated within two different test environments: First, the system is evaluated within a special domain, i.e. EMIS, which demands high recall. The second test bed comprises a large archive of unrestricted texts, the CLEF 2000 evaluation campaign, where the precision is in the focus. The chapter finishes with a discussion of which ways

the three different linguistic knowledge types contribute to achieve higher recall, and their impact on performance.

Chapter 8 sketches some enhancements envisaged for a further improvement of MPRO-IR and related to indexing, query expansion, and the search algorithm. Additonally, the benefit of linguistic processing by considering further information such as number, case and gender already provided by the morphological analysis, but also semantic information is examined.

The thesis finishes with a summary followed by an extensive bibliography.