Université de Neuchâtel

Faculté des sciences

Cette thèse intitulée:

Robust Multilingual Information Retrieval

présentée par:

Martin Christian Braschler

a été évaluée par un jury composé par les personnes suivantes:

Prof. Dr. Jacques Savoy (co-directeur de thèse)

Université de Neuchâtel, Suisse

Prof. Dr. Peter Kropf

Université de Neuchâtel, Suisse

Dr. Gareth Jones

Dublin City University, Irlande

Dr. Peter Schäuble (co-directeur de thèse)

Eurospider Information Technology AG, Suisse

Thèse acceptée: Neuchâtel, le 10 juin 2004.

IMPRIMATUR POUR LA THESE

Robust Multilingual Information Retrieval

M. Martin BRASCHLER

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel, sur le rapport des membres du jury

MM. J. Savoy (co-directeur de thèse), P. Kropf, P. Schäuble (co-directeur de thèse, Zürich) et G. Jones (Dublin, Ireland)

autorise l'impression de la présente thèse.

Neuchâtel, le 10 juin 2004

La doyenne:

Martine Rahier

Abstract

We present in this dissertation our combination approach for robust multilingual information retrieval (MLIR). Many people, especially those living in industrialized countries, are today exposed to an ever-increasing amount of (digital) information. Increasing globalization, among other factors, implies that more and more of this information is written in languages different from the ones preferred by the users of an information access system. Tools that allow access across languages ("bridging the language gap") therefore seem essential.

The goal of our multilingual information system is to make information accessible regardless of language differences between the searcher and the information. The system accepts queries in one language, and allows retrieval of documents written in one of several languages. For our experiments, we have adapted the system to five widely spoken Western European languages.

Early research in cross-language information retrieval (CLIR), the subfield of academia concerned with information access across languages, led to systems that, although using very different translation methods, performed at around 50%-60% of the effectiveness of similar monolingual information retrieval systems. Our first main goal was therefore to advance retrieval effectiveness of our system beyond the level reached by those early approaches. One of our observations was that different approaches led to very different performances for individual queries, even if their average performance was remarkably similar. We show that the use of a combination of several "simple" approaches can substantially increase retrieval effectiveness, drawing on the strengths of the individual translation methods. Besides the right approach to translation, good monolingual retrieval is also essential as a solid foundation for such a multilingual retrieval system, and we have investigated the issues encountered when adapting the system to new languages. To this end, we present an extensive study of German text retrieval.

The effectiveness of our methods has been evaluated by participating in a number of evaluation campaigns, and results were consistently competitive with those reported for the best performing systems. As determined at these evaluation campaigns, the level of CLIR performance has now risen to more than 80% of the best monolingual performance for selected pairs of widely spoken languages, thus systems such as ours have reached a high level of maturity in this respect.

Consequently, retrieval effectiveness was not our only goal and reason for choosing a combination approach. Rather, we wanted to develop robust methods that generalize well beyond specific test scenarios. We have designed our system with applications in operational settings in mind, avoiding over-tuning to laboratory settings during evaluation. We argue that while retrieval effectiveness has increased in recent years, commercial uptake of CLIR technology is still low. In operational settings, additional criteria come into prominent focus, and we demonstrate how we have designed our system with these criteria in mind. The stated goal is to achieve equilibrium between retrieval effectiveness as measured by laboratory experiments

and additional concerns, such as those encountered in operational settings. Among the issues that we have investigated are lexical coverage and the avoidance of negative outliers.

Table of Contents

1 Introduction

1.1 Motivation

This dissertation describes efforts to advance the state-of-the-art with respect to methodologies to access textual information "across languages", i.e. to access information from a multilingual "pool of texts" regardless of which language is used by the searcher for access.

The advent of the "information age", and recently the success story of the Internet, has brought exposure to an ever-increasing amount of (digital) information[1] for many people, mostly those living in industrialized countries. The ease and speed of transmission of digital data, and not least the continuing globalization, imply that more and more of this information is written in languages different from the ones preferred by the potential consumers. Some sources of global information, such as the World Wide Web, are open to a vast number of potential users. We postulate that tools that are capable of handling multiple languages for information access, bridging the language gap between a user and the content to be accessed, are therefore essential. To stay with the example of the Web, we observe that while initially most of the users came from English-speaking countries or from academic circles, making English the "lingua franca", the number of non-English-speaking users has risen dramatically from earlier years[2]. The marketing communications consultancy "Global Reach" estimates the number of non-English speaking Internet users to be around half a billion users for 2003, with more than 800 million projected for 2005. In 1996, according to the estimates, only 5 million non-English speaking users were online (Global Reach, 2004).

It is important to remember that the Web is just one factor in the increasing information load that users face today, albeit one that is very much in the public's eyes. It seems safe to assume that increasing globalization of markets has led to workers being confronted with more foreign-language information in general. Globalized market places imply that the inability to access information on a global scale is a serious distraction to commercial success. In particular, non-English

[1] An attempt to measure the amount of new information created every year has been made by researchers at the School of Information Management and Systems at the University of California at Berkeley both in 2000 (Lyman and Varian, 2000) and 2003 (Lyman and Varian, 2003). They estimate that the amount of newly stored information grew about 30% a year between 1999 and 2002.

[2] There is, however, no general agreement on whether the Web has continued to become more multilingual in recent years, e.g. in terms of the percentages of English-speaking users vs. non-English-speaking users or of content written in English vs. content written in other languages. As an indication to the contrary, the Web Characterization Project WCP (http://wcp.oclc.org/) found no change in the relative frequency of English pages on the Web between 1999 and 2002 (72% of all pages were written in English for both years) (O'Neill et al, 2003).

speakers need to have access to the vast amount of information that is only available in English.

Aside from pressure created by an information flood that is increasingly multilingual, there are sites that have long lived with the need to create, manage, store and retrieve multilingual information, such as large international corporations and organizations, and multilingual countries. In such environments, efficient and effective access to multilingual information is essential to guarantee cohesion.

1.2 Goal

The goal of the research underlying this study is to advance the state-of-the-art in electronically accessing multilingual information[3]. As will be demonstrated, research in cross-language information retrieval (CLIR), the academic field concerned with multilingual information access, had reached a point in 1997/98, where the "simple" methods being widely in use were performing at an effectiveness of around 50-60% of that of comparable monolingual systems, as determined by experiments in the context of evaluation campaigns[4]. This level was reached by several competing approaches, with none of the approaches clearly showing superior performance. We thus argue that a "glass ceiling" of performance for "early generation" approaches had been reached, and that the time was right to move forward to a new generation of more effective CLIR approaches.

Efforts to transfer CLIR technology that we have developed in the form of research prototypes to operational settings have led us to define additional criteria that we believe are important in order to be able to generalize beyond academic test collections and in order to establish commercial viability in real-world applications. We then set out to study how these additional criteria affect the design of our CLIR methods.

We have tried to address both concerns prominently in our work. As an answer to the challenges that faced us, we have started work on combination approaches that incorporate elements of several simpler CLIR methods. We feel that in this way we are able to establish equilibrium between the need to improve on the level of performance obtained by any single method on its own (therefore "breaking the

---

[3] While we do not specifically exclude any particular languages, we have focused mostly on frequently spoken Western European languages, namely English, French, German, Italian, and Spanish. The methods and conclusions do in principle often generalize beyond this set of languages, although there are obviously additional issues when e.g. adapting to East Asian languages, such as character encoding and segmentation.

[4] See Schäuble and Sheridan (1998) for a discussion of results in the TREC-6 evaluation campaign held in 1997, where a broad range of "simple" approaches was tested in a common evaluation scenario. A similar observation specifically with regard to query translation methods has been made by Oard (1998), who notes that "[...] simple query translation approaches suffer a severe penalty in effectiveness, usually achieving about half of the retrieval effectiveness of corresponding monolingual techniques when typical measures such as average precision are used".

glass ceiling") and the requirements of the additional criteria deemed essential for operational use. We will comment on our experiences with respect to these main goals throughout this study.

1.3 Definitions

Let us first clearly define some essential technical terms used freely throughout the dissertation.

- Information Retrieval (IR)

The academic field of "information retrieval" is concerned with the examination of methods to store and access large amounts of information. Access usually takes place in the form of formulating a "query" for the system, which is then used for searching. In a larger context, the query reflects an information need by the user, which the system answers by supplying items in a result set that are expected to contribute to the satisfaction of the user's information need. The term "information retrieval" is frequently used interchangeably with "document retrieval", emphasizing the fact that IR systems are most often designed to work with information stored in the form of documents, with the matching documents returned in the result set. Most research efforts so far have been spent on "text retrieval", i.e. information retrieval on textual information. If documents contain long, complex text passages using uncontrolled vocabulary, the term "full-text retrieval" or "full-text search" is also used. If nothing is specified to the contrary, the term "information retrieval" is used throughout this dissertation to describe full-text retrieval on (textual) documents. Information retrieval as an academic field today also comprises of related research areas that do not strictly use queries to access the information, but which are concerned with further questions of information access, such as (text) categorization, filtering, (text) summarization, and others.

- Document Relevance

By concentrating on full-text retrieval of textual documents, we have to deal with the special properties of such "retrieval items". The retrieval system returns result sets consisting of documents, and thus answers queries in an indirect way. Documents should be included in the result set if they contribute to the satisfaction of the information need that served as the basis of the user's query. It is important to note that a query often constitutes only an imperfect representation of the underlying information need: it would be paradox to expect a user that lacks information on a topic to formulate a comprehensive, targeted query. Furthermore, the unstructured nature of typical full-text documents presents additional problems, as the same facts can often be stated in numerous ways in natural language – different word inflections, synonyms and different paraphrases can, among others, lead to mismatches between query and documents. Again, a user can normally not be expected to supply examples of all possible formulations of the kind of information that is sought. Comparing a user's query to the documents accessible by the retrieval system is thus a complex task – based on the imperfect formulation of the information need that the query constitutes, the system has to select items that in

some, potentially paraphrased, form contain information that is "relevant" to the topic at hand. The imprecise nature of this process implies that not all documents that are returned are necessarily relevant in this sense. Even if a system were devised which solved all problems involved in matching the query perfectly to the documents, the relevance of documents would still depend on the individual user's interpretation of the information that is contained in them. This interpretation is usually dictated by personal preferences, prior background knowledge and other considerations. As a consequence, when evaluated using accepted measures, retrieval results score typically well below the theoretical optimum. Modern information retrieval systems typically react to the uncertainties involved in determining document relevance based on a query by calculating retrieval scores that allow ranking documents according to their probability of being relevant to a query.

- Monolingual Information Retrieval

If an information retrieval method or system does not make any provisions for handling potential language gaps between the language of the query (or, more generally, of the users) and the one or more languages in which the retrievable documents are written, we speak of "monolingual information retrieval" in this dissertation. Note that according to this definition, a system that can handle documents in many different languages is still a monolingual information retrieval system if it does not allow querying across languages. This is contrary to some definitions where a system integrating documents from multiple languages is referred to as a "multilingual information retrieval" system, regardless of its ability to cross any language barriers.

- Cross-Language Information Retrieval (also abbreviated as "CLIR" in the following)

Cross-Language Information Retrieval denotes the academic subfield concerned with developing information retrieval methods and systems that allow access to information across languages, e.g. retrieval of documents written in one or more languages different from the query language. Synonyms also found in the literature include "translingual information retrieval" and "cross-lingual information retrieval".

- Bilingual Information Retrieval

Bilingual information retrieval denotes a special case of cross-language information retrieval, where only two languages are considered, one as the language of the queries (or the users), and the other as the language all retrievable information is formulated in.

- Multilingual Information Retrieval ("MLIR")

Multilingual information retrieval denotes the generic set-up for cross-language retrieval, where, for a set of languages that the system can handle, the query and the

documents can be formulated in any of those languages, with the system retrieving information regardless of the language combinations involved[5].

- Combination Approaches

In the following, we make numerous references to combination approaches for CLIR and MLIR, as opposed to what is termed "simple" approaches. The focus is on the resources and methods that are used to bridge the language gap. We consider a system to use a combination approach if integrates at least two "simple" translation approaches that can also be used independently. The motivation for using a combination is to simultaneously "add up" the positive and dampen the negative contributions from each method.

1.4 Achievements

In this dissertation, we examine multilingual information retrieval, the generic case of cross-language retrieval. The dissertation makes two main contributions to the development of robust multilingual information retrieval systems:

1. Firstly, we observe that the main simple alternatives for cross-language information retrieval proposed in recent years yield limited performance when compared to similar, monolingual scenarios. We present a combination approach, using pieces from all major CLIR alternatives, that leads to superior retrieval effectiveness when evaluated on a large, standardized test collection.

2. Secondly, we examine to what extent additional criteria come into play when generalizing beyond specific test collections, e.g. when operational use of a multilingual information retrieval system is considered. In particular, we want to avoid over-tuning in response to characteristics that are merely artifacts of the collection. To this end, we list some of the issues involved in operational settings, and show how we addressed them.

In addition to this core focus of the dissertation, we explore the state-of-the-art of CLIR systems, analyzing common characteristics of successful systems. We believe that a CLIR system is only as good as the monolingual methodology that underlies it. Therefore, we carefully examine the issues involved in tuning an information retrieval system to a language other than English (in our case, German).

Last but not least, we have evaluated our approach extensively by participating in two TREC CLIR tracks and three CLEF evaluation campaigns, and we describe some of the experiments that we submitted. The CLEF campaigns (and the TREC

---

[5] As we have already noted in our discussion of monolingual IR, this is not the only possible definition of MLIR. An overview of different definitions is given in Hull and Grefenstette (1996). Our system and methods can support all five definitions they provide, although the experiments in our work presented here operate mostly on definition (4) "IR on a multilingual document collection, where queries can retrieve documents in multiple languages", due to the properties of the test data that we have used.

CLIR tracks we participated in) conduct comparative evaluation of cross-language IR systems for European languages, including the set of languages (English, French, German, Italian and Spanish) that our work was focused on.

1.5 Evaluation Methodology

The stated goal of our work to advance the state-of-the-art in multilingual information access, or, more specifically, to propose and implement mechanisms that improve the effectiveness of a multilingual information retrieval system, calls for a careful methodology to evaluate the results. Evaluation of information retrieval systems, and especially evaluation of retrieval effectiveness of such systems, has a long-standing tradition. The "Cranfield experiments" conducted by Cleverdon (1967) have had a great impact on the set-up of such evaluations and form the basis for what today is widely accepted as "best practice" in the field. Clearly, evaluation of information retrieval systems is not limited to assessing retrieval performance, but can also encompass other issues, such as studies of interaction between users and the system. As we concentrate in our work on the retrieval mechanism proper of the system, and on how to improve its effectiveness and robustness, we limit ourselves to system evaluation, that is, we abstract the retrieval process from particular users and equate good performance with the ability of the system to return "good" document sets in answers to incoming queries. While there is obviously much valuable insight to be gained from directly involving users in the evaluation of the system, this abstraction allows us more control over some of the variables that affect retrieval effectiveness in the form we aim to measure. Furthermore, such "laboratory tests" are less expensive and can, as we shall demonstrate, benefit from existing evaluation data.

All of our evaluations of retrieval effectiveness of the mechanisms we propose have been conducted either in the context of the TREC (Harman, 1995) and CLEF (Peters and Braschler, 2001) evaluation campaigns or have been based on using test data created by the organizers of these campaigns. Both campaigns are based on the Cranfield paradigm, which emphasizes comparative evaluation. A campaign channeling the evaluation effort of a large number of interested research teams is a natural set-up to maximize the benefits of this paradigm. The basic building blocks to receive comparable, reproducible evaluation results are a number of exemplary information needs (called "topics"), a set of documents to retrieve from, and manual assessments of relevance of these documents with respect to the topics, together forming what is known as a "retrieval test collection". The success of a retrieval method or system is assessed by its ability to select from the document set those that are deemed relevant to a specific "topic". Questions of different interpretations of relevance of documents with respect to the topic are usually abstracted from, choosing one representative interpretation.

The two measures "recall" and "precision" are most widely accepted for assessing the retrieval effectiveness based on the result set. Recall measures the ability of the system to provide comprehensive results: it measures the proportion of relevant documents that have been retrieved by the system versus the actual number

of relevant documents available in the document set for a topic. Precision measures the proportion of relevant document contained in the result set versus the overall size of the result set, including unwanted, irrelevant documents that were erroneously retrieved. Which of the two measures is the more appropriate depends on the user need that is modeled. The two measures can be contradictory: a system tuned to provide maximum recall often has to relax matching mechanisms to allow extra matches between queries and documents, which often leads to additional unwanted documents being retrieved as well. Similarly, a system tuned for maximum precision is usually configured in a restrictive manner, resulting in some relevant documents going undetected due to the loose relationship between them and the query. Systems providing ranked lists of documents as retrieval results allow users to determine the degree of recall and precision that best fits their information need: since documents with the highest probability of relevance, as determined by the retrieval mechanism, are ranked at the top, precision will be high if only few documents are inspected. Including more documents in the result set will likely result in degraded precision, but should lead to higher recall, with new relevant documents appearing at the low end of the result list. Automatic evaluation procedures calculate precision and recall at various levels, to cater for different underlying user preferences. The measure of average precision provides a single performance value based on an average of precision values taken at various levels of recall. It therefore embodies an average performance assessment of the system over various possible user preferences. Exact procedures for calculating the different measures can be found in Schäuble (1997). Recall and precision at different levels, as well as average precision, are used prominently in our work to measure performance increases due to our various modifications to our system.

Following the Cranfield paradigm, and using the test collections provided by the TREC and CLEF evaluation campaigns allows us to benefit from the extensive scrutiny that this methodology has been subject to over the years. A good article further discussing aspects of the set-up, underlying assumptions and interpretation of the results from such evaluations has been presented by Voorhees (2002).

1.6 Background Information

In the following, we attempt to give an overview of early developments in those fields of information retrieval and cross-language retrieval that are important for the understanding of our work. Excellent general introductions into the history of information retrieval can be found in Sparck Jones and Willett (1997) and Frakes and Baeza-Yates (1992), and into the early work in CLIR in Grefenstette (1998). Related work to specific research questions that we have worked on is detailed in the individual papers that constitute the core of this dissertation.

The academic field of information retrieval is relatively old, given its recent rise to prominence in the Internet age. By necessity, we restrict our discussion of the history of information retrieval to some of the most prominent early milestones in IR research. The beginnings date to the early 1950s, with the term "information retrieval" coined by Mooers in 1952 (Sparck Jones and Willett, 1997). Initially,

much of the focus was on working with controlled vocabulary used for library records in bibliographical databases. Basic notions such as document relevance, probabilistic retrieval, and relevance ranking were subsequently introduced (see e.g. Maron and Kuhns (1960)). The late 1960s brought experiments with full-text and queries formulated in natural language, and the introduction of such concepts as the vector space model popularized by Salton and his associates (see e.g. Salton and McGill (1983)). The basic procedures for evaluation of retrieval effectiveness date back to Cleverdon's experiments in the same time period (Cleverdon, 1967). These procedures form the basis for today's IR evaluation campaigns, such as TREC[6] in the United States (Harman, 1995), CLEF[7] in Europe (Peters and Braschler, 2001) and NTCIR[8] in Asia (Kando, 2003). Finally, another milestone was the careful analysis of the probability ranking principle by Robertson (1977) in the 1970s, which brought mathematical justification for new schemes for the weighting of query terms.

First experiments with cross-language retrieval are also quite old, and date back to the early 1970s, when Salton was experimenting with using multilingual thesauri for controlled vocabulary retrieval (Salton, 1970). The field lay dormant for a while before first experiments on full-text cross-language retrieval were conducted in the early 1990s, notably experiments with latent semantic indexing (Landauer and Littman, 1990) and experiments in the course of the EMIR project (EMIR Consortium, 1994; Fluhr et al., 1998)). It is likely that these early studies had an essential influence on the direction of much early work that followed in the second half of the decade. Landauer and Littman demonstrated that it is possible to build CLIR systems that base their translation process on training data. Evaluation was based on assessing the system's ability to correctly pick out translated documents from a parallel corpus. The EMIR project consortium aimed to build CLIR mechanisms that can be incorporated in operational systems, and developed mechanisms for dictionary-based cross-language retrieval. One focus of the work was on the (corpus-based) elimination of incorrect translation alternatives coming from the dictionary. The resulting prototype was compared to state-of-the-art machine translation, and was found superior[9], assumedly due to the MT system being limited to providing just one, potentially wrong, translation alternative.

It is generally acknowledged that the recent rapidly growing interest in CLIR in academic circles was initiated by a workshop held in conjunction with the SIGIR 1996 conference in Zurich. Some of the most promising results emerging from this workshop were published in Grefenstette (1998). The next year, the "AAAI Spring

[6] http://trec.nist.gov
[7] http://www.clef-campaign.org
[8] http://research.nii.ac.jp/ntcir/
[9] It must be noted, however, that the Cranfield test collection, which was used in the tests, must be considered as being of very limited size for today's standards in IR evaluation.

Symposium on Cross-Language Text and Speech Retrieval" brought the formulation of a "grand challenge" for CLIR[10] (AAAI, 1997).

Methods and systems for cross-language retrieval can be categorized according to how they cross the language barrier – whether they use translation of queries, translation of documents, or both (Braschler et al., 1999). In recent years, a small but active subfield has sprung up that attempts CLIR without using any translation at all[11]. Further division of methods is possible according to the type of translation resources used. Different ways of subdividing CLIR methods are possible, such as the attempt by Oard (1997), who uses a division into approaches using controlled vs. free vocabulary, followed by a grouping according to translation resources.

Early work published in the years immediately after the 1996 SIGIR workshop is characterized by the efforts of researchers to understand the behavior of the basic types of translation resources that had been identified for use in CLIR by that time: machine-readable dictionaries (MRDs), machine translation (MT) systems, and corpus-based, automatically generated data structures. Researchers generally used only one of these different types of resources, and measured the effectiveness of their prototypical systems against a monolingual baseline that was obtained by running a similar system without any translation steps. Systems were typically limited to bilingual retrieval, and mostly to one or two language pairs (usually involving English).

Good examples for early work in dictionary-based CLIR include Ballesteros and Croft (1997) and Hull and Grefenstette (1996). Among the issues identified as most important are coverage of the dictionary (dealing with out-of-vocabulary search terms), ambiguity of search terms (synonyms, homonyms), and the mapping of search terms onto base forms listed in the dictionaries. A first notable (and influential) break-through was work by Ballesteros and Croft (1997), who introduced pre-translation and post-translation feedback with the goal of broadening the query before and after retrieval, thus reducing the risk of lexical coverage problems, and producing more complete translations.

Machine translation systems have been targeted for CLIR mainly because of the ease of the CLIR system architecture that they imply: they make it possible to continue using a monolingual IR system while either translating the query or

---

[10] "Given a query in any medium and any language, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in different media or languages appropriately identified." (AAAI, 1997). This ambitious definition goes beyond the scope of our work, in that it also addresses the questions of multimedia and of presentation of results, which we exclude from our study.

[11] Experiments by Cornell University on English/French bilingual retrieval using cognates were an early notable attempt at CLIR without the use of any explicit translation (Buckley at al., 1998). Recently, a group at Johns Hopkins University has been reporting remarkable results when using n-grams to match across languages (McNamee and Mayfield, 2002).

documents before retrieval. Translating the documents has not been attempted often, due to problems with performance and stability of many MT systems when faced with the amount of data typically used in IR system development and evaluation campaigns. Examples of document translation using MT include Oard (1998), Chen and Gey (2004) and Braschler et al. (2003). Query translation using MT is more popular today, even though it had been mainly used as a comparative baseline in early experiments, such as those by Fluhr et al. (1998), Oard (1998) and Mateev et. al (1998). The comparatively low interest in MT for CLIR in early work is interesting, as MT output would seem like an obvious starting point for experimentation. The closed nature of most MT systems and the influence of previously published reports stemming from the EMIR project, which indicated relatively weak performance for MT-based CLIR (EMIR Consortium, 1994), may have influenced researchers in their choice of translation resources. Major problems identified with respect to using MT for CLIR include the bad handling of short, ungrammatical queries by MT systems, and performance problems.

Corpus-based, automatically generated data structures were identified early on as an attractive alternative to manually generated resources such as dictionaries. Algorithms that derive translation resources based on bilingual or multilingual training data have the advantage of potentially delivering retrieval-optimized translations that are tailored to specific collections. The ability to build such data structures automatically allows resources to be produced with very large coverage. Early work on CLIR using such corpus-based data structures includes work on latent semantic indexing (Landauer and Littman, 1990), similarity thesauri (Sheridan and Ballerini, 1996) and the generalized vector space model (GVSM) (Carbonell et al., 1997). The major problems are the acquisition of suitable training data, computational complexity and the minimizing of errors in the statistical process used to derive translations. Availability of suitable training data is a core issue, especially for pairs of less widely spoken languages. There have been attempts to extract content from the Web (see e.g. Nie et al. (1999)), but there are issues with quality and size (Nie and Simard, 2002). Available parallel corpora often cover content from specific domains or, as noted by Kwok et al. (2001), there may even be cultural issues if the data is acquired from foreign sources[12], which makes a transfer of the resulting data structures to text from different domains or origins difficult.

The cross-language track at the TREC-6 conference presented the first real opportunity to test the systems that resulted from the early work outlined above in a well-defined setting allowing a direct comparison of approaches (Schäuble and Sheridan, 1998). Much of the work up to that date was characterized by more limited evaluations based on whatever was available in terms of test data at the time. While some promising numbers were reported, it remained unclear how to interpret the monolingual baselines that were often used for comparison. The TREC-6

[12] Kwok et al. (2001) comment on the difference of Chinese as written in Mainland China versus the form written in Hong Kong, but similar problems can be expected for languages such as English and Spanish, which have a wide global distribution.

evaluation suffered from an unwieldy number of bilingual language pairs that could be chosen by participants for their experiments, but nonetheless a clearer picture emerged for the two language combinations that were most popular among the participants: English→French and English→German. For English→French, no team reached a CLIR performance that was within 50% of the best monolingual performance reported[13]. For English→German, the top result was roughly 64% of monolingual performance[14] (Voorhees and Harman, 1998). What makes this observation particularly stand out is the fact that examples of systems using all different types of translation resources participated in the evaluation:

- Systems using machine-readable dictionaries (MRDs), e.g. (Gaussier et al., 1998),
- Systems using machine translation (MT), e.g. (Oard and Hackett, 1998).
- Systems using corpus-based approaches, e.g. (Mateev et al., 1998).

We therefore derive two major lessons from the TREC-6 results:

1. It appeared at the time that a "glass ceiling" of around 50%-60% of monolingual performance had been hit, which was not easily breakable by any of the principal early approaches to CLIR. TREC-6 was thus a good starting point for the community to investigate how to move beyond these "simple" approaches.
2. Results from small-scale evaluations with little or no possibility of comparisons with other systems, such as those reported prior to the TREC-6 CLIR track, have to be interpreted very carefully. While they serve to indicate promising directions for further research, the conclusions that can be drawn are usually limited to the "special setting" that was used for evaluation, and may not generalize to other situations.

The first of these two issues we address directly, as it was a stated goal of our work to investigate ways to improve the effectiveness of CLIR systems beyond what simple approaches can offer. However, we also view the second issue as being of major importance, and have therefore evaluated the systems resulting from our efforts by participating in evaluation campaigns and using the corresponding large, reliable test collections.

---

[13] The comparison is in terms of mean average precision (MAP). A group from Xerox Research Centre Europe reported the best monolingual French retrieval performance, with a MAP of 0.4073 (Gaussier et al., 1998). The best CLIR performance was reported by the Cornell University group, with a MAP of 0.1982, using a comparatively simple cognate matching strategy (Buckley et al., 1998).

[14] The best monolingual German performance was reported by the Swiss Federal Institute of Technology, with a MAP of 0.3349. The best CLIR performance came from the same group (MAP of 0.2135), using machine translation (Mateev et al., 1998).

1.7 The Structure of this Dissertation

The structure of the remainder of this dissertation is the following. In Section 2. we present the published papers that form the core of the dissertation. We start the section by providing the necessary context to understand the role of the individual papers. The rest of the section is then divided into individual subsections for each paper. In each of these subsections. we will briefly summarize the main aspects of the paper, especially with respect to its importance to our goal of developing a robust multilingual information retrieval system that generalizes to a large range of settings. Section 3 summarizes the results of our studies, before examining more closely the significance of our findings. and their relevance to the state-of-the-art in CLIR. We comment on the applicability of our results in a broader sense. e.g. to operational settings, which is one of our main concerns. Finally, we take a look into the future, seeking ways to improve CLIR effectiveness by looking at various interesting issues that remain more or less unresolved. The dissertation finishes with acknowledgments (Section 4) and a list of references.

2 Presentation of papers

The core of this dissertation is formed by four published papers:

1. Martin Braschler, Carol Peters: *Cross-Language Evaluation Forum: Objectives, Results, Achievements*, Information Retrieval, pages 7-31 (Paul B. Kantor, Josiane Mothe, Justin Zobel (Eds.), Information Retrieval, Volume 7, Issue 1/2, Kluwer Academic Publishers, ISSN 1386-4564) (Braschler and Peters, 2004)

2. Martin Braschler, Bärbel Ripplinger: *How Effective is Stemming and Decompounding for German Text Retrieval?*, Information Retrieval, pages 291-316 (Paul B. Kantor, Josiane Mothe, Justin Zobel (Eds.), Information Retrieval, Volume 7, Issue 3/4, Kluwer Academic Publishers, ISSN 1386-4564) (Braschler and Ripplinger, 2004)

3. Martin Braschler: *Combination Approaches for Multilingual Text Retrieval*, Information Retrieval, pages 183-204 (Paul B. Kantor, Josiane Mothe, Justin Zobel (Eds.), Information Retrieval, Volume 7, Issue 1/2, Kluwer Academic Publishers, ISSN 1386-4564) (Braschler, 2004)

4. Martin Braschler, Anne Göhring, Peter Schäuble: *Eurospider at CLEF 2002*, CLEF 2002, pages 164-174 (Carol Peters, Martin Braschler, Julio Gonzalo, Michael Kluck (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Rome, Italy, September 2002, Revised Papers, Lecture Notes in Computer Science, Vol. 2785, Springer, 2003, ISBN 3-540-40830-4) (Braschler et al., 2003)

These papers answer different questions with regard to the goal of building an effective, robust multilingual information retrieval system. We will comment on specific questions addressed by the individual papers in the following subsections.

The main motivation for choosing our direction of work was the relatively poor performance of CLIR methods around 1997, when systems using a wide range of different approaches all obtained a retrieval effectiveness of only around 50-60% of the best monolingual systems. One of our observations was that different approaches led to very different performances for individual queries, even if their average performance was remarkably similar. We were therefore convinced that the best way to reach the next level in terms of effectiveness was not to keep tuning (and possibly over-tuning) the individual methods or translation resources, but instead to shift the focus to a larger framework that would allow the incorporation of multiple "simple" approaches, combining and merging aspects from each of the methods.

From the beginning, we have evaluated our work in the context of evaluation campaigns, first at the TREC conferences in the United States (TREC-7 cross-language track in 1998 and TREC-8 cross-language track in 1999), then later at the first three CLEF campaigns (CLEF 2000, CLEF 2001, and CLEF 2002). Working in the context of an evaluation campaign allows comparisons across systems, and helps to avoid comparisons to weak baselines. It became quickly apparent that high-quality monolingual retrieval is an essential basis for an effective CLIR system. Indeed, we observe today that many participants in CLEF start out with monolingual systems and then gradually proceed to bilingual and multilingual tasks within successive campaigns. We have therefore structured the exposition of our work in a similar fashion: while we start by commenting on the state-of-the-art in multilingual information retrieval, we discuss the implications of tuning a system for non-English monolingual retrieval before we turn to the description of our approach to multilingual information retrieval.

The landscape of CLIR systems and approaches has changed drastically during the course of our work, not least due to the influence of evaluation campaigns such as TREC, CLEF and NTCIR. We feel that combination systems have become an important direction in CLIR research, and have found that some "blueprints" for successful cross-language retrieval seem to be emerging. For the most popular language pairs, especially the widely spoken Western European languages, CLIR systems have clearly moved past the effectiveness level of 50-60%, today reaching performance regions where we feel that other issues gain increasing importance, such as the considerations that come into play in operational settings. While we feel that from the standpoint of laboratory testing, the field has reached a high degree of maturity in terms of retrieval performance for selected language pairs, it is obvious that much still remains to be done with respect to the uptake of CLIR technology in the "real world".

2.1 State-of-the-Art in Multilingual Information Retrieval: "Cross-Language Evaluation Forum: Objectives, Results, Achievements" (with Carol Peters)

Cross-Language Information Retrieval as an academic field has made substantial progress during the course of our research that led to this dissertation, and we want to start the exposition of our work with a paper that pursues two main goals: firstly, to outline the state-of-the-art in CLIR systems and CLIR evaluation, and secondly, to search for a blueprint of an effective CLIR system.

The paper was co-authored by Carol Peters, and both authors have been directly involved in organizing the CLEF ("Cross-Language Evaluation Forum") campaigns, which gave valuable insights into the workings of such an evaluation exercise and the directions of research that were pursued by its participants.

CLEF is a successor of the cross-language track at the TREC ("Text Retrieval Conference") series of conferences (Harman, 1995). The cross-language track was first held in 1997, offering bilingual retrieval tasks for Western European languages (English, French and German documents), and was very influential for our work in that it demonstrated the limits of early CLIR approaches (Schäuble and Sheridan, 1998). We subsequently became involved both in the organization of the two following TREC CLIR tracks (Braschler et al., 1999; Braschler et al., 2000) and in participating with our own systems, based on combination approaches (Braschler et al., 1999b; Braschler et al., 2000b). An overview of work in the three cross-language tracks from TREC-6 through TREC-8 can be found in Harman et al. (2001).

While successful in their own right, it was felt after TREC-8 that the cross-language tracks at TREC attracted too few participants from Europe, a region it was believed would benefit enormously from more work on CLIR. The activities for CLIR involving Western European languages were therefore extended and moved to Europe, with a new multi-national consortium set up as organizers, and funding provided by the European Commission.

The CLEF campaigns are structured into tracks, areas for experiments on different research questions. From the beginning, CLEF has offered tracks for monolingual, bilingual and multilingual retrieval on news texts, as well as a track for retrieval on domain-specific content (Kluck and Gey, 2001). In later years, tracks for interactive retrieval (Oard et al., 2004), speech retrieval (Jones and Federico, 2003), image retrieval (Clough and Sanderson, 2003), and question answering (Magnini et al., 2003) have been added. We comment on the structure of CLEF and the different tracks in the paper.

The main tracks of CLEF (monolingual, bilingual, and especially multilingual retrieval on news texts), like TREC, follow the Cranfield methodology (Cleverdon, 1967). We have discussed the set-up of an evaluation following this paradigm earlier in this dissertation. The paper gives more details on how the paradigm was adapted for a multilingual setting and for evaluation of cross-language retrieval.

For scoring systems, the main tracks at CLEF use a test collection consisting of a set of retrievable documents ("document collection", up to 1.6 million documents for CLEF 2003), a set of formulations of "information needs" by hypothetical users[15] (up to 60 for CLEF 2003) and measures for evaluating the effectiveness of the system in answering these information needs on the basis of the document collection. By 2003, CLEF supported nine document languages: Dutch, English, Finnish, French, German, Italian, Russian, Spanish and Swedish. The detailed characteristics of the corresponding document sets are given in the paper.

Calculation of the effectiveness measures is based on the document relevance with respect to the corresponding information need, which is defined in terms of topical similarity as determined by domain experts ("relevance assessors"). As main measures, CLEF has adopted precision and recall, as introduced earlier in the dissertation. Both measures are influenced heavily by the interpretation that is used in determining the relevance of a document with respect to a specific query. For reasons of practicality, a fixed interpretation of relevance is used, and closer examination of the resulting implications can be found in the paper. The paper also details the process of topic creation, i.e. the selection of exemplary information needs used in the evaluation. Translation of the topics into multiple languages allows evaluation of various language pairs in CLR systems, but introduces problems in topic consistency. It is important, especially in context of the limitations of the underlying methodology, to stress that CLEF emphasizes comparative evaluation. Absolute scores of individual experiments derived in CLEF are not meaningful in isolation, but serve for relative comparisons to other systems and methods only.

The existence of a campaign concentrating on cross-language retrieval was essential to our efforts, as it allowed us to benchmark the system against realistic monolingual baselines, and thus supported our goal of working toward more effective multilingual text retrieval. We adopted the idea of using combination approaches for CLIR early (in TREC-7, see Braschler et al. (1999b)) by using a combination of two translation methods: query terms coming from either a similarity thesaurus or a bilingual wordlist combined with terms derived from a pseudo relevance feedback method using aligned documents (Braschler and Schäuble, 1998). In general, combination approaches have become increasingly popular at recent campaigns. The question arises whether, after seven years of CLIR evaluation campaigns, a blueprint has emerged on how to build a system that is effective for multilingual text retrieval. As we argue in the paper, several participants have obtained good results using combination approaches in recent campaigns. In the CLEF 2002 campaign, for example, the three groups with the top performing systems for the multilingual track all used combination approaches (Savoy, 2003; Savoy, 2004; Chen, 2003; Chen and Gey, 2004; Braschler et al., 2003). The paper presents a potential "recipe" for success in the CLEF multilingual track, where

[15] Known as "topics" in CLEF and TREC terminology. They form the basis for the formulation of queries, which are then run against the document collection.

systems are based on effective monolingual retrieval, a combination of multiple types of translation resources or methods, and the merging of intermediate results generated by query translation for bilingual retrieval. We see this conclusion, which is consistent with our own research efforts, as a validation of the solidity of our approach.

2.2 Monolingual Information Retrieval as the Base: "How Effective is Stemming and Decompounding for German Text Retrieval?" (with Bärbel Ripplinger)

Our work that went into this paper, which is co-authored by Bärbel Ripplinger, is based on two convictions:

1. Firstly, that non-English monolingual text retrieval is not necessarily equivalent to English-language text retrieval, and that some of the conventional wisdom that went into designing systems for English text retrieval needs to be re-examined when working on other languages, and
2. Secondly, that effective monolingual text retrieval is an essential basis for effective multilingual text retrieval.

Two main areas for investigation of the implications of adapting retrieval mechanisms to a new language seem obvious starting points: handling of word forms (e.g. morphology) during the indexing phase, and weighting of features during subsequent retrieval. While we feel that these two areas are probably the most promising, there are of course many other issues that also deserve attention, some of which become prominent in certain language families only, such as issues of text segmentation in Eastern Asian languages. We have chosen to fix the weighting scheme in our system to the popular Lnu.ltn scheme (Singhal et al., 1996). Lnu.ltn is very similar to Lnu.ltc, which in turn is reported to give good results in all the languages surveyed (Savoy, 2003). This weighting scheme also gave us competitive performance in our monolingual retrieval experiments for the CLEF campaigns.

Concentrating on the stemming issue, we have conducted a detailed study of stemming and decompounding for text retrieval in German, the language we were most actively working on. Stemming is probably the most prominent issue to be re-examined when adapting a system from one language to the next. A system using stemming conflates derived word forms to a common stem (for example, but not necessarily, through applying morphological analysis). There are several reasons for using a stemming procedure in a retrieval system, including the effects on the efficiency of the system. By conflating several forms to the same stem, the number of entries and therefore the size of the search index is reduced. More importantly from our viewpoint, stemming is potentially helpful to increase the effectiveness of free-text retrieval, where search terms can occur in various different forms in the document collection. Stemming makes retrieval of such documents independent of the specific word form used in the query.

The value of stemming for English-language text retrieval systems is controversial. There are some claims of performance improvements, but any such improvement is modest (see e.g. Frakes (1992) and Hull (1996)). Harman failed to find any benefit from using stemming in her experiments (Harman, 1991). In recent years, a number of studies on stemming behavior for languages other than English have been published. In these studies, performance improvements in terms of effectiveness are often reported that far surpass those observed for English. Often this effect is attributed to the larger declensional richness of the languages that have been studied. The paper lists a number of relevant publications for various languages.

An additional issue in German is the possibility of concatenating words to form new, more complex terms ("Compounding process")[16]. Concepts that are expressed as multi-word terms in languages such as English can then be captured by a single compound word. In German, this compounding process is very productive, with speakers producing new terms on the fly, making it impossible to compile an exhaustive list of all possible compounds. Since the same concepts can often also be expressed by a phrase, serious implications for retrieval arise. If a searcher enters only parts of a compound, or if a document contains a query concept only in phrasal form, mismatch will result during retrieval. The solution is to split compounds into their constituents, a process we will refer to as "decompounding".

Not much was published in terms of German stemming and decompounding before the TREC CLIR and CLEF campaigns started, probably due to the lack of test collections. Most notable were the PADOK studies (Womser-Hacker 1989; Krause and Womser-Hacker 1990), carried out in the late 1980s, which compared several approaches to automatic indexing on patent data. These studies observed beneficial effects from morphological analysis when using titles and abstracts as retrieval items, but not when using full texts.

Participants in CLEF have occasionally reported on their results when applying stemming and decompounding; but their experiments have been typically limited to a single approach, usually compared to the baseline of not using stemming at all. Sometimes the effect of stemming is obscured by other components that were deployed simultaneously, such as blind relevance feedback, making the analysis of those results harder. Examples from recent CLEF campaigns include work by Moulinier et al. (2001), who reported a performance gain of 20% using linguistic analysis (stemming plus compound analysis), work by Monz and de Rijke (2002), who show a gain of 25% using blind relevance feedback in addition to a linguistic processing, and work by Tomlinson (2002), who applied linguistic stemming including decompounding, and obtained a performance improvement of 43% in terms of retrieval effectiveness.

---

[16] The same phenomenon can be observed in other Germanic languages (e.g. Swedish), and some other, selected European languages, such as Finnish.

The defining characteristic and main contribution of our paper is that we were able to conduct a large-scale evaluation of the effect of stemming on the German language, comparing a whole range of different stemming approaches, which gave us valuable additional insights into the properties of German text retrieval. While the experiments conducted within CLEF cited above all hinted at the benefits of applying stemming and decompounding to German-language texts, it remained unclear which stemming method maximizes the benefits, and how much language-specific linguistic knowledge is necessary for effective retrieval.

In the paper, we examine five basically different stemming methods, ranging from approaches with no linguistic knowledge to a complex morphological analyzer, on the CLEF test collection. The core result is that most of these approaches give a sizeable boost to retrieval effectiveness when applying stemming (though still not always to a degree that is statistically significant), and that decompounding is even more helpful, but more difficult to implement correctly. For decompounding, we found that the correct balance must be established between splitting the right compounds for better matching, and keeping some compounds intact to reduce noise. When using stemming and decompounding in combination, we observed gains in effectiveness in terms of mean average precision of up to 60% (these gains are statistically significant).

These findings, plus the fact that other participants that obtained top performance in the multilingual track at CLEF also adapted their systems for retrieval in the individual languages, convince us that careful re-examination of monolingual retrieval for the languages involved in a CLIR system leads to a solid base for successful multilingual retrieval. The challenge remains to do this in a way that addresses the characteristics of the language itself, and not of the test collection.

2.3 Robust Multilingual Information Retrieval: "Combination Approaches for Multilingual Text Retrieval"

As mentioned, our reaction to limitations in terms of retrieval effectiveness of early CLIR approaches has been the use of combination approaches that incorporate multiple simpler methods.

Our early efforts in producing such combination methods date back to 1998 (Braschler and Schäuble, 1998; Braschler et al., 1999b). The main issue in using combination approaches is how to weigh the contribution by each of the simpler approaches that are incorporated. This problem is analogous to problems encountered earlier by IR researchers attempting to combine multiple sources of evidence for document relevance. In these studies, the corresponding problems are also referred to as "data fusion" (merging of multiple result lists calculated on the same document collection), "collection fusion" (merging of multiple result lists calculated on disjoint document collections), and "query combination" (merging of multiple query representations before retrieval). Experiments in those areas are based on the observation that the estimation of relevance of documents improves when multiple sources of evidence are available (see e.g. Saracevic and Kantor

(1988) and Belkin et al. (1993)). Early experiments in TREC conferences generally gave good results (see e.g. Belkin et al. (1994); Fox and Shaw (1994)).

We use such combination mechanisms to incorporate multiple types of translation resources in our system, and also to be able to handle multiple languages one bilingual pair at a time, and then later combine back into multilingual result lists. The latter is most similar to collection fusion, as defined above. Earlier work on the monolingual problem of merging the output from separate document collections can be found e.g. in Callan et al. (1995) and Voorhees et al. (1995), both of which also provide good introductions to the problem.

While retrieval effectiveness as evaluated using test collections in laboratory settings was a primary focus of our work, we were also keenly interested in what considerations are needed when adapting a CLIR system to generalize beyond the context of popular test collections, e.g. what happens when a system is deployed for operational use. In this context, we have identified a number of additional issues. The combination of multiple translation methods directly addresses one of the key issues in this list: user acceptance of an IR system suffers greatly from the presence of "negative outliers" in terms of retrieval quality – i.e. queries that lead to unsatisfactory results. In the CLIR context, such results are often caused by missing or incorrect translations. If we combine more than one translation method, the risk of retrieval failure is decreased. Furthermore, the lexical coverage of the system is greatly enhanced, which is essential in order to process a wide range of possible queries, including technical terminology, fashionable terms and names.

More generally speaking, we believe combination systems give us:

1. Increased lexical coverage, e.g. through the combination of several translation resources of the same type.
2. Robustness with regard to negative outliers, e.g. through the combination of multiple types of translation resources, each capable of handling different types of queries.
3. Broader applicability of the system to a range of situations, such as operational settings, e.g. through the combination of different domain-specific translation resources. We also ensure broad applicability by the avoidance of tuning that requires resources that are only available in specific situations.

We comment on each of these aspects in the paper. We demonstrate our combination approach, which can incorporate translation resources of several types (machine-readable dictionaries, machine translation output, and corpus-based approaches, such as the similarity thesaurus), both document and query translation, as well as multiple languages which are handled in the form of individual bilingual pairs.

We have used translation resources of all three above-mentioned types in the CLEF campaigns, but have recently abandoned dictionaries for our CLEF

experiments, having not been able to detect any improvements in retrieval effectiveness in an earlier evaluation campaign. We assume that the translation information provided by the general-purpose dictionaries we had available was too similar to the information provided by machine translation systems, which also incorporate such dictionaries as a core component in their translation process. Domain-specific dictionaries, however, remain essential when a CLIR system has to be adapted to domain-specific terminology by providing exact translations of the most critical terms, and our system remains prepared for their inclusion, when appropriate.

As a corpus-based technology, we use bilingual similarity thesauri, an automatically generated data structure that for every term in the source language links all terms in the target language that are statistically similar as determined from suitable training data. The monolingual version of this data structure can be used for query expansion and has been described in Qiu and Frei (1993). We build the data structure using bilingual data previously aligned on the document level (Braschler and Schäuble, 1998). Semi-automatically constructing such a "comparable corpus", which contains pairs of documents that treat roughly the same topic, is much easier than obtaining scarce parallel corpora that contain real translations of every item, especially if less frequently spoken languages are involved.

We use machine translation output for both query and document translation. During document translation, more contextual information is available for each term, which makes this an interesting alternative if translating the document collection is feasible.

In the paper, we present multiple approaches to merge the output from these different translation resources and from the individual bilingual retrieval results, among them a new method based on estimating the query coverage ("feedback merging"). We compare these merging approaches to the theoretical optimum and note that there is still ample potential to improve the effectiveness of the merging process. An important characteristic of our merging experiments is the avoidance of the use of auxiliary information that we do not expect to be available in many operational settings. Specifically, we avoid the use of large sets of training data and relevance assessments, such as are sometimes available for evaluation campaign experiments.

Our experiments in the paper are conducted using the CLEF 2000-2002 test collections, and we give specific results on how many topics (queries) were affected by the individual combinations.

2.4 Concrete Experiments with Multilingual Information Retrieval: "Eurospider at CLEF 2002" (with Anne Göhring and Peter Schäuble)

The CLEF campaign has accompanied and influenced much of our work in this dissertation, and we have used the CLEF test collections to obtain many of the results in the other papers.

We have, however, also participated directly in CLEF, in order to validate our twin goals of building an effective and robust multilingual text retrieval system. The measurement of retrieval effectiveness is one of the main concerns of CLEF, and by participating we were able to directly compare our system with the systems built by many other capable researchers in the field. However, especially in 2002, we did not put our sole emphasis on tuning the system for maximum effectiveness such as measured by CLEF, but instead tried to build the system with additional criteria that we deemed essential for operational use in mind, while not overly compromising CLEF-style retrieval performance.

We report on our experiences in this regard in the final published paper that is included with this dissertation, which is co-authored by Anne Göhring and Peter Schäuble. The paper describes our third participation in CLEF, while the two earlier efforts are reported in Braschler and Schäuble (2001) and Braschler et al. (2002).

The experiments used either a combination of document translation and two forms of query translation (QT) - machine translation (MT) and similarity thesaurus (ST) - or document translation only.

The system we used produces several intermediate search results, either due to using different translation methods, or due to handling only a subset of the five languages at a time. These intermediate search results are then "merged" into the multilingual result list necessary for submission to CLEF. For this we experimented with several merging methods, some of which we developed in the course of our CLEF participation. The system builds on the results we obtained from our experiment in monolingual retrieval, and consequently employs stemming for all languages involved (and decompounding for German). The weighting scheme was fixed to Lnu.ltn (Singhal et al., 1996).

The results of our participation were very encouraging. We were able to conclude that our combination approach is robust, with 80-90% of all queries performing on or above median performance of all systems in CLEF. This good performance was achieved even though we used no collections-specific fine-tuning for individual languages to avoid potential over-tuning to characteristics of the CLEF collection. We also did not use any collection-specific training of the merging algorithms, which we felt was not applicable to operational settings and therefore not compatible with our overall goals. It is worthy of note that, despite these restrictions, our system was third in the list of the top performing groups for the multilingual track.

3 Conclusions and Outlook

3.1 Summary

In this dissertation, we present our combination system for multilingual information retrieval, which allows users to state information needs in their preferred language, while retrieving from a document pool that contains texts in multiple languages. The

result is displayed as a unified list, containing items from all languages where appropriate.

The choice to examine combination methods was made on the basis of the observation that at the time of the start of our work, simple approaches showed limited potential for increased retrieval effectiveness. Building on our observation that different methods and translation resources produce good results for different queries, we speculated that a combination of multiple approaches should be beneficial, analogously to the earlier observation in information retrieval that multiple sources of evidence help in determining document relevance.

We have built a system that combines output from different types of translation resources (machine-readable dictionaries, output from machine translation systems, corpus-based data structures), different bilingual language pairs, and different forms of bridging the language gap (document translation and query translation). The question of how to best merge these "building blocks" remains without clear solution, and we propose and analyze several different methods.

Our methods and prototypical systems were continuously evaluated through participations in the TREC CLIR tracks and CLEF evaluation campaigns from 1998 to 2002. Our combination approach was shown to be effective, and our systems obtained results that were competitive with the top performing systems for all of our CLEF participations.

Retrieval effectiveness was not our only goal and reason for choosing a combination approach. Rather, we wanted to develop methods that generalize well beyond specific test collections. We have therefore designed our system with applications in operational settings in mind, considering additional criteria of such environments and avoiding over-tuning to specific test collections. We argue that our combination approach generalizes well, specifically by greatly increasing lexical coverage of the system, reducing the number of negative outliers, and better incorporating the necessary technical terminology. The limitations in the type of information that can be used to tune the system become especially evident when investigating the merging problem. We describe the implications that arise for our system design in our studies.

3.2 Conclusions, Comments & Contributions

Combination systems have been used in the recent CLEF campaigns with notable success, not just by us, but also by other groups. The combination of multiple methods has clearly stabilized the performance of CLIR systems and helps to provide good retrieval results for a wide range of queries. In this sense, one can argue that for pairs of widely spoken Western European languages, such as those covered at CLEF, cross-language retrieval has matured, and retrieval effectiveness in terms of recall and precision as measured in laboratory settings is satisfactory. Where at TREC-6, we observed for bilingual retrieval that "simple" approaches resulted in an effectiveness of only about 50-60% of that of comparable monolingual systems, the same figure is more than 80% for systems using well-tuned

combination methods in the latest CLEF campaign[17]. Most of the groups with top performing bilingual systems apply the same methods in the larger multilingual track, where our system consistently places among the best performers. We therefore feel that, while room for improvements remain, we have made substantial progress in our efforts to build an effective multilingual retrieval system.

We are aware that this success does not necessarily translate to any language pair, especially for languages that have characteristics that are very different from those of the languages we have been working on. New languages introduce new problems, and systems tend to score lower on recently added languages in the CLEF campaign[18]. Fewer translation tools, very different morphologies, and character coding problems are among issues that emerge when working with languages from other regions of the world. Moreover, the lack of evaluation resources, such as adequately sized test collections, hinders progress in designing effective CLIR systems for such languages. Some of the most widely spoken languages worldwide have so far received little attention by CLIR researchers (for example, languages from the Indian subcontinent). We have countered this issue by choosing methods and a framework that are open to adaptations to further languages by dynamically altering the combination depending on the resources that are available for a specific language pair. The ability to include corpus-based resources, machine-readable dictionaries and machine translation output into our set-up should provide a sound basis for good retrieval results. Resources of similar type (such as e.g. thesauri) can usually also be integrated, if it is possible to "flatten" and convert their contents to what is essentially a dictionary with associated similarity or probability values. While extensions will be necessary for new languages or very dissimilar types of translation resources, many of the core principles used for overall design should remain valid.

Cross-Language Information Retrieval remains a very active field today, as is attested by the continuously high interest in the CLEF, TREC and NTCIR campaigns. While we were more focused in our work in the combination of multiple "simple" approaches, investigating the individual approaches in the context of this overall goal, work on the individual use of the different types of translation resources, and their best application in retrieval systems continues. This can either be work that is focused on fine-tuning and improving existing mechanisms, or

---

[17] CLEF 2003 offered bilingual retrieval for selected language pairs only, which makes comparisons somewhat difficult. For the German target collection, only Finnish was allowed as a query language, which was a language that was added only recently to the CLEF campaigns, and is not yet well understood in the CLIR context. Also, Russian was only added at the last minute in CLEF 2003. We disregard those two target languages. For the remaining three target languages offered in 2003, we observe a CLIR performance in terms of mean average precision of 83% of monolingual for Spanish, of 81% of monolingual for Italian and of 82% of monolingual for Dutch, when we compare the best performing system for bilingual retrieval with the best performing monolingual system (not necessarily both from the same group) (Cross-Language Evaluation Forum, 2003)

[18] See our comments above on Finnish and Russian.

experimental work on new directions. There has been evidence of both trends in recent CLEF campaigns.

A promising development of the former is the increased research into using language models for CLIR. Language models have attracted considerable attention in the IR community recently as a very effective alternative to established retrieval models. The basic idea is to view the query as a "corrupted" version of a relevant document, and to view the retrieval system as a "decoder" which tries to reproduce the initial message, i.e. to find the document that is relevant to the query (Hiemstra, 2000). While initially proposed for monolingual retrieval (Ponte and Croft, 1998), this model is attractive for CLIR due to its inclusion of a query reformulation model, which can be used to incorporate translation probabilities, such as those coming from suitable machine-readable dictionaries (Hiemstra et al., 2001). Language models therefore provide the basis for a direct, "natural" integration of the information retrieval model and the translation model, something that is not the case for the weighting mechanisms that we use in our work.

The good performance that state-of-the-art CLIR now obtains for Western European languages, and the fact that we were able to build a competitive system while also considering additional issues that we deem essential in operational settings, is important. Because while we argue for maturity of CLIR systems in terms of retrieval effectiveness for some popular language pairs, we observe that the uptake of this technology in the commercial world is still very limited. None of the large web search engines use CLIR technology, nor does any search technology vendor aggressively market CLIR features. This runs counter to our observation that pressure to adapt CLIR technology should be increasing, with users being increasingly exposed to multilingual information. We could speculate that the right solution for the right price is still missing. Specifically, much remains to be done in terms of better understanding how the cross-language aspect influences the more complex information needs that we encounter in many operational settings, where search technology is embedded in larger solutions and has to support core work processes by accessing business-critical data. Furthermore, presentation of multilingual search results is largely an open issue, which is not actively addressed through the pursuit of better retrieval effectiveness and corresponding laboratory testing. The CLEF campaign has started to encourage research in interactive cross-language retrieval, but our understanding of how these issues are interlinked with the requirements of complex business applications remains incomplete.

In terms of work on the translation methods and retrieval methods such as we use in our system, in our opinion the challenge will be to proceed with improvements that are sustainable and not artifacts of over-tuning the systems to the test collections. As we are approaching monolingual performance, the effort to tune the algorithms will increase in relation to the resulting benefit. We are far from optimal retrieval performance, as anyone will attest that works regularly with state-of-the-art search engines, but the question now shifts to how we can derive additional information about the information need of the user, be it from the user itself or from other sources, which will help us to produce more appropriate result

sets. By optimizing retrieval effectiveness without sacrificing applicability to many settings, we hope to have contributed to a better understanding of what are the promising options for tuning.

3.3 Future Work

When we look into the future, we see an even more ambitious set of challenges. While translation methods have matured recently as measured by evaluation campaigns such as CLEF, we see potential for looking beyond the main issues reflected in CLEF topics as a way of improving CLIR systems. The fact that evaluation campaigns have to base their analysis on a limited set of topics makes it inherently hard to investigate research questions that reflect in only few or none of the specific topics at hand. We see the transliteration of person names from different writing systems as a good example of this dilemma: if only a handful of topics mention person names, and an even smaller subset of these names is transliterated, a different treatment of such features will likely lead to small effects in average performance. Furthermore, the small number of topics affected makes it also hard to derive conclusive information from a query-by-query analysis. Other examples of issues that are only reflected in few CLEF topics, but that we deem worthy of future work are issues of different formulations, different metaphors, and different regulations when documents from separate cultural areas are considered.

It remains one of our main goals to adapt our system to meet a wide range of information needs from many different application areas. This means that the issues that are "leveled out" by the measures employed at evaluation campaigns become more important. Especially names are critical entities in business processes, and search technology that aspires to support such processes needs provisions for handling them. Names by nature are often not connected to any "literal" meaning, they do not adhere to spelling rules, and thus are often not translated across languages, while they are transliterated if necessary.

Better handling of names in the CLIR context presents two main problems: firstly, how to detect names, and avoid accidental translation where applicable, and secondly, how to match names across languages, especially where transliteration occurs. The first problem is investigated in the research field of "named entity recognition" (NER), with much of the work dealing with English-language text. Lately, however, there has been increasing work on how to adapt NER systems to multilingual data collections, and how to extract bilingual lists of potential names (see e.g. Cucerzan and Yarowsky (1999) and Saito and Nagata (2003)). Work on the transliteration problem has been recently presented by Pirkola et al. (2003) and Qu et al. (2003). We are currently working on the second issue, investigating to what degree phonetic information about likely pronunciation of names in a target language helps in matching transliterated names. The basic idea is based on the observation that all the individual transliterations of a name in different languages aim to capture the "original sound" of the name. Thus, if compared phonetically, they should be similar even if their spelling is comparatively dissimilar due to differing pronunciation rules across languages. First trials indicate that it the

weighting of the phonetic information versus spelling similarity is a main problem in this line of research.

Looking further, many other business-critical entities probably need to be evaluated in the CLIR context on our way to an even more robust, versatile multilingual retrieval system. Numbers, dates, and numerous other entities all have their own small and large issues when crossing language barriers that can become critical in operational settings. Dates are written in different formats, swapping the order in which day, month and year are given. Likewise, times are given either in 12 hour or 24 hour format, and numbers use different characters for the thousands separator and the decimal point. A language-specific normalization is necessary if such entities are to be used as integral retrieval features.

With regards to work on the issue of stemming and decompounding, issues we have left to be addressed in later investigations include the weighting of components after decompounding and compound splitting based on corpus statistics. The most successful stemming components used in our study work independently of the specific documents that are indexed. Alternatively, decisions on applying stemming and decompounding to individual terms could be made based on their use in the corpus of documents being indexed. Linguistica (Goldsmith, 2001) is an example of such an approach, but it did not produce competitive results in our experiments. An alternative approach has been proposed by Agosti et al. (2003). A main motivation for these proposals was the provision of an alternative in cases where no suitable manual stemmer was available. But potential may also be found in following a slightly different line of work, where manual stemmers are enhanced by a component that modifies the stemming rules based on corpus statistics. Such a stemming approach has been presented by Xu and Croft (1998). For decompounding, there are indications that it may be beneficial to leave compounds intact that have a special meaning in the context of specific domains. Our work on the NIST decompounding component is an initial step in that direction, in that it only splits compounds that actively co-occur with their individual components. In the case of domain-specific text, such co-occurrence statistics can be expected to show different patterns.

A further area where we still see considerable potential in the development of mechanisms for CLIR and multilingual retrieval is merging, where we have demonstrated that simple methods give far from optimum performance, and more complex methods often use information that is unlikely to generalize beyond the test collections that are investigated. The incomparability of retrieval scores across separate document collections, and the different vocabularies of the individual languages handled by the system contribute to the difficulty of the merging problem. One way to address the issue is determining a mapping between the different document collections to be merged, thus allowing the comparison of retrieval scores of "similar" documents. Our earlier attempts at this showed promising results (Braschler and Schäuble, 1998), but the resulting mechanisms were ultimately too complicated. More work in this direction would have to concentrate on simplifying the approach, and specifically minimizing the amount of training data that is

necessary. A new merging alternative called "two-step RSV" was recently presented by Martínez-Santiago et al. (2003). They propose re-indexing some of the documents that are retrieved from the individual subcollections, using word-level alignments of query terms between the different languages.

Looking further ahead, inclusion of all intermediate retrieval results during merging is not necessarily the only option. Merging mechanisms could also be used to dynamically influence the combination settings themselves, e.g. excluding results derived from some types from translation resources for certain types of queries. This can be viewed as an "aggressive" merging approach, where very different weights are assigned to the individual result lists. Our feedback merging method (Braschler, 2004) is a step in this direction, as it assigns different weights for individual queries. Additionally, a query categorization mechanism could be envisioned, which could assist in this task.

We feel that a combination architecture is well suited as a sound base to incorporate new building blocks such as outlined above.

4 Acknowledgements

No work of the scope conducted for this dissertation is possible without the help of numerous others. Any attempt to list all people that have contributed directly or indirectly to my understanding of the problems described in this report, or to my development as a researcher, is by necessity incomplete.

In particular, I want to stress the contribution by three people that accompanied me on the path that led to this dissertation, in chronological order: Peter Schäuble, who at the time as an assistant professor at ETH Zurich was responsible for my diploma thesis, which was my introduction to research, and who has supported my work at Eurospider, a company which he founded, over the past six years; Donna Harman from NIST, who as the supervisor of my guest researcher's stay at her institution allowed me to conduct my first independent research and guided my first steps in that direction; and last but not least, Jacques Savoy, who generously accepted to be the supervisor of this dissertation and has given me much advice and encouragement.

This dissertation has been evaluated by a "jury de thèse" consisting of Jacques Savoy, Peter Kropf, Gareth Jones and Peter Schäuble. I want to thank them for agreeing to serve in this role, and for the comments and suggestions that have formed the basis for the revised version of this text.

This being a dissertation based on published papers, I want to specifically mention my collaborators on those papers, people with whom I had the opportunity to work closely on CLIR research problems: Carol Peters, Bärbel Ripplinger, Anne Göhring and Peter Schäuble. I have written several other papers in the same field, which were valuable steps along the way to this study, and I am equally grateful to my co-workers on those projects.

Working at Eurospider Information Technology gave me the opportunity to link my CLIR research work with work I conducted for the company, something which was essential to broaden the scope of the work to encompass issues beyond core academic questions. I would like to thank all people at the company that I have collaborated with over the years, especially those working with me in the small, but fine research group.

Lastly, no undertaking of this magnitude can be conducted solely with academic or technical help. The support of my family and my friends was and is invaluable, and I want to specifically stress the importance of the unwavering support that my parents gave me all the years up to this present day.

5 References

AAAI (1997). See http://www.glue.umd.edu/~dlrg/filter/sss/ (last visited Jan 31, 2004)

Agosti, M., Bacchin, M., Ferro, N. & Melucci, M. (2003). Improving the Automatic Retrieval of Text Documents. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Lecture Notes in Computer Science, Vol. 2785 (pp. 279-290), Springer.

Ballesteros, L. & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In N. J. Belkin, D. Narasimhalu, P. Willett (Eds.), Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 84-91).

Belkin, N. J., Cool, C., Croft, W. B. & Callan, J. P. (1993). The Effect of Multiple Query Representations on Information Retrieval System Performance. In R. Korfhage, E. M. Rasmussen, P. Willett (Eds.), Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 339-346).

Belkin, N. J., Kantor, P., Cool, C. & Quatrain, R. (1994). Combining Evidence for Information Retrieval In D. K. Harman (Ed.), The Second Text REtrieval Conference (TREC-2). NIST Special Publication 500-215 (pp. 35-43).

Braschler, M. & Schäuble, P. (1998). Multilingual Information Retrieval Based on Document Alignment Techniques. In Research and Advanced Technology for Digital Libraries, Second European Conference, ECDL '98, Lecture Notes in Computer Science, Vol. 1513 (pp. 183-197), Springer.

Braschler, M., Krause, J., Peters, C. & Schäuble, P. (1999). Cross-Language Information Retrieval (CLIR) Track Overview. In E. M. Voorhees and D. K. Harman (Eds.), Information Technology: The Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242 (pp. 1-8).

Braschler, M., Mateev, B., Mittendorf, E., Schäuble, P. & Wechsler, M. (1999b). SPIDER Retrieval System at TREC7. In E. M. Voorhees and D. K. Harman (Eds.), Information Technology: The Seventh Text REtrieval Conference (TREC-7), NIST Special Publication 500-242 (pp. 446-454).

Braschler, M., Peters, C. & Schäuble, P. (2000). Cross-Language Information Retrieval (CLIR) Track Overview. In E. M. Voorhees, D. K. Harman (Eds.), Information Technology: The Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-246 (pp. 25-33).

Braschler, M., Kan, M.-Y., Schäuble, P. & Klavans, J.-L. (2000b). The Eurospider Retrieval System and the TREC-8 Cross-Language Track. In E. M. Voorhees, D. K. Harman (Eds.), Information Technology: The Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-246 (pp. 367-375).

Braschler, M. & Schäuble, P. (2001). Experiments with the Eurospider Retrieval System for CLEF 2000. In C. Peters (Ed.), Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000. Lecture Notes in Computer Science, Vol. 2069 (pp. 140-148), Springer.

Braschler, M., Ripplinger, B. & Schäuble, P. (2002). Experiments with the Eurospider Retrieval System for CLEF 2001. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406 (pp. 102-110), Springer.

Braschler, M., Göhring, A. & Schäuble, P. (2003). Eurospider at CLEF 2002. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Lecture Notes in Computer Science, Vol. 2785 (pp. 164-174), Springer.

Braschler, M. & Ripplinger, B. (2004). How Effective is Stemming and Decompounding for German Text Retrieval?, In Information Retrieval, Volume 7, Issue 3/4, 291-306, Kluwer Academic Publishers.

Braschler, M. & Peters, C. (2004). Cross-Language Evaluation Forum: Objectives, Results, Achievements. In Information Retrieval, Volume 7, Issue 1/2, 7-31, Kluwer Academic Publishers.

Braschler, M. (2004). Combination Approaches for Multilingual Text Retrieval. In Information Retrieval, Volume 7, Issue 1/2, 183-204, Kluwer Academic Publishers.

Buckley, C., Mitra, M., Walz, J. & Cardie, C. (1998). Using Clustering and SuperConcepts within SMART: TREC 6. In E. M. Voorhess, D. K. Harman (Eds.), Information Technology: The Sixth Text Retrieval Conference (TREC-6), NIST Special Publication 500-240 (pp. 107-124).

Callan, J. P., Lu, Z. & Croft, W. B. (1995). Searching Distributed Collections with Inference Networks. In E. A. Fox, P. Ingwersen, R. Fidel (Eds.), Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 21-28).

Carbonell, J. G., Yang, Y., Frederking, R. E., Brown, R. D., Geng, Y. & Lee, D. (1997). Translingual Information Retrieval: A Comparative Evaluation (a modified version of the IJCAI'97 paper). In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97), IJCAI, 1997 (modified version at http://citeseer.nj.nec.com/carbonell97translingual.html).

Chen, A. (2003). Cross-Language Retrieval Experiments at CLEF 2002. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Lecture Notes in Computer Science, Vol. 2785 (pp. 28-48), Springer.

Chen, A. & Gey, F. C. (2004). Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding. In Information Retrieval, Volume 7, Issue 1/2, pp. 149-182, Kluwer Academic Publishers.

Cleverdon, C. W. (1967). The Cranfield tests on index language devices. Aslib Proceedings. 19, 173-192. Reprinted in (Sparck Jones and Willett, 1997)

Clough, P. & Sanderson, M. (2003). The CLEF 2003 Cross-Language Image Retrieval Task. In C. Peters, F. Borri (Eds.). Working Notes for the CLEF 2003 Workshop (pp. 379-388). See also http://www.clef-campaign.org [last visited Feb 3, 2004].

Cross-Language Evaluation Forum (2003). Results of the CLEF 2003 Cross-Language System Evaluation Campaign. In C. Peters, F. Borri (Eds.). Working Notes for the CLEF 2003 Workshop. See also http://www.clef-campaign.org [last visited Feb 3, 2004].

Cucerzan, S. & Yarowsky, D. (1999). Language independent named entity recognition combining morphological and contextual evidence. Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora (pp. 90-99).

EMIR Consortium (1994). Final Report of the EMIR Project Number 5312. For the Commission of the European Union, Brussels.

Fluhr, C., Schmitt, D., Ortet, P., Elkateb, F., Gurtner, K. & Radwan, K. (1998). Distributed Cross-Lingual Information Retrieval. In Cross-Language Information Retrieval (pp. 41-50), Kluwer Academic Publishers.

Fox, E. A. & Shaw, J. A. (1994). Combination of Multiple Searches. In D. K. Harman (Ed.), The Second Text REtrieval Conference (TREC-2), NIST Special Publication 500-215 (pp. 243-252).

Frakes, W. B. (1992). Stemming Algorithms. In (Frakes and Baeza-Yates, 1992) (pp. 131-160), Prentice Hall, Eaglewood Cliffs, NJ, USA.

Frakes, W. B. & Baeza-Yates, R. (1992). Information Retrieval. Data Structures & Algorithms. Prentice-Hall.

Gaussier, E., Grefenstette, G., Hull, D. A., Schulze, B. M. (1997). Xerox TREC-6 Site Report: Cross Language Text Retrieval. In E. M. Voorhess, D. K. Harman (Eds.), Information Technology: The Sixth Text Retrieval Conference (TREC-6), NIST Special Publication 500-240 (pp. 775-788).

Global Reach (2004). Online at http://glreach.com/globstats/

Goldsmith J (2001) Unsupervised Learning of the Morphology of a Natural Language. In Computational Linguistics, 27(2), 153-198, MIT Press. URL: http://humanities.uchicago.edu/faculty/goldsmith/Linguistica2000/ (last visited 5/27/2004)

Grefenstette, G. (1998). Cross-Language Information Retrieval. Kluwer Academic Publishers.

Harman, D. K. (1991). How Effective is Suffixing?. In Journal of the American Society for Information Science, 42(1), 7-15.

Harman, D. K. (1995). The TREC Conferences. In Hypertext - Information Retrieval - Multimedia: Synergieeffekte Elektronischer Informationssysteme, Proceedings of HIM '95 (pp. 9-28). Universitätsverlag Konstanz.

Harman, D. K., Braschler, M., Hess, M., Kluck, M., Peters, C., Schäuble, P. & Sheridan, P. (2001). CLIR Evaluation at TREC. In C. Peters (Ed.), Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069 (pp. 7-23). Springer.

Hiemstra, D. (2000). Using Language Models for Information Retrieval. CTIT PhD Thesis Series No. 01-32.

Hiemstra, D., Kraaij, W., Pohlmann, R. & Westerveld, T. (2001). Translation Resources, Merging Strategies, and Relevance Feedback for Cross-Language Information Retrieval. In C. Peters (Ed.), Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069 (pp. 102-115). Springer.

Hull, D. A. (1996). Stemming Algorithms - A Case Study for Detailed Evaluation. In Journal of the American Society for Information Science, 47(1), 70-84.

Hull, D. A. & Grefenstette, G. (1996). Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval. In H.-P. Frei, D. K. Harman, P. Schäuble, R. Wilkinson (Eds.), Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 49-57).

Jones, G. J. F. & Federico, M. (2003). Cross-Language Spoken Document Retrieval Pilot Track Report. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Lecture Notes in Computer Science, Vol. 2785 (pp. 446-457), Spinger Verlag.

Kando, N. (2003). CLIR at NTCIR Workshop 3. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Lecture Notes in Computer Science, Vol. 2785 (pp. 485-504), Spinger Verlag.

Kluck, M. & Gey, F. C. (2001). The Domain-Specific Task of CLEF – Specific Evaluation Strategies in Cross-Language Information Retrieval. In C. Peters (Ed.), Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069 (pp 48-56). Springer.

Krause, J. & Womser-Hacker, C. (1990). Das Deutsche Patentinformationssystem. Entwicklungstendenzen, Retrievaltests und Bewertungen. Carl Heymanns. [in German]

Kwok, K. L., Grunfeld, L., Dinstl, N. & Chan, M. (2001). TREC-9 Cross Language, Web and Question-Answering Track Experiments using PIRCS. In E. M. Voorhees, D. K. Harman (Eds.), Information Technology: The Ninth Text

REtrieval Conference (TREC-9), NIST Special Publication 500-249 (pp. 417-426).

Landauer, T. K. & Littman, L. M. (1990). Fully automatic cross-language document retrieval using latent semantic indexing. In Proceedings of the 6th Conference of University of Waterloo Centre for the New Oxford English Dictionary and Text Research (pp. 31-38).

Lyman, P. & Varian, H. R. (2000). How Much Information, 2000. Retrieved from http://www.sims.berkeley.edu/how-much-info on [February 12, 2004].

Lyman, P. & Varian, H. R. (2003). How Much Information, 2003. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on [February 12, 2004].

Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F. & de Rijke, M. (2003). The Multiple Language Question Answering Track at CLEF 2003. In C. Peters, F. Borri (Eds.). Working Notes for the CLEF 2003 Workshop. See also http://www.clef-campaign.org [last visited Feb 3, 2004].

Maron, M. E. & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. In Journal of the Association for Computing Machinery, Vol. 7. 216-244. Reprinted in (Sparck Jones and Willett, 1997).

Martínez-Santiago, F., Montejo-Ráez, A. & Ureña-López, L. A. (2003). SINAI at CLEF 2003: decompounding and merging. In C. Peters, F. Borri (Eds.). Working Notes for the CLEF 2003 Workshop (pp. 99-107). See also http://www.clef-campaign.org [last visited May 21, 2004].

Mateev, B., Munteanu, E., Sheridan, P., Wechsler, M. & Schäuble, P. (1998). ETH TREC-6: Routing, Chinese, Cross-Language and Spoken Document Retrieval. In E. M. Voorhess, D. K. Harman (Eds.), Information Technology: The Sixth Text Retrieval Conference (TREC-6), NIST Special Publication 500-240 (pp. 623-635).

McNamee, P. & Mayfield, J. (2002). JHU/APL Experiments at CLEF: Translation Resources and Score Normalization. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406 (pp. 193-208), Springer.

Monz, C. & de Rijke, M. (2002). Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German and Italian. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406 (pp. 262-277), Springer.

Moulinier, I., McCulloh, J. A. & Lund, E. (2001). West Group at CLEF 2000: Non-English Monolingual Retrieval. In C. Peters (Ed.), Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069 (pp. 253-260).

Nie, J. Y., Simard, M., Isabelle, P. & Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel

texts from the Web. In M. Hearst, F. C. Gey, R. Tong (Eds.), Proceedings for the 22nd Annual International Conference of the ACM-SIGIR '99 (pp. 74-81).

Nie, J. Y. & Simard, M. (2002). Using statistical translation models for bilingual IR. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406 (pp. 137-150), Springer.

Oard, D. W. (1997). Alternative approaches for cross-language text retrieval, In AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for rtificial Intelligence. Electronic working notes at (AAAI, 1997).

Oard, D. W. (1998). A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval, In Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA-98). Lecture Notes in Artificial Intelligence, Vol. 1529 (pp. 472-483), Springer.

Oard, D. W. & Hackett, P. (1998). Document Translation for Cross-Language Text Retrieval at the University of Maryland. In E. M. Voorhess, D. K. Harman (Eds.), Information Technology: The Sixth Text Retrieval Conference (TREC-6). NIST Special Publication 500-240 (pp. 687-696).

Oard, D. W., Gonzalo, J., Sanderson, M., López-Ostenero, F. & Wang, J. (2004). Interactive Cross-Language Document Selection. In Information Retrieval, Volume 7, Issue 1/2, 205-228, Kluwer Academic Publishers.

O'Neill, E. T., Lavoie, B. F. & Bennett, R. (2003). Trends in the Evolution of the Public Web 1998 – 2002. D-Lib Magazine, April 2003, Volume 9 Number 4

Peters, C. & Braschler, M. (2001) European research letter: Cross-language system evaluation: The CLEF campaigns, Journal of the American Society for Information Science and Technology, Volume 52, Issue 12, 1067-1072, John Wiley & Sons.

Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K. & Järvelin, K. (2003). Fuzzy Translation of Cross-Lingual Spelling Variants. In J. Callan, G. Cormack, C. Clarke, D. Hawking, A. Smeaton (Eds.), Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 345-352).

Ponte, J. M. & Croft, W. B. (1998). A Language Modeling Approach to Information Retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen (Eds.), Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 275-281).

Qiu, Y. & Frei, H.-P. (1993). Concept Based Query Expansion. In R. Korfhage, E. M. Rasmussen, P. Willett (Eds.), Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 160 – 169).

Qu, Y., Grefenstette, G. & Evans, D. A. (2003). Automatic Transliteration for Japanese-to-English Text Retrieval. In J. Callan, G. Cormack, C. Clarke, D. Hawking, A. Smeaton (Eds.), Proceedings of the 26th Annual International

ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 353-360).

Robertson, S. E. (1977). The probability ranking principle in IR. In Journal of Documentation, Vol. 33, 294-304. Reprinted in (Sparck Jones and Willett, 1997).

Saito, K. & Nagata, M. (2003). Multi-Language Named-Entity Recognition System based on HMM. Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition (pp. 41-48).

Salton, G. & McGill, M. J. (1983). The SMART and SIRE Experimental Retrieval Systems (pp. 118-155), McGraw-Hill.

Salton, G. (1970). Automatic processing of foreign language documents. Journal of the American Society for Information Science, 21(3), 187-194.

Saracevic, T. & Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches and overlap. In Journal of the American Society for Information Science, 39, 197-216.

Savoy, J. (2003). Report on CLEF 2002 Experiments: Combining Multiple Sources of Evidence. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum. CLEF 2002, Lecture Notes in Computer Science, Vol. 2785 (pp. 66-90), Springer.

Savoy, J. (2004). Combining Multiple Strategies for Effective Monolingual and Cross-language Retrieval. In Information Retrieval, Volume 7, Issue 1/2, 121-148, Kluwer Academic Publishers.

Schäuhle, P. (1997). Content-Based Information Retrieval from Large Text and Audio Databases. Section 1.6 Evaluation Issues.(pp. 22-29), Kluwer Academic Publishers.

Schäuble, P. & Sheridan, P. (1998). Cross-Language Information Retrieval (CLIR) Track Overview. In E. M. Voorhess, D. K. Harman (Eds.), Information Technology: The Sixth Text·Retrieval Conference (TREC-6). NIST Special Publication 500-240 (pp. 31-43).

Sheridan, P. & Ballerini, J.-P. (1996). Experiments in Multilingual Information Retrieval using the SPIDER System. In H.-P. Frei, D. K. Harman, P. Schäuble, R. Wilkinson (Eds.). Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 58-65).

Singhal, A., Buckley, C. & Mitra, M. (1996). Pivoted Document Length Normalization. In H.-P. Frei, D. K. Harman, P. Schäuble, R. Wilkinson (Eds.), Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 21-29).

Sparck Jones, K. & Willett, P. (Eds) (1997). Readings in Information Retrieval, Morgan Kaufmann Publishers.

Tomlinson, S. (2002). Stemming Evaluated in 6 Languages by Hummingbird SearchServerTM at CLEF 2001. In C. Peters, M. Braschler, J. Gonzalo, M.

Kluck (Eds.), Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406 (pp. 278-287), Springer.

Voorhees, E. M., Gupta, N. K., Johnson-Laird, B. (1995). Learning Collection Fusion Strategies. In E. A. Fox, P. Ingwersen, R. Fidel (Eds.), Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 172-179).

Voorhees, E. M. & Harman, D. K. (Eds) (1998). Information Technology: The Sixth Text Retrieval Conference (TREC-6), Appendix A (pp. A-163 – A-198). A more complete version is available online at http://trec.nist.gov/pubs/trec6/-t6_proceedings.html (last visited Feb 20, 2004).

Voorhees, E. M. (2002). The Philosophy of Information Retrieval Evaluation. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406 (pp. 355-370).

Womser-Hacker, C. (1989). Der PADOK-Retrievaltest. In Sprache und Computer, Band 10, Georg Olms Verlag. [in German]

Xu, J. & Croft, W. B. (1998). Corpus-Based Stemming using Co-occurrence of Word Variants. ACM Transactions on Information Systems, Volume 16, Issue 1, 61-81, ACM Press.

Cross-language Evaluation Forum: Objectives, Results, Achievements

Martin Braschler[1], Carol Peters[2]

[1] Eurospider Information Technology AG
Schaffhauserstr. 18, 8006 Zürich. Switzerland
Université de Neuchâtel, Institut interfacultaire d'informatique
Pierre-à-Mazel 7, CH-2001 Neuchâtel. Switzerland
martin.braschler@eurospider.com

[2] ISTI-CNR, Area di Ricerca, 56124 Pisa, Italy
carol.peters@isti.pi.cnr.it

Abstract. The Cross-Language Evaluation Forum (CLEF) is now in its fourth year of activity. We summarize the main lessons learned during this period, outline the state-of-the-art of the research reported in the CLEF experiments and discuss the contribution that this initiative has made to research and development in the multilingual information access domain. We also make proposals for future directions in system evaluation aimed at meeting emerging needs.

Keywords: evaluation campaign, cross-language information retrieval, evaluation methodology, CLIR state-of-the-art

1 Introduction

The Cross-Language Evaluation Forum is in its fourth year of activity – the results of the CLEF 2003 campaign have now been judged and will be presented in the annual workshop in August. However, the history of CLEF can be traced back to 1997 with the first track for the evaluation of cross-language systems at the Text REtrieval Conference (TREC) (Schäuble and Sheridan 1998). We are thus now in the position to attempt an assessment of the results achieved in nearly seven years of activity.

The declared high-level objectives of CLEF are threefold:

1. to provide an infrastructure for the testing and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts.
2. to construct test-suites of reusable data that can be employed by system developers for benchmarking purposes.
3. to create an R&D community in the cross-language information retrieval (CLIR) sector.

The aim of this paper will be to describe how CLEF has worked towards realizing these objectives, what has been achieved so far, and what lessons we feel have been learnt from this experience. It is often claimed (see e.g. Smeaton and Harman 1997) that evaluation campaigns really do play an active part in advancing system development. A question we will thus be asking is what has been the impact of CLEF on cross-language system development in this period. However, before we begin to assess this impact, we start by examining the status of R&D in this sector and the relationship between the research world and the application communities.

Although first experiments with cross-language information retrieval date back to the early 70s (Salton 1970), recognition of CLIR as a separate area of research probably began at SIGIR 1996 with the organisation of a Workshop on "Cross-Linguistic Information Retrieval" (Grefenstette 1998); several other activities and events followed on closely – showing that there was much interest in the research community in this "new" area. The growing popularity of the Internet and the consequent wide availability of (free) networked information sources for an increasingly vast public have clearly fuelled academic interest in CLIR. The World Wide Web phenomenon, coupled with increasing globalization of corporations and organizations, has led to strong demand for tools that permit the user to find information wherever and however it is stored, regardless of language boundaries. The demand has also been stimulated by more non-English-speaking end users going online and corporations finding themselves competing in a world-wide marketplace driven by foreign language information. There would thus appear to be a strong potential for effective and efficient systems that allow users to search document collections in multiple languages and retrieve relevant information in a form that is useful to them, even when they have little or no linguistic competence in the target languages. However, such systems are not easy to develop and although there has been much work on system development since 1996, as evidenced by the papers in this special issue, as yet there has been no strong commercial take-up. The intention of this paper is to summarize the lessons learnt from the first three CLEF evaluation campaigns (CLEF 2000 – CLEF 2002), the results of the fourth still under analysis at the time of writing, and to address the question of what remains to be done in order to bridge the gap between research and application, i.e. between system developer and system user. The paper also provides the necessary

background information for a full understanding of the experiments reported in the other papers included in this special issue dedicated to CLEF.

The rest of this paper is organised as follows. Section 2 describes the background to CLEF, from its origins to the present, whereas Section 3 presents details on the evaluation methodology, the test collections, and the techniques used for results calculation and analysis. In Section 4, we analyse the main findings from this series of campaigns and investigate whether CLEF has a direct impact on the development of the systems of the groups that participate. In the final section, we summarise our achievements so far and outline our ideas for future directions.

2 Seven Years of Activity

2.1 CLIR at TREC – 1997-1999

The surge of interest in the CLIR research area after 1996 led to the organization of a first track for CLIR system evaluation in 1997 at the TREC-6 conference (Schäuble & Sheridan, 1998). The following year, the NTCIR Workshop began cross-language evaluation for systems working on Asian languages (Kando 2003).

In the first year of the CLIR track at TREC, the participating groups performed a bilingual task – searching target document collections in French, German or English with queries formulated in a different language. Results were to be returned in the form of lists of document identifiers, ranked by decreasing probability of their relevance to the query. Thirteen groups took part, using all possible combinations of source and target languages, with the consequence that it was very difficult to compare results over different systems. The track took a big step forward in the following year with the introduction of a new task in which the systems had to search a multilingual collection of documents in four languages – the new language was Italian – for relevant items, using a single query language. Of the nine groups participating in the CLIR track in TREC-7, five groups tried this task. Its introduction forced developers to study the most appropriate indexing, transfer and retrieval strategies when handling collections in multiple languages simultaneously, ranking results across collections and languages in a single list – a complex exercise (Braschler et al. 1999). The multilingual task was repeated in 1999 with eight groups attempting it. Bilingual tasks continued to be offered as secondary exercises. An overview of the CLIR tracks at TREC is given in Harman et al. (2001).

The experience at TREC showed that an activity of this type, which involved handling, processing and understanding text in many languages, needed the collaboration of native speakers of the languages involved. For this reason, from 1998 on, the CLIR track at TREC was organised on a distributed basis, with sites in different countries being responsible for handling the work on the document collections in each language. At the end of 1999, it was decided that it would be more appropriate to centre the coordination of the evaluation activity for European languages in Europe while TREC would shift its attention to other language groups (Chinese in 2000, Arabic in 2001 and 2002). TREC, CLEF and NTCIR agreed to coordinate their activities and their schedule to avoid, as far as possible, overlapping

of dates, in order to facilitate groups that wished to participate in more than one of the cross-language campaigns.

2.2 Cross-Language System Evaluation moves to Europe

The move to Europe and the launching of an independent project, known as the Cross-Language Evaluation Forum (CLEF), has made it possible to build on and extend the results achieved within TREC. The multilingual environment provided by Europe has facilitated the addition of new languages and has stimulated participation. European businesses and institutions are accustomed to working in a multilingual context and need tools that help them to do this. The European Union currently recognises 13 official languages – this number is expected to become 25 in 2004 with the addition of ten new member states (Pieters 2002). It is one of the guiding policies of the EU that linguistic and cultural diversity should be safeguarded and access to knowledge should be guaranteed in all the languages of the Union. Support for CLEF was thus sought from the European Commission on the grounds that an initiative that involves an increasing number of European languages should be organised on the European rather than the national level[1].

The first campaigns of CLEF had the main goals:

- to accommodate as many European languages as possible;
- to encourage the participation of European groups (disappointingly low during the three years of activity coordinated in the US);
- to provide facilities for monolingual system testing and tuning in European languages other than English, which was already well covered by TREC;
- to stimulate systems to move from monolingual searching to the implementation of a full multilingual retrieval service
- to study the needs of both system developers and system users in order to promote the introduction of new tasks, designed to meet newly identified requirements.

The results have been encouraging from the start. Separate tracks to test monolingual, bilingual and multilingual systems were provided with the aim of allowing groups to work their way up gradually from mono- to multilingual retrieval. Additional tracks have been added to supplement the core tracks. The test collection has continued to grow. Participation of both academic and industrial groups, and especially of European groups, has increased rapidly (see Table 1). In

[1] CLEF 2000 and CLEF 2001 were sponsored by the European Commission under the Information Society Technologies programme and within the framework of the DELOS Network of Excellence for Digital Libraries (IST-1999-12262); CLEF 2002 and 2003 have been funded as an independent project (IST-2000-31002). The consortium members are ISTI-CNR, Italy (coordinators); ELDA, France; Eurospider Information Technology, Switzerland; IZ-Bonn, Germany; LSI-UNED, Spain; NIST, USA.

the following section, we describe the organization of these evaluation campaigns and explain the underlying motivations for certain choices and decisions.

Table 1. Growth in number of participants and experiments over the years.

Year	# Participants (for all tracks)	# Participants (core tracks only)	# Experiments (core tracks only)
TREC-6 (1997)	13	13	(95)[2]
TREC-7 (1998)	9	9	27
TREC-8 (1999)	12	12	45
CLEF 2000	20	20	95
CLEF 2001	34	31	198
CLEF 2002	37	34	282
CLEF 2003	42	33	415

3 Organisation of the CLEF Evaluation Campaign

CLEF aims at providing a forum that is not so much a competition, but rather a place where people can objectively evaluate their systems and ideas. It is left open to the participants whether they want to perfect a tried-and-true approach, or try "revolutionary" new ideas, even though some of these may quickly disappear. In this sense, the intentions of participants, and the amount of effort they invest in their CLEF experiments differ vastly. The campaign as a whole benefits from this fact, as newcomers and creative, "daring" groups find as much a place for their work as veteran participants that have built elaborate, finely tuned systems.

3.1 Comparative Evaluation and the Cranfield Paradigm

For the core activities, CLEF adopts a corpus-based, automatic scoring method, based on ideas first introduced in the Cranfield experiments (Cleverdon 1997) in the late 1960s. This methodology is widely used and accepted in the information retrieval community. Its properties have been thoroughly investigated and are well understood. This approach is also used by the popular series of TREC conferences (Harman 1995), which are the "gold standard" for this form of evaluation campaigns. "End-users" are not directly involved in the evaluation when following this "Cranfield" paradigm. While user-based evaluation aims to directly measure the user's satisfaction with a particular system, the CLEF core activities have been limited to system evaluation. One of the reasons for adopting system evaluation lies

[2] In TREC6, only bilingual retrieval was offered, which resulted in a large number of runs combining different pairs of languages (Schäuble and Sheridan 1998). Starting with TREC7, multilingual runs were introduced (Braschler et al. 1999), which usually consist of multiple runs for the individual languages that are merged. The number of experiments for TREC6 is therefore not directly comparable to later years.

in the costs and complexities incurred by conducting user-based evaluations. While much is to be said for involving end users in the evaluation of systems (and the "interactive track" of CLEF indeed tries to work in this direction), besides significantly lowering the cost of conducting the evaluation, abstracting the evaluation process can help to control some of the parameters that affect retrieval performance, and thus may increase the power of comparative experiments. In CLEF, following the Cranfield paradigm, good system performance is equated with good retrieval effectiveness (in terms of returning lists of documents). There are a host of assumptions underlying the "laboratory setting" used for system evaluation in CLEF, and we will briefly discuss some of their implications. For a more detailed discussion of the Cranfield paradigm see (Voorhees 2002).

The CLEF campaigns use a combination of a set of retrievable documents, a set of formulations of "information needs" and measures for evaluating the effectiveness of the system in answering these information needs on the basis of the set of documents. The measures used by CLEF, mainly recall and precision[3], are based on the "relevance" of a document to the corresponding information need, which is defined on topical similarity as determined by expert relevance assessors. For the sake of practicality, CLEF must make multiple assumptions with regard to the concept of relevance it adopts: first, that all "relevant documents" contribute equally to the performance measures, second, the relevance of a document is independent of the relevance of other documents, and third, that all potential users agree on the relevance of a document with respect to an information need. For some performance measures, such as recall, it is further assumed that all relevant documents in the collection are known. Relevance assessments were chosen to be binary: a document is either relevant or irrelevant. CLEF emphasizes comparative evaluation. An important consequence of the abstractions used in its methodology is that absolute scores of individual experiments are not meaningful in isolation.

Of course, the distinguishing feature of CLEF is that it applies this evaluation paradigm in a multilingual setting. This means that the criteria normally adopted to create a test collection consisting of suitable documents, sample queries and relevance assessments must be adapted to satisfy the particular requirements of this context. In the rest of this section, we will thus discuss how the CLEF evaluation campaigns have been designed and set up to meet the special needs of multilingual information retrieval. In Section 3.2 we describe how the evaluation tasks have been studied to meet the needs of the multilingual system developers community, and in Section 3.3 we describe the measures taken to create a test collection that can be used in a multilingual, multicultural context and can be trusted to give consistent, unbiased results. This is particularly important with respect to the relevance assessments whose objectivity must be guaranteed. Finally, in Section 3.4, we comment on the completeness of the relevance assessments as used in CLEF.

---

[3] Recall measures the proportion of relevant documents retrieved with respect to the total number of relevant documents in the entire document collection. Precision measures the proportion of relevant documents retrieved with respect to the total number of documents retrieved from the collection.

3.2 Evaluation Tracks

Over the years, the range of activities offered to participants of the initial TREC CLIR and subsequent CLEF campaigns has expanded and been modified in order to meet the needs of the research community. Consequently, the campaigns are structured into several distinct tracks. Some of these tracks are in turn structured into multiple tasks. Almost from the very beginnings in TREC, the main focus of the campaigns has been the multilingual retrieval track, in which systems must use queries in one language to retrieve items from a test collection that contains documents written in a number of different languages (four in CLEF 2000, five for CLEF 2001 and 2002, and either four or eight in CLEF 2003). Participants are actively encouraged to work on this, the hardest task offered. A stated goal of the CLEF campaign is to allow groups to gradually move from "easier" tracks/tasks in their first participation to eventually work with as many languages as possible, joining the multilingual track. To this end, bilingual and monolingual tracks are also offered. These smaller tracks serve additional important purposes, for example in terms of helping to better understand the characteristics of individual languages, and to fine-tune procedures.

In CLEF, we distinguish between the *core tracks*, which are those that are offered regularly each year (the monolingual, bilingual, multilingual and domain-specific tracks) and *additional tracks*, which tend to be organized on a more experimental basis and have the objective of identifying new requirements and appropriate methodology for their testing. The core tracks are coordinated by the members of the CLEF consortium, whereas the additional tracks are organized by interested associated groups, under the CLEF umbrella, on a voluntary basis. Here below we describe the tracks and tasks offered by the campaigns from CLEF 2000 through CLEF 2003. For each of the core tracks, the participating systems construct their queries (automatically or manually) from a common set of statements of information needs (known as topics) and search for relevant documents in the collections provided, listing the results in a ranked list. With the exception of the paper by Oard et al which discusses work done in the Interactive CLEF track, the articles in this special issue describe experiments on the core tracks.

CORE TRACKS

Multilingual Information Retrieval. This is the main task in CLEF. It requires searching a multilingual collection of documents for relevant items, using a selected query language. Multilingual information retrieval is a complex task, testing the capability of a system to handle a number of different languages simultaneously, ordering them according to relevance. In CLEF 2000, the multilingual collection for this track contained English, German, French, and Italian documents. In CLEF 2001 and 2002, it also contained Spanish texts. For CLEF 2003, two distinct multilingual tasks were offered: multilingual-4 and multilingual-8. The collection for multilingual-4 contained English, French, German and Spanish documents. Multilingual-8 involved searching a collection containing documents in eight languages: Dutch, English, Finnish, French, German, Italian, Spanish and Swedish.

For each campaign, a common set of topics (i.e. structured statements of information needs from which queries are extracted) has been prepared in up to twelve languages: Dutch, English, Finnish, French, German, Italian, Spanish, Swedish, Russian, Portuguese, Japanese and Chinese. Topics have been offered in other languages on demand.

Bilingual Information Retrieval. In this track, any query language can be used to search a single target document collection. Many newcomers to CLIR system evaluation prefer to begin with the simpler bilingual track before moving on to tackle the more complex issues involved in truly multilingual retrieval. CLEF 2000 offered the possibility for a bilingual search on an English target collection, using any other language for the queries. CLEF 2001 offered two distinct bilingual tracks with either English or Dutch target collections. In response to considerable pressure from the participants, in CLEF 2002 we decided to extend the choice to all of the target document collections, with the single limitation that only newcomers to a CLEF cross-language evaluation task could use the English target document collection. This decision had the advantage of encouraging experienced groups to experiment with "different" target collections, rather than concentrating on English, but it had the strong disadvantage that the results were harder to assess in a comparative evaluation framework. There were simply too many topic-target language combinations, only receiving a few experiments each. Consequently, for CLEF 2003, we offered a very different choice. The main objective in the 2003 bilingual track was to encourage the tuning of systems running on challenging language pairs that do not include English, but also to ensure comparability of results. For this reason, runs were only accepted for one or more of the following source → target languages pairs: Italian → Spanish, German → Italian, French → Dutch, Finnish → German. Newcomers only (i.e. groups that have not previously participated in a CLEF cross-language task) could choose to search the English document collection using a European topic language. At the last moment, we acquired a Russian collection and thus also included Russian as a target collection in the bilingual task, permitting any language to be used for the queries.

Monolingual (non-English) IR. Until recently, most IR system evaluation focused on English. However, many of the issues involved in IR are language dependent. CLEF provides the opportunity for monolingual system testing and tuning, and for building test suites in other European languages apart from English. Each of the CLEF campaigns has thus provided the opportunity for monolingual system testing and tuning on any of the target collections available, with the exception of the English collection.

Domain-Specific Mono- and Cross-Language Information Retrieval. The rationale for this task is to study CLIR on other types of collections, serving a different kind of information need. The information that is provided by domain-specific scientific documents is far more targeted than news stories and contains much terminology. It is claimed that the users of this type of collection are typically interested in the completeness of results. This means that they are generally not satisfied with finding just some relevant documents in a collection that may contain

many more. Developers of domain-specific cross-language retrieval systems need to he able to tune their systems to meet this requirement. See (Kluck and Gey 2001) for a discussion of this point.

ADDITIONAL TRACKS

Interactive CLIR (iCLEF). The aim of the tracks listed so far is to measure system performance mainly in terms of its effectiveness in document ranking. However, this is not the only issue that interests the user. User satisfaction with an IR system is based on a number of factors, depending on the functionality of the particular system. For example, the ways in which a system can help the user when formulating a query or the ways in which the results of a search are presented are of great importanec in CLIR systems where it is common to have users retrieving documents in languages with which they are not familiar. An interactive track that has focused on both user-assisted query formulation and document selection has been implemented with success since CLEF 2001 (see Oard et al. 2004).

Multilingual Question Answering (QA at CLEF). This is a completely new track introduced for the first time at CLEF 2003. It consists of several tasks and offers the possibility to test monolingual question answering systems running on Spanish, Dutch and Italian texts, and cross-language systems using questions in Dutch, French, German. Italian and Spanish to search an English document collection. This track is an important innovation for CLEF as successful question answering systems need to integrate IR technology together with sophisticated natural language processing (NLP) procedures. The aim of this track is both to stimulate monolingual work in the question answering area on languages other than English and to encourage the development of the first experimental systems for cross-language QA.

Cross-Language Spoken Document Retrieval (CL-SDR). The current growth of multilingual digital material in a combination of different media (e.g. image, speech, video) means that there is an increasing interest in systems capable of automatically accessing the information available in these archives. For this reason, the DELOS Network of Excellence for Digital Libraries[4] supported a preliminary investigation aimed at evaluating systems for cross-language spoken document retrieval in 2002. The aim was to establish baseline performance levels and to identify those areas where future research was needed. The results of this pilot investigation were first presented at the CLEF 2002 Workshop and are reported in Jones and Federico (2003). Cross-language spoken document retrieval has been offered as a pilot experiment in CLEF 2003.

Cross-Language Retrieval in Image Collections (Image CLEF). This track was offered for the first time in CLEF 2003 as a pilot experiment. The aim is to test the effectiveness of systems to retrieve as many relevant images as possible on the basis of an information need expressed in a language different from that of the document collection. Queries were made available in five languages (Dutch, French, German,

[4] More information on DELOS can be found at http://delos-noe.iei.pi.cnr.it/

Italian and Spanish) to search a British-English image collection. Searches could make use of the image content, the text captions or both.

These last two tracks show the effort that is now being made by CLEF to progress from text retrieval tasks to tasks that embrace multimedia.

3.3 The Test Collections

The CLEF test collection for the core tracks is formed of sets of documents in different European languages but with common features (same genre and time period, comparable content); a single set of topics rendered in a number of languages; relevance judgments determining the set of relevant documents for each topic. A separate test collection has been created for systems tuned for domain-specific tasks.

3.3.1 Document collections. The main document collection now consists of well over 1.5 million documents in nine languages – Dutch, English, Finnish, French, German, Italian, Russian, Spanish and Swedish. This collection has been expanded gradually over the years. The 2000 collection consisted of newspaper, news magazine and news agency articles mainly from 1994 in four languages: English, French, German and Italian. Two languages were added in 2001: Spanish because of its global importance, and Dutch, partly to meet the demands of a considerable number of Dutch participants – a very active community in early CLEF, but also because it provided a challenge for those who wanted to test the adaptability of their systems to a new, less widespread language. Swedish and Finnish were introduced for the first time in CLEF 2002 for different reasons. Swedish was chosen as a representative of the Nordic languages, whereas Finnish was included both because it was a representative of a different language group (the Uralic languages) and also because its complex morphology makes it a particularly challenging language from the text processing viewpoint. Russian was an important addition in 2003 as it is the first collection in the CLEF corpus that does not use the Latin-1 (ISO-8859-1) encoding system.

The domain-specific collection consists of the GIRT database of German social science documents, with a controlled vocabularies for English-German and German-Russian. The GIRT texts were first used in the TREC CLIR tracks and have been expanded for CLEF. In 2003 a third, even more extensive version of the GIRT database was introduced, consisting of more than 150,000 documents in an English-German parallel corpus. In CLEF 2002, this track also used the Amaryllis scientific database of approximately 150,000 French bibliographic documents, and a controlled vocabulary in English and French.

Table 2. Sources and dimensions of the main CLEF 2003 document collection

Collection	Added in	Size (MB)	No. of docs	Median size of docs. (Bytes)	Median size of docs. (Tokens[5])	Median size of docs. (Features)
Dutch: Algemeen Daghlad 94/95	2001	241	106483	1282	166	112
Dutch: NRC Handelsblad 94/95	2001	299	84121	2153	354	203
English: LA Times 94	2000	425	113005	2204	421	246
English: Glasgow Herald 95	2003	154	56472	2219	343	202
Finnish: Aamulehti late 94/95	2002	137	55344	1712	217	150
French: Le Monde 94	2000	158	44013	1994	361	213
French: ATS 94	2001	86	43178	1683	227	137
French: ATS 95	2003	88	42615	1715	234	140
German: Frankfurter Rundschau 94	2000	320	139715	1598	225	161
German: Der Spiegel 94/95	2000	63	13979	1324	213	160
German: SDA 94	2001	144	71677	1672	186	131
German: SDA 95	2003	144	69438	1693	188	132
Italian: La Stampa 94	2000	193	58051	1915	435	268
Italian: AGZ 94	2001	86	50527	1454	187	129
Italian: AGZ 95	2003	85	48980	1474	192	132
Russian: Izvestia 95[6]	2003	68	16761			
Spanish: EFE 94	2001	511	215738	2172	290	171
Spanish: EFE 95	2003	577	238307	2221	299	175
Swedish: TT 94/95	2002	352	142819	2171	183	121

SDA/ATS/AGZ = Schweizerische Depeschenagentur (Swiss News Agency)
EFE = Agencia EFE S.A (Spanish News Agency)
TT = Tidningarnas Telegrambyrå (Swedish newspaper)

[5] The number of tokens extracted from each document can vary slightly across systems, depending on the respective definition of what constitutes a token. Consequently, the number of tokens and features given in this table are approximations and may differ from actual implemented systems.

[6] Figures for Russian are not comparable due to different encoding system.

Table 2 gives details of the source and dimensions of the main multilingual document collection used in CLEF. Most papers in this special issue make reference to the collection as of 2002. The table gives the overall size of each subcollection, the year of its addition, number of documents contained, and three key figures indicating some typical characteristics of the individual documents: the median length in bytes, tokens and features. Tokens are "word" occurrences, extracted by removing all formatting, tagging and punctuation, and the length in terms of features is defined as the number of distinct tokens occurring in a document.

3.3.2 Topics. For the core tasks, the participating groups derive their queries in their preferred language from a set of topics that were created to simulate user information needs. Following the TREC philosophy, each topic consists of three parts: a brief title statement; a one-sentence description; a more complex narrative specifying the relevance assessment criteria. The title contains the main keywords, the description is a "natural language" expression of the concept conveyed by the keywords, and the narrative adds additional syntax and semantics, stipulating the conditions for relevance assessment. Queries can be constructed from one or more fields. Here below we give the English version of a typical topic from CLEF 2002.

<top>

<num> C091 </num>

<EN-title> AI in Latin America </EN-title>

<EN-desc> Amnesty International reports on human rights in Latin America. </EN-desc>

<EN-narr> Relevant documents should inform readers about Amnesty International reports regarding human rights in Latin America, or on reactions to these reports. </EN-narr>

</top>

The motivation behind using structured topics is to simulate query "input" for a range of different IR applications, ranging from very short to elaborate query formulations, and representing keyword-style input as well as natural language formulations. The latter potentially allows sophisticated systems to make use of morphological analysis, parsing, query expansion and similar features. In the cross-language context, the transfer component must also be considered, whether dictionary or corpus-based, a fully-fledged MT system or other. Different query structures may be more appropriate for testing one or the other methodology.

The creation of a topic set in a multilingual context necessitates a very rigorous procedure in order to ensure consistency and coherency of the topic sets in the different languages. CLEF topics are developed on the basis of the contents of the multilingual document collection. For each language, native speakers propose a set

of topics covering events of local, European and general importance. The topics are then compared over the different sites to ensure that a high percentage of them will find some relevant documents in all collections, although the ratio can vary considerably. The fact that the same topics are used for the mono-, bi-, and multilingual tracks is a significant constraint[7]. While in the multilingual task it is of little importance if a given topic does not find relevant documents in all of the collections, in both the bilingual and monolingual tracks, where there is a single target collection, a significant number of the queries must retrieve relevant documents.

Other criteria must be met to provide a full range of cross-language testing possibilities for the participating systems. For example, it is important to include names of locations (translatable or not), of people (where some kind of robust matching may be necessary, e.g. Eltsin, Ieltsin, or Yeltsin), important acronyms, some terminology (testing lexical coverage), syntactic and semantic equivalents. The goal is to achieve a natural, balanced topic set accurately reflecting real world user needs while at the same time testing a system's processing capabilities to the full.

Once the topics have been selected, they are prepared in all the collection languages by skilled translators translating into their native language. They can then be translated into additional languages, depending on the demand from the participating systems. In all cases, the aim is to produce natural language renderings of the concepts expressed rather than literal translations.

The main CLEF topic set currently consists of 200 topics for eight different languages (Dutch, English, Finnish, French, German, Italian, Spanish, Swedish), and a subset of them for additional topic languages (Chinese, Japanese, Portuguese, Russian, Thai). 40 topics were created for CLEF 2000, 50 topics each for 2001 and 2002, and 60 for CLEF 2003. Separate topic sets have been developed for the GIRT task in German, English and Russian and in French and English for Amaryllis. The CLEF topic generation process and the issues involved are described in detail in Womser-Hacker (2002) and Mandl and Womser-Hacker (2003).

The size of the topic set in each campaign is dictated by limited evaluation resources. With the kind of effort that is possible within the CLEF campaigns, it is impractical to produce exhaustive relevance assessments for larger topic sets. Using such a topic set as a (small) sample of all potential information needs that end users might have has implications on the interpretation of the experimental results. Indeed, statistical analysis of the results in the multilingual tracks of the 2001 and 2002 campaigns has shown that performance changes need to be rather large to find statistically significant differences between experiments (Braschler 2002, 2003). The best remedy is to use a larger topic set for experiments, which CLEF facilitates by keeping portions of the document collection unchanged from campaign to campaign.

[7] This is essential for economic reasons and important in order to create a reliable test collection. Relevance assessment is a very resource consuming task. By using the same topics, the relevance assessment for the three tracks can be done simultaneously and the document pools per topic for each language collection are sufficiently large.

effectively allowing post-campaign experiments to use multiple topic sets from several campaigns. A recent discussion of the influences of different topic set sizes can be found in Voorhees and Buckley (2002).

3.3.3 Relevance Judgments. The relevance assessments are produced in the same distributed setting and by the same groups that work on the topic creation. CLEF uses methods adapted from TREC to ensure a high degree of consistency in the relevance judgments. All assessors follow the same criteria when judging the documents. An accurate assessment of relevance of retrieved documents for a given topic implies a good understanding of the topic. This is much harder to achieve in the distributed scenario of CLEF where understanding is influenced by language and cultural factors. Rules are established to ensure, as far as possible, that the decisions taken as to relevance are consistent over sites, and over languages.

The practice of assessing the results on the basis of the "Narrative" means that only using the "Title" and/or "Description" parts of the topic implicitly assumes a particular interpretation of the user's information need that is not (explicitly) contained in the actual query run in the experiment. The fact that the information contained in the title and description fields could have additional possible interpretations has influence only on the absolute values of the evaluation measures, which in general are inherently difficult to interpret. However, comparative results across systems are usually stable when considering different interpretations. These considerations are important when using the topics to construct very short queries to evaluate a system in a web-style scenario.

The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead, approximate recall figures are calculated by using pooling techniques. The results submitted by the participating groups are used to form a "pool" of documents for each topic and for each language by collecting the highly ranked documents from all the submissions. The implications of using this pooling procedure on the validity of the results published by CLEF are discussed in the following section. Table 3 gives an overview of the number of topics and documents used in the campaigns for the core tracks, and of the size of the corresponding document pools.

Table 3. Number of documents assessed for the CLEF campaigns (core tracks).

Collection	# Lang.	# Docs.	Size in MB	# Docs. assessed	# Topic.	# Assessed/ topic
CLEF 2003	9	1,611,178	4124	~188,000	60*	~3100
CLEF 2002	8	1,138,650	3011	140,043	50*	~2900
CLEF 2001	6	940,487	2522	97,398	50	1948
CLEF 2000	4	368,763	1158	43,566	40	1089

* only 30 topics assessed for Finnish in 2002 and only 37 topics assessed for Russian in 2003.

3.4 Results Analysis

As was mentioned when introducing the Cranfield paradigm used by the CLEF campaigns in Section 3.2, performance measures reported for the CLEF experiments depend heavily on relevance assessments. Using a pooling methodology, i.e. only judging documents for relevance that have been highly ranked in at least one of the experiments considered, while essential to make relevance assessment feasible for large collections, incurs the risk that a substantial portion of relevant documents goes undetected, i.e. that the relevance assessments are not sufficiently complete. Since such a situation would cast serious doubts on the validity of the conclusions derived on the basis of the CLEF results, tests investigating pool quality have been run since the first CLEF campaign in 2000. The situation is particularly critical with respect to the reusability of the test collections produced by CLEF: an incomplete pool may put experimenters whose systems did not contribute to the pool during the campaigns at a disadvantage.

We have adopted an idea originally put forward by Zobel (1998): to get an indication of the completeness of the pool, individual participants are in turn removed from the pool, each time re-evaluating results. If results remain stable throughout this process, evidence is gained that the pool is sufficiently complete so that new participants would add little in terms of more relevant documents to the pool. That is, the pool contains the vast majority of relevant documents. We have calculated the mean and maximum difference observed when using reduced relevance assessments for the multilingual document pool for all three campaigns held to date. For 2002, we have also calculated the respective numbers for the subcollections formed by the individual languages. In all cases, we have found that the CLEF collections compare favourably to other test collections produced by similar campaigns, with a maximum difference for the multilingual experiments of 5.99% observed for the 2000 campaign, which has been steadily lowered to 1.76% in 2002 thanks to even larger pools based on more different systems and experiments being used (Table 4). The mean differences are in the order of 1% or less, meaning that conclusions with respect to one system outperforming another system are only affected in cases where differences are too small to be of any likely statistical significance. For the subcollections formed by individual languages, somewhat larger differences are observed, mainly for the newer languages that have been introduced in the later campaigns (see Table 5). This is to be expected due to fewer participants and systems contributing to the respective pools. Still, the numbers are sufficiently favourable to make it unlikely that conclusions based on the test collections are invalidated, provided that care is taken in considering the inherent limitations of system evaluations such as those conducted by CLEF.

Table 4. Multilingual track of the CLEF 2000, CLEF 2001 and CLEF 2002 campaigns. Key values of the pool quality analysis: mean and maximum change in average precision when removing the pool contribution of one participant, and associated standard deviation.

Campaign	Mean difference	Max. difference	Std. dev. difference
CLEF 2000	0.0013 (0.80%)	0.0059 (5.99%)	0.0012 (1.15%)
CLEF 2001	0.0023 (1.02%)	0.0076 (4.50%)	0.0051 (2.37%)
CLEF 2002	0.0008 (0.48%)	0.0030 (1.76%)	0.0018 (1.01%)

Table 5. CLEF 2002: subcollections for individual languages. Key values of the pool quality analysis: mean and maximum change in average precision when removing the pool contribution of one participant, and associated standard deviation.

Track	Mean Difference	Max. Difference	Std. Dev. Difference
DE German	0.0025 (0.71%)	0.0095 (5.78%)	0.0054 (1.71%)
EN English	0.0023 (1.14%)	0.0075 (3.60%)	0.0051 (2.60%)
ES Spanish	0.0035 (0.87%)	0.0103 (2.52%)	0.0075 (1.86%)
FI Finnish	0.0021 (0.82%)	0.0100 (4.99%)	0.0049 (2.05%)
FR French	0.0019 (0.54%)	0.0050 (1.86%)	0.0038 (1.08%)
IT Italian	0.0008 (0.22%)	0.0045 (0.93%)	0.0016 (0.46%)
NL Dutch[8]	0.0045 (1.26%)	0.0409 (9.15%)	0.0116 (3.09%)
SV Swedish	0.0082 (3.32%)	0.0306 (10.19%)	0.0182 (7.51%)

3.5 Workshops

Each CLEF evaluation campaign culminates in a two-day workshop. The objective of the workshops is to bring together the groups that have participated in that year's campaign so that they can report on the results of their experiments. These workshops are an essential part of the CLEF experience: they provide the opportunity for researchers working on common problems to get together and exchange ideas and opinions which not only regard current approaches and techniques but also future directions for research in this field. Participants also have the chance to make proposals for new tasks to be introduced in future campaigns. The workshops play a strong role in the creation of a CLIR research community around the CLEF activity. It is easy to witness their impact on successive campaigns as we see groups experimenting with approaches they have seen presented in previous years and also sharing tools and resources. Copies of the Working Notes and the presentations given at the workshop are made publicly available on the CLEF website.

[8] One experiment that was an extreme outlier in terms of performance was removed before calculation of the Dutch figures to avoid a non-representative skew in the numbers.

4 Current State of CLEF systems

In the three CLEF campaigns that have been completed to date (and in the three CLIR tracks held at the TREC conference before that), numerous ideas and methods have been used by the participants. Some of these methods have quickly faded away after one campaign, whereas others have been eagerly adopted and expanded on by other participants. This sharing of ideas is a very important aspect of the CLEF activities, which we will expand on later in this section. We start with the perhaps rather ambitious goal of describing the blueprint for "a successful CLEF system" - of course any such attempt is by necessity limited by what has been learnt by the participants in the campaigns that have taken place so far. Even so, the sharing of ideas implies a certain amount of convergence toward systems that use one of the (or the only) blueprints that have proven successful so far. Clearly, such convergence could lead to the risk of a "monoculture" of CLIR systems; this evidences the value of participants that dare to "think outside the box" and try new approaches. Luckily, the CLEF campaign has seen a number of experiments by such participants. We think this may be helped by the fact that CLEF is set up not to be a purely competitive forum, but rather as a place for both introducing and fine-tuning ideas.

4.1 A Successful Blueprint for a Multilingual Retrieval System

When analyzing the results of the CLEF 2002 campaign (the CLEF 2003 campaign has not yet concluded), it becomes apparent that there are strong similarities between the systems of the three participants that submitted the top performing experiments for the multilingual retrieval task. Our interpretation of this fact is that through participation in earlier campaigns and learning from the experiences reported from them, these three groups have found a "blueprint" for what constitutes – in terms of the state-of-the-art – a type of CLIR system that is successful for the CLEF task. We will now try to outline our understanding of this blueprint. These three systems have been built by Université de Neuchâtel (Savoy 2004), University of California at Berkeley Group 2 (Chen and Gey 2004) and Eurospider (Braschler 2004) and are each described in articles included in this special issue. All three systems are based on a strong foundation in monolingual retrieval for some or all of the languages in the multilingual track in CLEF 2002 (English, French, German, Italian, Spanish). This does not necessarily mean that an extraordinary amount of linguistic knowledge or language-specific processing is used, but that the methods employed for monolingual retrieval are robust and well-tuned, leading to performance in the monolingual retrieval tracks by these groups which either outperforms other participants or is very close to top performance. All three groups used stemming for all five languages and decompounding for German words. However, they did not use sophisticated morphological analyzers or tools such as part-of-speech taggers. For term weighting, well-known weighting schemes that have previously been shown to be successful in English language evaluations such as TREC were applied (BM25, Lnu.ltn, Berkeley ranking). Blind feedback (i.e. automatic query expansion by terms collected from the documents ranked at the top after initial retrieval) was used by all three groups. This "formula" for monolingual retrieval (robust stemming,

well-known weighting schemes, and blind feedback) was only outperformed for Italian by two Italian groups (Amati et al. 2003, Bertoldi and Federico 2003), which may hint at potential for some more language-specific fine-tuning.

In terms of strategies used to cross the language barrier, the approaches of these three groups also show parallels. All use the combination of more than one type of translation resource, and in some cases also more than one translation resource of the same type. Université de Neuchâtel uses a range of machine translation systems, supplemented by an electronic dictionary. UC Berkeley also uses a dictionary, plus some of the same machine translation systems, but combines these resources with a corpus-based translation resource built from parallel texts. Eurospider combines machine translation using different systems with a "similarity thesaurus" derived from suitable training data. All three systems use query translation, although Eurospider also combines this with document translation through machine translation.

For multilingual retrieval, several alternatives for the handling of all the languages exist. They can be handled simultaneously, or they can be handled one at a time, through a succession of bilingual retrieval steps, and then subsequently merged into one, multilingual result. All three groups have used the latter approach for query translation (when using document translation, this problem does not arise, since retrieval on the translated document collection is monolingual). However, the effective merging of the various bilingual results has proven to be a difficult challenge for all participants in the last two campaigns, and it appears that no robust, well-performing methods have been found. While all use several (mostly simple) methods for merging, the merging performance of these three groups appears to be still far below the theoretical optimum.

In summary, and looking at these three systems, we conclude that one possible blueprint for building a system that is successful for the CLEF 2002 multilingual track could well consist of:

- effective monolingual retrieval for most or all of the languages involved. This is achieved through use of robust stemming, well-known weighting schemes and blind feedback.
- a combination of translation information derived from multiple types of translation resources. The right parameterization and combination of these elements leads to effective retrieval results on the CLEF test collection.
- using query translation for a series of bilingual retrieval steps for individual language pairs. The results are then merged into one, multilingual result set.

A total of eleven groups submitted experiments to the multilingual track in CLEF 2002, and many alternative ideas to the ones outlined above were proposed, some radically different, and also successful. We will address some of these approaches in the following sections. Even so, elements of the same blueprint can be detected in some further experiments by other participants in both the multilingual and bilingual tracks, such as Océ (Brand and Brünner 2003), University of Exeter

(Lam-Adesina and Jones 2003) and others. The challenge remains to prove that the success of this specific blueprint is generalizable to other test collections and operational settings.

4.2 Other Approaches

In 2000, combination systems, i.e. systems combining more than one type of translation resource, were already used to some extent by groups from Eurospider (Braschler and Schäuble 2001), from Johns Hopkins University (JHU-APL) (McNamee et al. 2001), the Twenty-One group at TNO-TPD and University of Twente (Hiemstra et al. 2001) and the Laboratoire RALI at Université de Montreal (Nie et al. 2001). However, in general, combination approaches were not widely used during the first CLEF campaign in 2000, indicating that participants have adopted them over the years on the basis of experiences reported on in earlier campaigns. The typical CLEF 2000 system was based on only one type of translation resource, most likely machine-readable dictionaries (MRD). There was considerable work on using corpus-based techniques for disambiguating different translation alternatives, but less effort on using such corpus-based techniques directly for translation. Then, as now, query translation was by far the preferred approach. In 2001, a massive increase in interest in corpus-based approaches for translatinn could be observed (This, and other fluctuations in the type of resources used hy participants for their experiments, is documented in Section 4.5 of this paper). The number of groups using some type of corpus-based resource, either exclusively or in combination with other alternatives, douhled. Even so, machine-readahle dictionaries remained the most popular choice, which we believe has a lot to do with newcomers often adopting them as their first choice when starting out. Indeed, the fluctuation of groups using dictionary-based approaches is unusually high compared to the alternatives of corpus-based resources and machine translation. The 2001 campaign seems to have been the campaign where participants experimented with the largest overall number of different resources, while scaling back somewhat for 2002, keeping the resources and methods that worked well for them in 2001. A possible interpretation is that for many groups, 2000 was a starting point for their work in CLIR within an evaluation setting, meaning that they started out with rather simple systems. In 2001, these groups frequently adopted many of the ideas proposed by other groups, trying a large range of alternatives. Based on their experiences in that campaign, they then kept the pieces that worked well for them, and concentrated on making them work even better in 2002. The emergence of well-performing systems with strong parallels for the multilingual track would seem to be consistent with this possible pattern.

4.3 Learning Curve

Apart from adapting and enhancing other group's ideas, over the years participants have also moved from simpler to more complex systems, and from easier (monolingual) to harder (bilingual, or even multilingual) tasks. There are a number of groups that have progressed from only participating in monolingual retrieval in CLEF 2000 to producing full-hlown multilingual experiments in 2002. We are very

happy to see this effect, as it is an indication that participation in CLEF can stimulate groups to expand or enhance their systems. Examples of groups that have progressed like this are the groups from Istituto Trentino di Cultura (ITC-irst) (Bertoldi and Federico 2004), from University of Amsterdam (Hollink et al. 2004) and from University of Tampere (bilingual to multilingual) (Hedlund et al. 2004), which are all represented by papers in this special issue.

4.4 Thinking outside the Box - Some More Unusual Approaches

As previously stated, the CLEF campaigns are not set up to be a competition. While many participants seek to perfect their systems with respect to the effectiveness they attain in CLEF tracks and tasks, this is by no means the only valuable use of the evaluation resources provided by the CLEF consortium. A substantial number of groups want to try new, unproven ideas, regardless of the possibility that these methods may not be successful on the CLEF task. Not only are such contributions invaluable in stimulating new strands of research and avoiding a "monoculture" of look-alike CLIR systems, they also tend to help to increase the quality of the evaluation resources produced by the campaign, as they increase the breadth of results that get judged. We cannot in this restricted space do justice to all the creative ideas that have been proposed so far in the first three years of CLEF campaigns, but nevertheless we try to give a suggestion of the diversity of CLEF experiments by highlighting some of the work that fell outside of the norm for the respective year.

1. no translation. It seems a fair assumption that for successful cross-language retrieval, the system must translate either the query, the documents, or both to bridge the language gap. Contrary to this, the group at Johns Hopkins University (JHU-APL) has produced CLIR experiments that use no translation at all, instead matching on character *n*-grams shared between words in the respective languages. The method works best when the two languages are closely related etymologically, and when the query can be massively expanded prior to retrieval, in order to obtain as many *n*-gram matches as possible. For more details see McNamee and Mayfield (2004).

2. random indexing. The Swedish Institute for Computer Science (SICS) has demonstrated a new corpus-based approach that uses bit vectors of comparatively small length to represent relations between terms. They have used this technique both to calculate term-term similarities across languages on multilingual training data (Sahlgren and Karlgren 2002) and for query expansion in monolingual retrieval (Sahlgren et al. 2003).

3. lexical triangulation. When no translation resources are available for direct translation between two languages, one alternative is the use of a "pivot" intermediary language, translating from the source language to the pivot (often English), and then translating from the pivot to the target language. The drawback of this method is the amplification of translation ambiguities due to the multiple translation steps. A group from the University of Sheffield (Gollins and Sanderson 2001) demonstrated a technique that tries to minimize this

additional ambiguity by using multiple pivot languages, and combining the information derived from them.

4. translation selection (interactive). Most participants view the main tracks of the CLEF campaign as a "batch processing setting", deriving queries entirely automatically from the topics, and then using no manual intervention during the retrieval process. The introduction of the interactive track has allowed interested groups to investigate some of the additional problems that arise when humans enter the "equation", such as how to facilitate document selection for users that have little or no knowledge of the language the documents are written in. Even before the interactive track started, a group from New Mexico State University (Ogden and Du 2000) investigated the question of whether a monolingual user can aid the query translation phase of cross-language retrieval, by disambiguating the query terms during interaction with the system, thus allowing more precise translation

5. automorphology. One of the obstacles for scaling multilingual retrieval systems is the complexity involved in adding more languages to the system. Even systems that do not use elaborate linguistic knowledge and language-specific processing usually need at least some resources for new languages, such as a stemmer, or stopword lists. A group from University of Chicago (Goldsmith et al. 2001) investigated the question of how to automatically derive morphological information from a document corpus, without human supervision. They used their method as a stemming component for monolingual retrieval in CLEF.

4.5 Summarizing the Use of Different Translation Resource Types in CLEF

So far, we have discussed the different approaches used in CLEF, the blueprints derived from them, and the learning curves demonstrated by some groups as we can derive them from the reports published during the past campaigns. In the following, we will attempt to summarize our discussion by analyzing the adoption of different types of translation resources in CLEF.

For this purpose, we have divided the translation resources in four rough categories:

1. (Manually produced) Machine-readable dictionaries (MRD). All translation resources that consist of (manually assembled) lists of words and phrases and their associated translations. Manually produced thesauri are also included in this category.

2. Corpus-based approaches. All translation resources that have been automatically derived from suitable multilingual training data, such as parallel or comparable corpora. This includes methods such as latent semantic indexing, similarity thesauri, statistical translation models, translation probabilities, etc.

3. Machine Translation (MT). Full machine translation, i.e. the (attempt at) grammatically correct translation of entire sentences and documents.

4. No Translation at all. Approaches that use no direct translation resources, such as *n*-gram matching or cognate matching.

Not all methods and approaches can easily be classified according to this scheme. Some methods use a primary resource, such as a manually generated bilingual dictionary, coupled with a secondary resource, such as corpus-based translation probabilities. In these cases, when a type of translation resource is used only in a very limited fashion, we will note these secondary resources separately. Groups may use multiple types of primary and secondary resources in each campaign, either in a single experiment or in different experiments.

In an informal sense, we have for long time observed that participants learn from each other's experiences and adopt successful ideas. In addition to tallying the uses of different resource types, our goal was thus to investigate whether we can find evidence of "flows" between these four types of translation resources, i.e. if participants abandon one type of translation resource in order to move to a different type in the following campaign.

When looking at the numbers in Table 6, we can see that machine-readable dictionaries were the most used type of translation resource in all three campaigns covered by the analysis. A more in-depth analysis of this fact reveals that a large number of newcomers tend to adapt MRDs every year, boosting this number (Table 7). This may well be due to the fact that newcomers find it easiest to start out with dictionaries for their initial CLIR work (MT possibly being too much a "black box" to be integrated into their experiments). Even so, groups using (only) MRDs are also the most likely not to return for further campaigns, possibly indicating that these groups had limited agendas with regard to CLIR research.

Corpus-based approaches have seen an extraordinary surge between 2000 and 2001, doubling the number of participants using them. This is a good example of participants taking up (successful) ideas from the earlier campaign, and trying to expand on them. The rise in the number of groups using such approaches was also facilitated by the emergence of combination approaches as an attractive option of CLIR.

It is interesting to note that in 2001 the highest total number of different types of resources was used by the participants. Apparently, after the initial 2000 campaign participants started investigating as many translation resource types as possible, determining in 2001 what works well for them. The 2002 campaign then seems to have brought some consolidation.

Table 6. Use of different types of translation resources over the years (all participants). The table shows the number of participants using a specific type of resource either as primary resource or as a secondary resource (figure in brackets).

Resource type	CLEF 2000	CLEF 2001	CLEF 2002
Machine-readable dictionary (MRD)	13(1)	15(1)	11(0)
Corpus-based	4(2)	8(4)	6(2)
Machine translation	5(0)	7(0)	6(0)
No translation	0(0)	0(0)	1(0)

Table 7. Use of different types of translation resources over the years (newly adopted resources). The table shows the number of newly adopting a specific type of resource either as primary resource (first figure) or as a secondary resource (figure in brackets).

Resource type	CLEF 2000	CLEF 2001	CLEF 2002
Machine-readable dictionary (MRD)	-	9(1)	4(0)
Corpus-based	-	4(2)	1(1)
Machine translation	-	4(0)	2(0)
No translation	-	0(0)	1(0)

As mentioned, we have also attempted to detect flows between different types of translation resources, i.e. evidence of participants switching between translation technologies. The conclusion after our analysis has to be that there is in fact very little "flow" between translation resources from one campaign to the next. This may seem contrary to our informal observation that there is substantial uptake of ideas between groups. We attribute this phenomenon however to a trend we distinguish towards the use of combination systems: participants tend to add new types of translation resources to their systems rather than replacing them. This is not directly captured by tallying the use of individual resources. Furthermore, participants may shift the focus of their effort to different types of translation resources between years, even if they do not completely abandon them. Again, simply counting the number of groups using the different types of resources cannot uncover such subtle shifts. In fact, it is very hard to accurately derive the information necessary to track such minor shifts for as large a pool of participants as have participated in all the CLEF campaigns. Thus, while acknowledging the shortcomings of our analysis, we believe to have found evidence that participants tend to shift the focus of their work instead of abandoning selected translation resources. Apart from possibly being a result of more robust and flexible combination approaches being used, this may also be an indication of the difficulty in acquiring more and better suited translation resources: resources are too valuable to throw away.

5 Directions for the Future

In this paper, we have described the organization of the CLEF evaluation campaigns and the main results achieved so far. It is now the moment to ask whether we have gone far enough and whether our initial objectives as stated in the Introduction have been achieved:

- The infrastructure for the testing of multilingual information retrieval systems has been set up and has been shown to be operating well, scaling up to the more than 400 experiments submitted for the 2003 campaign.
- Test-suites of reusable data for a large number of European languages have been created and assessed for reliability: the end product of the EC-funded CLEF activity is envisaged to be test-suites that are publicly available rather than being restricted to CLEF participants as in the present.
- A strong CLIR research community, involving both academic and industrial participants, which recognizes CLEF as a forum for cross-language research and development activities, has been created.
- Research into improving the performance of systems for monolingual information retrieval for European languages other than English has been stimulated.
- A successful blueprint for the building of effective multilingual text retrieval systems appears to be in existence.

The question is thus whether there is any real need to continue with the CLEF campaigns.

To some extent, it is fair to say that research in the cross-language information retrieval sector is currently at a crossroads. In a recent workshop at SIGIR 2002[9] the question asked was whether the CLIR problem can now be considered as solved. The answers given were mixed: the basic technology for cross-language text retrieval systems is now in place as is clearly evidenced by the papers in this special issue. But if this is so, why has this technology not been adopted by any of the large Web search engines and why do most commercial information services not offer CLIR as a standard service? Although there is a strong market potential, the actual systems are still not ready to meet the needs of the generic user. For a commercial CLIR system to be successful, it needs to be versatile, efficient when working on-line, accommodate many ("all") languages, present its results in a sufficiently user-friendly fashion, and possibly handle multimedia. It is clear that much work remains to be done to address these points and bridge the present gap between the CLIR R&D community and the application world.

[9] See http://ucdata.berkeley.edu/sigir-2002/ for details on the workshop and, in particular, the position paper by Douglas W. Oard.

What role should CLEF play in this scenario? It seems evident that in the future, we must go further in the extension and enhancement of CLEF evaluation tasks, moving gradually from a focus on cross-language text retrieval and the measurement of document rankings to the provision of a comprehensive set of tasks covering all major aspects of multilingual, multimedia system performance with particular attention to the needs of the end-user. This is not unlike what has happened in recent years with the TREC conferences.

Some of the further challenges left to address by CLEF include:

- What kind of evaluation methodologies should be developed to address more advanced information requirements?
- How can we cover the needs of all European languages – including minority ones?
- What type of coordination and funding model should be adopted – centralized or distributed?
- How can we best reduce the gap between research and application communities?

The most recent campaign, CLEF 2003, has attempted to move in the direction outlined above. It offered four additional tracks on top of the traditional core tracks for mono-, bi- multilingual and domain-specific systems. The aim has not only been to diversify the tasks offered but also to stimulate system developers to work in those areas that should find take-up by the application world. For this reason, in addition to continuing with the very successful interactive track, CLEF 2003 also included an activity for multilingual question answering (QA). The development of a successful question answering system implies not only a finely tuned IR engine but also sophisticated NLP functionality. Considerable work on QA system development has already been undertaken for English (with evaluation tracks at TREC) and has recently begun for Asian languages (with evaluation for systems running on Japanese at NTCIR). CLEF is attempting to encourage similar experiments for other European languages and also the development of QA systems that search across languages. Such systems recognise that users frequently want to retrieve certain specific pieces of information, and only that information, without having to wade through a large number documents to find it.

We are also beginning to take the first steps towards the evaluation of systems that run on collections of documents in more than one medium. Two pilot experiments were thus set up to test cross-language spoken document retrieval systems and cross-language retrieval on an image collection in the CLEF 2003 campaign. Both activities actually test the performance of particular types of text retrieval systems: automatic speech transcriptions in the first case, image captions in the second. Our aim is to stimulate system developers into examining and handling the problems involved in searching over multilingual, multimedia collections where language dependent and language independent factors interplay. These new tracks were organised as a result of proposals made at the CLEF 2002 workshop and

consequent to the interest expressed in the ensuing discussion. Preliminary findings have been presented at the CLEF 2003 workshop.

We believe that if CLEF is to continue in the future it will be necessary to continue in this direction – being increasingly aware of and adapting to the needs of commercial system developers and offering a variety of tasks designed to meet these needs.

For more information on the CLEF evaluation campaigns, see http://www.clef-campaign.org.

Acknowledgments

The authors would like to acknowledge the help, support and advice of numerous individuals and organizations which has been invaluable in the organization of the CLEF campaigns. Many aspects of the CLEF campaign have been modelled after successful blueprints that were previously tested and developed within the TREC conferences. CLEF itself started its "life" as a track organized within the TREC framework. We are particularly grateful to Donna Harman and Ellen Voorhees from NIST, organizers of TREC, for their tireless support. CLEF is an activity conducted by the CLEF project consortium, of which the two authors are but two members. We are greatly indebted to all the other consortium members, both individuals and their organizations. Lastly, we want to acknowledge all our data providers; too numerous to list here. Without their generous support the CLEF evaluation activity would be impossible.

References

Amati, G., Carpineto, C., Romano, G. (2003): Italian Monolingual Information Retrieval with PROSIT. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785, Spinger Verlag, pp. 257-264.

Bertoldi, N., Federico, M. (2003): ITC-irst at CLEF 2002: Using N-best Query Translations for CL-IR. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign. LNCS 2785, Spinger Verlag, pp. 49-58

Bertoldi, N., Federico, M. (2004): Statistical Models for Monolingual and Bilingual Information Retrieval, 7:51-70.

Brand, R., Brünner, M. (2003): Océ at CLEF 2002. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785, Spinger Verlag, pp. 59-65.

Braschler, M., Krause, J., Peters, C., Schäuble, P. (1999): Cross-Language Information Retrieval (CLIR) Track Overview. In: Proceedings of the Seventh Text REtrieval Conference (TREC-7), NIST Special Publication 500-242, pp. 25-32.

Braschler, M., Schäuble, P. (2001): Experiments with the Eurospider Retrieval System for CLEF 2000. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Revised Papers. pp. 140-148.

Braschler, M. (2002): CLEF 2001 – Overview of Results. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.): Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Revised Papers.

Braschler, M. (2003): CLEF 2002 – Overview of Results. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785, Spinger Verlag, pp. 9-27.

Braschler, M. (2004): Combination Approaches for Multilingual Text Retrieval. Information Retrieval, 7:181-202.

Chen, A., Gey, F. C. (2004): Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Word Decompounding. Information Retrieval, 7:147-180.

Cleverdon, C. (1977) The Cranfield Tests on Index Language Devices. In: K. Sparck-Jones and P. Willett, eds. Readings in Information Retrieval. Morgan Kaufmann, 1997. pp. 47-59.

Goldsmith, J. A., Higgins, D., Soglasnova, S. (2001): Automatic Language-Specific Stemming in Information Retrieval. In C. Peters (Ed.). Cross-Language

Information Retrieval and Evaluation. Lecture Notes in Computer Science 2069, Springer Verlag, pp. 273-283.

Gollins, T., Sanderson M. (2001): Sheffield University: CLEF 2000 Submission – Bilingual Track: German to English. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Revised Papers. pp. 245-252.

Grefenstette, G. (ed.) (1998). Cross-Language Information Retrieval. Kluwer Academic Publishers.

Harman, D. (1995): The TREC Conferences. In R. Kuhlen and M. Rittberger (Eds.): Hypertext - Information Retrieval - Multimedia: Synergieeffekte Elektronischer Informationssysteme, Proceedings of HIM '95. Universitätsverlag Konstanz, pp. 9-28.

Harman, D., Braschler, M., Hess, M., Kluck, M., Peters, C., Schäuble, P., Sheridan, P. (2001): CLIR Evaluation at TREC. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Revised Papers. pp. 7-23.

Hedlund, T., Airio, E., Kekustalo, H., Lehtokangas, R., Pirkola, A., Järvelin, K. (2004): Dictionary-Based Cross-Language Information Retrieval: Learning Experience from CLEF. Information Retrieval, 7:97-117.

Hiemstra, D., Kraaij, W., Pohlmann, R., Westerveld, T. (2001): Translation Resources, Merging Strategies, and Relevance Feedback for Cross-Language Information Retrieval. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Revised Papers. pp. 102-115.

Hollink, V., Kamps, J., Monz, C., de Rijke, M. (2004): Monolingual Retrieval for European Languages. Information Retrieval, 7:31-50.

Jones, G. J. F., Federico, M. (2003): Cross-Language Spoken Document Retrieval Pilot Track Report. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785. Spinger Verlag, pp. 446-457.

Kando, N. (2003). CLIR at NTCIR Workshop 3. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign. LNCS 2785. Spinger Verlag. pp. 485-504.

Kluck, M., Gey, F.C. (2001). The Domain-Specific Task of CLEF – Specific Evaluation Strategies in Cross-Language Information Retrieval. In C. Peters (Ed.). Cross-Language Information Retrieval and Evaluation. Lecture Notes in Computer Science 2069, Springer Verlag, pp. 48-56.

Lam-Adesina, A. M., Jones, G. J. F. (2003): Exeter at CLEF 2002: Experiments with Machine Translation for Monolingual and Bilingual Retrieval. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785. Spinger Verlag, pp. 127-146.

Mandl, T., Womser-Hacker, C. (2003): Linguistic and Statistical Analysis of the CLEF Topics. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign. LNCS 2785, Spinger Verlag, pp. 505-511.

McNamee, P., Mayfield, J., Piatko, C. (2001): A Language-Independent Approach to European Text Retrieval. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Revised Papers. pp. 129-139.

McNamee, P., Mayfield, J. (2004): Character N-gram Tokenization for European Text Retrieval. Information Retrieval, 7:71-95.

Nie, J.-Y., Simard, M., Foster, G. (2001): Multilingual Information Retrieval Based on Parallel Texts from the Web. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Revised Papers. pp. 188-201.

Oard, D. W., Gonzalo, J., Sanderson, M., López-Ostenero, F., Wang, J. (2004): Interactive Cross-Language Document Selection. Information Retrieval, 7:203-226.

Ogden, B., Du, B. (2000): Can Monolingual Users Create Good Multilingual Queries Without Machine Translation? In Peters, C. (Ed): First Results of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign. Working Notes for the CLEF 2000 Workshop, ERCIM-00-W01, pp. 133-134.

Pieters, D. (2002). The Languages of the European Union, Publication of the European Commission, IBSN 90-807420-3-1. Available at http://europa.eu.int/futurum/documents/offtext/espdiscuss10_en.pdf

Sahlgren, M., Karlgren, J. (2002): Vector-based Semantic Analysis using Random Indexing for Cross-lingual Query Expansion. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.): Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Revised Papers, pp. 169-176.

Sahlgren, M., Karlgren, J., Cöster, R., Järvinen, T. (2003): SICS at CLEF 2002: Automatic Query Expansion using Random Indexing. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign. LNCS 2785, Spinger Verlag, pp. 311-320.

Savoy, J. (2004): Combining Multiple Strategies for Effective Monolingual and Cross-language Retrieval. Information Retrieval, 7:119-146.

Schäuble, P., Sheridan P. (1998). Cross-Language Information Retrieval (CLIR) Track Overview. Proceedings of the Seventh Text Retrieval Conference (TREC-7). NIST Special Publication 500-422, pp. 25-32.

Smeaton. A.F., Harman, D. (1977). The TREC Experiments and their Impact on Europe, Journal of Information Science, Vol. 23, 1997, pp.169-174.

Salton, G. (1970): Automatic processing of foreign language documents. Journal of the American Society for Information Science, 21(3):187-194.

Voorhees, E. (2002): The Philosophy of Information Retrieval Evaluation. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.): Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Revised Papers, pp. 355-370.

Voorhees, E., Buckley, C. (2002): The Effect of Topic Set Size on Retrieval Experiment Error. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 316-323.

Womser-Hacker, C. (2002): Multilingual Topic Generation within the CLEF 2001 Experiments. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.): Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Revised Papers, pp. 389-393.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.

How Effective is Stemming and Decompounding for German Text Retrieval?

Martin Braschler[1,2], Bärbel Ripplinger[1]

[1] Eurospider Information Technology AG
Schaffhauserstrasse 18, CH-8006 Zürich, Switzerland
martin.braschler@eurospider.com, BRipplinger@web.de

[2] Université de Neuchâtel, Institut Interfacultaire d'Informatique
Pierre-à-Mazel 7, CH-2001 Neuchâtel, Switzerland
martin.braschler@unine.ch

Abstract. Information retrieval systems operating on free text face difficulties when word forms used in the query and documents do not match. The usual solution is the use of a "stemming component" that reduces related word forms to a common stem. Extensive studies of such components exist for English, but considerably less is known for other languages. Previously, it has been claimed that stemming is essential for highly declensional languages. We report on our experiments on stemming for German, where an additional issue is the handling of compounds, which are formed by concatenating several words. The major contribution of our work that goes beyond its focus on German lies in the investigation of a complete spectrum of approaches, ranging from language-independent to elaborate linguistic methods. The main findings are that stemming is beneficial even when using a simple approach, and that carefully designed decompounding, the splitting of compound words, remarkably boosts performance. All findings are based on a thorough analysis using a large reliable test collection.

Keywords: stemming, decompounding, German, evaluation, morphological analysis

1 Introduction

Most modern information retrieval (IR) systems implement some sort of what is commonly known as "stemming". A system using stemming conflates derived word forms to a common stem. The benefits of such a procedure is two-fold: by conflating several forms to the same stem, the number of entries and the size of the search index is reduced. More importantly, stemming is potentially helpful for free text retrieval, where search terms can occur in various different forms in the document

collection. Stemming makes retrieval of such documents independent from the specific word form used in the query. In our experiments, we will concentrate on this second aspect, the impact of stemming on retrieval effectiveness.

The main reason for the use of stemming is the hope that through the increased number of matches between search terms and documents, the quality of search results is improved. In terms of the most popular measures for determining retrieval effectiveness, precision (amount of relevant documents retrieved compared to all documents retrieved) and recall (amount of relevant documents retrieved compared to total number of relevant documents in the collection), stemmed terms retrieve additional relevant documents that would have otherwise gone undetected, i.e. they improve recall. There is also potential for improved precision, since additional term matches can contribute to a better weighting for a query/document pair.

Viewing stemming as a vehicle to enhance retrieval effectiveness necessitates an analysis of the produced stems with regard to overstemming or understemming. If a stemmer conflates terms too aggressively, thereby conflating unrelated or loosely related words, many extraneous matches between the query and irrelevant documents are produced; we talk of "overstemming". Even though some of these stems may be correct from a linguistic viewpoint, they are not helpful for retrieval. In contrast, if crucial relevant documents are missed because of a conservative stemming strategy, we speak of "understemming". A good stemmer has to find the right balance of conflation for effective retrieval.

Where many previous studies compare their stemming method only to no stemming or to simple affix stripping methods such as proposed by Lovins (1968) and Porter (1980), the experiments described here compare a complete spectrum of methods, ranging from language-independent to sophisticated linguistic analysis. We consider this an important step forward, since while smaller studies may demonstrate the benefit of stemming for a certain language in principle, they cannot answer the important question of how to build a component that maximizes the potential of stemming for that language. Only an exhaustive comparison can give an indication of the right amount of linguistic knowledge and the right balance of conflation that is necessary.

The paper deals with the German language, where, besides stemming, decompounding seems to be an additional issue in retrieval. In most Germanic languages (e.g. German, Dutch, Swedish), but also in some other languages (e.g. Finnish), it is possible to build compounds by concatenating several words. Performance, especially recall, may be negatively affected if this word formation process is not taken into account. If the user enters only parts of compound words or replaces the compound with a phrasal construction, relevant documents may not be found.

Similar to stemming, there are different approaches to analyze and split compounds ("decompounding"). We concentrate on the usefulness of compound splitting for information retrieval, and not as a linguistic exercise. The splitting of some compounds words can cause a shift of meaning. Such compounds should be

left intact ("conservative decompounding"). On the other hand, the productive nature of German compound formation, which allows for ad-hoc building of new compounds, makes it impossible to enumerate all potential compound words. "Aggressive decompounding" attempts to produce a maximum number of splittings by using heuristics. In our experiments, we investigated different approaches to decompounding, ranging from surface-based to well-developed linguistic methods.

As a consequence of limited resources for some languages, many studies on stemmer and decompounding behavior have been conducted using small test collections, containing short documents. Both the size of the collection and the length of the documents can influence observations on stemming. Especially the length of the retrieval items is important, since short documents (e.g. only titles, or titles and abstracts) increase the likelihood for word mismatch if no stemming is used. It is therefore not immediately obvious if a performance improvement measured on collections with short documents, such as widely used before 1990, translates into an improvement on larger collections with long documents. Furthermore, large test collections may be needed to draw more reliable conclusions. In our work we used a part of the large German collection from the CLEF corpus to get appropriate statistical data. The collection contains lengthy newspaper and news magazine articles.

Similarly, the characteristics of the queries can impact the conclusions of the stemming experiments. Again, longer queries will have more likelihood of having multiple forms of search terms included, which may affect the amount of new information that a stemming component can contribute. As a consequence, we have used two different query lengths in all our experiments.

The remainder of the paper is structured as follows: we will first give an overview of related work (Section 2) and a summary of some relevant key characteristics of the German language (Section 3), before detailing the experimental setup in Section 4 (test collection), Section 5 (stemming and decompounding approaches) and Section 6 (retrieval system). The experiments themselves are then discussed, along with a careful statistical analysis and a query-by-query analysis (Section 7 through Section 9). The paper closes with conclusions and an outlook.

2 Related Work

The role of stemming is well explored for English, even though results are controversial. Where Harman (1991) reported that stemming gives no benefit, Frakes (1992) and Hull (1996) claim at least a small benefit. Especially notable is the study by Hull (1996), which describes a very careful analysis of different stemmers and provides an ambitious attempt to compare a wide range of different approaches to stemming. Furthermore, in this study, data from the popular TREC test collections (Harman 1997) is used, which are much larger than earlier collections and contain lengthy documents. Because English has no productive compounding process, we are not aware of studies into the effects of decompounding for this language.

In recent years, more studies on stemming behavior for languages other than English have been published. While simple stemmers are considered sufficient for languages such as English, it is claimed that rich declensional languages require more sophisticated stemming approaches. Studies exist on Slovene (Popovic and Willet 1992), Hebrew (Choueka 1992), Dutch (Kraaij and Pohlmann 1996b), German (Womser-Hacker 1989, Moulinier et al. 2001), Italian (Sheridan and Ballerini 1996), French (Savoy 1999), and others. For most of these languages, significant benefits (from 10% to 130%) can be observed. Decompounding has been investigated, at least for German (Moulinier et al. 2001) and Dutch (Kraaij and Pohlmann 1996a). The improvements reported are about 17% to 54%.

Previous work on what could be termed as "German stemming" was mainly conducted in the context of studies comparing manual and automatic indexing in databases. The PADOK studies (Womser-Hacker 1989, Krause and Womser-Hacker 1990), carried out in the late eighties, compared several approaches to automatic indexing on patent data. The baseline system indexed every token in the original form as it appears in the document. However, queries could make use of string truncation and wildcards. This baseline system was compared to the two systems "PASSAT", comprising morphological analysis, including decompounding (used in PADOK I & PADOK II), and "CTX" which additionally applied a syntactical analysis (used in PADOK I only). In PADOK I (Womser-Hacker 1989), documents consisting of titles and abstracts only were used. It was found that PASSAT achieved higher recall and CTX achieved higher precision compared to the baseline system. In PADOK II (Krause and Womser-Hacker 1990), the full text of the documents was used. In this new setting, no significant gain from using the morphological analysis employed by PASSAT could be detected.

The later GIRT pretest used texts from the field of social sciences (Frisch and Kluck 1997). Two systems were compared: "freeWAISsf" and "Messenger", differing in more areas than just in the stemming approaches employed (freeWAISsf uses weak morphological analysis, Messenger apparently no stemming). This makes it hard to draw conclusions on the effectiveness of the stemming component based on the published results.

Recent work by Moulinier et al. (2001) reported a performance gain of 20% using linguistic analysis (stemming plus compound analysis) for German. A gain of 89% in recall together with a loss of 24% in precision are the results obtained by Ripplinger (2002) in a small experiment within a domain specific environment. The experiments by Monz and de Rijke (2002) show a gain of 25% to 69% for German and Dutch respectively using blind relevance feedback in addition to a linguistic processing. Tomlinson (2002) applied linguistic stemming including decompounding, and obtained a performance improvement of 43% for German, and 30% for Dutch. He also conducted experiments for English and Romance languages (French, Italian, and Spanish) with a performance gain of 5% to 18%.

Most of this recent work came in the context of the TREC CLIR task and the CLEF campaigns. The experiments are usually confined to a comparison between one particular stemming or decompounding approach and not using any form of

stemming at all. It is therefore unclear how the findings can assist in the choice of a particular stemming approach or how to assess the absolute potential of stemming for performance improvement in the German language.

3 Characteristics of the German Language

German, in contrast to English, is a highly declensional language, a fact expressed by a rich system of inflections and cases. Depending on the word class, i.e. noun, verb, or adjective, there is a set of possible inflections for each particular word (e.g. 144 forms for verbs). For instance, "informiert", "informiere", "informierte", "informierten" are forms of the verb "informieren" (to inform). Furthermore, words can be formed by attaching multiple derivational inflections to a stem in order to build new forms. For instance, the lexeme "inform" is the stem for "informieren", "informierend", "Information", "Informant", "informative", "informatorisch", but not "informal". To ensure efficient retrieval all these forms should be conflated to the same stem.

Additionally, in most Germanic languages (e.g. German, Dutch, Swedish), as well as in some other languages (e.g. Finnish), compounds can be built by concatenating several words, such as e.g., "Haus|tür" (house door), "Wind|energie" (wind energy), "Früh|stück" (breakfast) or "Unglücks|fall" (accident). Such compound formation occurs in almost all languages, such as "hair|dresser", "speed|boat" in English, and "porte|feuille" (wallet), "hon|homme" (fellow) in French. In English and in Romance languages, however, these words are lexicalized, i.e. they cannot be expressed by a nominal phrase the way compounds in Germanic languages could be ("Haustür" vs. "Tür des Hauses"). Because Germanic languages also know lexicalized compounds (e.g. "Wettbewerb", "Frühstück"), the correct treatment of such words is quite complicated. Therefore, we expect that compound analysis requires more linguistic knowledge to be of benefit than stemming does.

In principle, only words with certain types of part-of-speech can be coupled, for instance noun/noun, adjective/noun, or verb/noun. Some analyzers split only those compounds where the constituents have the same part-of-speech (noun/noun, adjective/adjective), and could be thus classified as "conservative". Others split the compounds into all possible word forms (often by means of a lexicon lookup). Because these methods do not consider the part of speech of the constituents, i.e. also splitting into pronouns, prepositions or articles, this approach can be described as "aggressive". Linguistic compound analyzers lie in between, splitting compounds only into valid word forms, i.e. nouns, verbs, adjectives.

4 The Test Collection

Many of the studies on stemming behavior for languages other than English suffer from scarce availability of suitable test collections. Whereas for English the well-known TREC collections exist, comparable test data became available only recently for important European and Asian languages (through the CLEF and NTCIR campaigns). For their 2001 evaluation campaign, the CLEF consortium distributed a

multilingual document collection of roughly 1,000,000 documents written in one of six languages (Dutch, English, French, German, Italian and Spanish). For the German part of this data, CLEF 2000 used articles from the daily newspaper "Frankfurter Rundschau" and the weekly news magazine "Der Spiegel". CLEF 2001 used a superset of this, adding newswire articles taken from the "Schweizerische Depeschenagentur SDA". To go with this data, there is a total of 90 topics[1] (40 for 2000, 50 for 2001) with corresponding relevance assessments. By eliminating the SDA articles, we were able to form a unified German test collection, using the topics from both years. Of the 90 topics, 85 have matches in the German data that we used. This comparatively high number of topics (many studies use only 50 queries or less) facilitates the detection of significant differences in stemming and decompounding behavior (see also Table 1 for more details of the test collection).

5 Stemming and Decompounding Approaches

One of the main objectives of this study is to compare approaches from a spectrum as wide as possible. This spectrum covers methods ranging from completely language-independent methods to components that use elaborate linguistic knowledge. In our experiments, apart from using different stemming and decompounding methods, all other indexing parameters remain constant (tokenization, stopword list, etc.).

To keep things readable, the names of the different approaches are abbreviated in some places using a letter-based abbreviation scheme. The respective "codes" are given in the following subheadings. Approaches applying stemming only are denoted by lower case letter codes, while those deploying decompounding are marked by capitalizing the initial letter.

The approaches used in this study are:

5.1. No stemming ("n" "nostem" run)

As a baseline, we indexed all documents without using any form of stemming or decompounding.

---

[1] Topics are "statements of user needs". They form the basis for the formulation of queries, which are then run against the document collection. In all our experiments, we constructed the queries without manual intervention ("automatic experiments"), by indexing all or part of the topic text.

Table 1. Key characteristics of the test collection

Document source	Frankfurter Rundschau 1994, Der Spiegel 1994, 1995
Number of documents	153,694
Size (MByte)	383
Number of topics	85[2]
Number of indexing tokens per document (tokens minus stopwords)	
Maximum	2.885
Minimum	1
Mean	156.09
Median	128
Relevance assessments:	
Number of assessments:	20,980
Number of relevant documents	1.790
Maximum relevant (one topic)	109
Minimum relevant (one topic)	1
Mean relevant (one topic)	21.06
Median relevant (one topic)	12

5.2. Combination of word-based and n-gram based retrieval ("6" "6gram+word" run)

The use of combined character *n*-gram and word-based indexing was reported as a successful approach to German text retrieval by Mayfield et al. (1999) and Savoy (2003). Based on their findings, we chose to combine 6-grams and unstemmed words. The individual 6-grams, built on the unstemmed words, potentially span word boundaries. A main benefit of this approach is its complete language independence - no specific linguistic knowledge is needed to form the *n*-grams, and the word-based indexing is done without attempting conflation. On the other hand, the method is storage-intensive: the large number of different *n*-grams leads to a massively increased index size (roughly three times that of an unstemmed word-based index).

5.3. Linguistica: Automatic machine learning ("l" "linguistica" run)

Linguistica (Goldsmith 2001) performs a morphological segmentation based on unsupervised learning. The aim is to find the correct morphological splits for individual words, in a language-independent way. The program first establishes for all words a "best splitting" into two parts of two or more letters (not to be confused with splitting compounds). These are then treated as a "stem" and a "suffix", even

[2] 90 from CLEF multilingual topic sets, minus 5 topics without any relevant documents in the subset of the German data we used

though they may in reality represent other units, such as a prefix and a stem. If multiple candidates for such a splitting exist, heuristics based on word frequency and the number of possible parses are applied. In a second step, the program determines a list of suffixes that can occur with each stem, called the "signature" of a stem. Each stem is associated with exactly one signature. For example, in our experiments, the stems "erzeug", "verpack", "befrag" and "beschaff" were associated with the signature "en.er.t.te.ung", meaning that Linguistica learnt splittings such as "verpack-en" and "befrag-ung" from the test collection.

As a result, Linguistica outputs a lists of signatures, possible affixes, and the list of words contained in the processed corpus together with their possible splits (the "lexicon"). For the test collections (consisting of roughly 1.4 million unique terms) the lexicon contains about 123.000 entries, for instance "machbar" (feasible), "machbar-en" (feasible), "machbar-es" (feasible), "machbar-keit" (feasibility). For our experiment we used these entries, which often do not really denote a proper stem from a linguistic viewpoint.

Decompounding is not a feature in Linguistica, however, the program is sensitive to words used frequently in compounds such as "bank" (bank), "partei" (party) and "teil" (part). These are categorized as affixes, leading to signatures such as:

NULL.bank.bild.dienste.haus.partei.plan.rat.staat.teil.viertel

(NULL.bank.picture.service.house.party.plan.council.state.part.quarter)

All values are valid lexemes (i.e. nouns) in German. Such signatures result in accidental decompounding: since suffixes are discarded after stemming, compounds such as "Datenbank" (data base), "Anklagebank" (dock), "Handelsbank" (merchant bank) are conflated to "Daten" (data), "Anklage" (charge) and "Handel" (trade). The second component of the compound noun is lost and cannot be used for subsequent retrieval.

5.4. NIST stemmer: Rule-based approach ("t" "nist_stem" run, "T" "nist_dec" run)

The NIST German rule-based stemmer has been constructed through analysis of the frequency of German suffixes in large wordlists. Its approach is similar to the Porter English stemmer (Porter 1980). The rules were hand-crafted to produce as many valid conflations of high-frequency word forms as possible, while keeping the rate of incorrect conflations low. Rules operate based on the suffix to be replaced, the new suffix used as the replacement, the length of the word, and additional, special criteria. The length of the word is calculated using a special measure introduced by Porter that counts the number of consonant/vowel sequences (informally connected to the number of syllables a word contains). Additional

criteria include among others whether the word contains at least one vowel, or if the word ends in special character combinations. The stemmer attempts to iteratively strip suffixes from a word, applying the rules cyclically. The resulting stems are often not correct from a linguistic viewpoint. For instance the word "glück-lich-er-weis-e" (luckily) is reduced to "gluck" (same stem as for "luck"). The stemmer is available as a part of the NIST ZPrise 2 retrieval system.

The stemmer can be combined with a corpus-based decompounding component based on co-occurrence analysis. After collecting a list of candidate nouns, the component tries to find valid splittings by looking for potential constituents that co-occur in the same documents. In case more than one potential splitting is identified, a heuristic is used for selection. Highest preference is given to splittings that result in a maximum number of constituents. This purely corpus-based approach produces a number of errors, but is overall rather conservative in the number of splittings generated.

For our experiments, we used both the pure stemming approach ("t" "nist_stem" run) and the combination with decompounding ("T" "nist_dec" run). This study represents the first careful evaluation of the NIST German stemmer.

5.5. Spider stemmer: Commercially motivated (rule -based and lexicon) approach ("sn" "spider_NS", "Sf" "spider_FS", "Sc" "spider_CS", "Ss" "spider_SS" runs)

For our experiments we had access to the stemming component used in the commercial "*relevancy*" retrieval system by Eurospider, the "Spider stemmer". The approach is based on a combination of a lexicon and a set of rules that are used for suffix stripping and unknown words (Wechsler et al. 1997).

The Spider stemmer has been used for over five years in all commercial installations of the system, and has therefore been constantly adapted according to customer feedback. Extensive performance figures for this stemmer are published publicly for the first time in this study.

The component includes optional decompounding, which can be applied in one of three modes, from conservative to aggressive splitting. For our runs we used the version without decompounding (run "sn" spider_NS) as well as three forms of decompounding: a conservative option (splits words only if all the components have the same part-of-speech as the overall compound, run "Ss" "spider_SS"), a more relaxed option (at least one component has to match the overall compound with regard to part-of-speech, run "Sc" "spider_CS") and an aggressive version (split whenever possible, run "Sf" "spider_FS").

For example, "Umweltaspekte" (environmental aspects) is reduced to "umweltaspekt" (run "sn") and split to "um-welt-aspekt" (run "Sf").

5.6. MPRO: Morpho-syntactic analysis ("mu" "MPRO_lu" run, "ms" "MPRO_ls" run, "Mu" "MPRO_lu_dec" run, "Ms" "MPRO_ls_dec" run)

MPRO, a development by the IAI (Maas 1996), performs a morpho-syntactic analysis consisting of lemmatization, part-of-speech tagging, and, for German, a compound analysis. To lemmatize and tag words with their part-of-speech(s), MPRO uses general morphological rules in form of small subroutines that co-operate with a morphological dictionary. As a result, for each word a set of attribute-value pairs describing inflectional attributes (e.g. gender, number, tense, mood, etc.), word structure and semantics (lexical base form, derivational root form, compound constituents, semantic class, etc.) is produced. For instance, the analysis of the word "Kollisionen" (collisions) results in

{string=kollisionen,c=noun,lu=kollision,nb=pl,g=f,t=kollision,

ls=kollidieren,s=ation,...}

and produces the lexical unit "kollision" and the root form "kollidieren" (collide). For the compound "Umweltaspekte" (environmental aspects) MPRO generates a splitting based on lexical base forms, "umwelt-aspekt" (t feature), and one based on derivational root forms of the compound constituents, (ls feature; in this example identical).

{string=umweltaspekte,c=noun,lu=umweltaspekt,nb=pl,g=f,

t=umwelt_aspekt,ls=umwelt_aspekt,...}

With this tool, the corpus has been analyzed and for each word, information about the lexical base form (lu), and the derivational root (ls) is used to generate two lexical resources. The derivational root form presents a kind of "aggressive" stemming and splitting because the morphological derivation is more general than the lexical base form. For a more detailed description how to deploy MPRO for IR purposes see Ripplinger (2002).

We have used either the lexical base forms or the root forms and no decompounding information for the stemming only runs, "mu" "MPRO_lu" and "ms" "MPRO_ls", and the splitting based on lexical base forms or on derivational information for the decompounding runs, "Mu" "MRPO_lu_dec" and "Ms" "MPRO_ls_dec".

6 The Retrieval System

For retrieval, the commercial "*relevancy*" system from Eurospider is used. Since the system uses a modular approach, the lexical resources generated from the individual approaches described above can easily be used for indexing. "*relevancy*" also offers the possibility of indexing *n*-grams as described above. For weighting of both documents and queries, we chose the Lnu.ltn weighting scheme (Singhal et al. 1996), with the "slope" parameter set to 0.15 (determined empirically to be the optimal value). Lnu.ltn is widely used and has given competitive performance on the CLEF collection (see e.g. Braschler and Schäuble 2001). The scheme includes document length normalization, a factor influenced by stemming, and some caution may be appropriate when generalizing the results to approaches that use vastly different collection statistics for weighting. There is no different weight for compound constituents as proposed by Hull et al. (1996).

7 Experiments

For each of the approaches described above, we produced several results sets ("runs"), using different sections of the CLEF topics. The topics are structured into title, description and narrative fields, which express a user's information need in different length. The title section (T) typically consists of one to three keywords, the description section (D) typically consists of one sentence, and the narrative section (N) usually is several sentences long, giving detailed instructions on how to determine the relevance of retrieved documents. Using these fields, seven different combinations are possible (TDN, TD, TN, DN, N, D, T). In this paper, we report on the results obtained using the longest possible queries TDN and the shortest possible queries T, two popular choices for retrieval experiments.

To clearly present the results in the following tables and graphs the runs have been divided into several groups. In tables, runs of the two query lengths TDN and T are distinguished. Tables are also divided into approaches using decompounding (top half) and those using stemming only (lower half), separated by a dotted line. The baseline of no stemming is given at the bottom of the table. The runs have been structured further for the graphs. We give graphical results for both query lengths, one overview graph each for runs with and without decompounding. In these overview graphs, the representation of the Spider and MPRO stemmers is limited to one variant each. We have chosen the best performing options: "Sf" "spider_FS" for Spider/decompounding, "sn" "spider_NS" for Spider/stemming only, "Ms" "MPRO_ls_dec" for MPRO/decompounding, and "mu" "MPRO_lu" for MPRO/stemming only. A graphical presentation of the results of the other variants is given in separate graphs. In all graphs, "no stemming" is represented as the baseline for reference.

Table 2. Retrieval results (mean average precision). Runs using decompounding are given in the upper half, whereas the runs in the lower half use stemming only.

Run (T)	Avg prec (T)	Rrun (TDN)	Avg prec (TDN)
"Sf" spider_FS	0.3650 (+60.4%)	"Sf" spider_FS	0.4471 (+34.6%)
"Ss" spider_SS	0.3586 (+57.6%)	"Ms" MPRO_ls_dec	0.4440 (+33.7%)
"Ms" MRPO_ls_dec	0.3547 (+55.9%)	"Ss" spider_SS	0.4431 (+33.4%)
"Sc" spider_CS	0.3546 (+55.9%)	"Sc" spider_CS	0.4415 (+32.9%)
"Mu" MPRO_lu_dec	0.3385 (+48.8%)	"Mu" MPRO_lu_dec	0.4234 (+27.5%)
"I" nist_dec	0.3240 (+42.4%)	"T" nist_dec	0.4022 (+21.1%)
"6" 6gram+word	0.2757 (+21.2%)	"6" 6gram+word	0.3219 (-3.1%)
"t" nist_stem	0.2792 (+22.7%)	"t" nist_stem	0.3682 (+10.9%)
"sn" spider_NS	0.2722 (+19.6%)	"sn" spider_NS	0.3616 (+8.9%)
"mu" MPRO_lu	0.2682 (+17.9%)	"mu" MPRO_lu	0.3461 (+4.2%)
"ms" MPRO_ls	0.2353 (+3.4%)	"l" linguistica	0.3435 (+3.4%)
"l" linguistica	0.2302 (+1.2%)	"ms" MPRO_ls	0.3396 (+2.3%)
"n" nostem	0.2275	"n" nostem	0.3321

In Table 2, we give the mean average precision numbers for all runs. This is the most popular single-valued measure for TREC-style retrieval experiments, and is defined as non-interpolated average precision over all relevant documents. According to these numbers, methods that use decompounding (lower half of Table 2) perform better than methods that do not split compound words (upper half of Table 2). An exception is the combined 6-gram/word-based indexing run, which we treat as a "decompounding run", since the 6-grams allow matching of compound parts. This is the only run that for long queries (TDN) performs worse than the baseline of no stemming. The best variants of the two top methods, "Sf" "spider_FS" and "Ms" "MPRO_ls_dec", show almost identical performance, while "T" "nist_dec" performs slightly worse (see also figures 2 and 4). They obtain *performance gains of between 33% and 60% depending on query length when* compared to no stemming

The gap to the best method without decompounding ("t" "nist_stem") is fairly large – "Sf" "spider_FS" outperforms "t" "nist_stem" by 30.7% (T queries) and 21.4% (TDN queries), respectively. The individual variants of the Spider and MPRO stemmers show very little difference. For both query lengths, the most aggressive "Full Split" variant of the Spider stemmer slightly outperforms the other two decompounding variants (see also figures 5 and 6). For MPRO, the difference between the "Ms" "MPRO_ls_dec" and "Mu" "MPRO_lu_dec" decompounding variants is somewhat larger (see also figures 7 and 8). Again, the "Ms" "MPRO_ls_dec" variant using the root forms can be regarded as the more "aggressive" variant, and has the upper hand. The reverse, however, is true for the runs that use stemming only: the "mu" "MPRO_lu" run using the lexical units outperforms the "ms" "MRPO_ls" run for both query lengths.

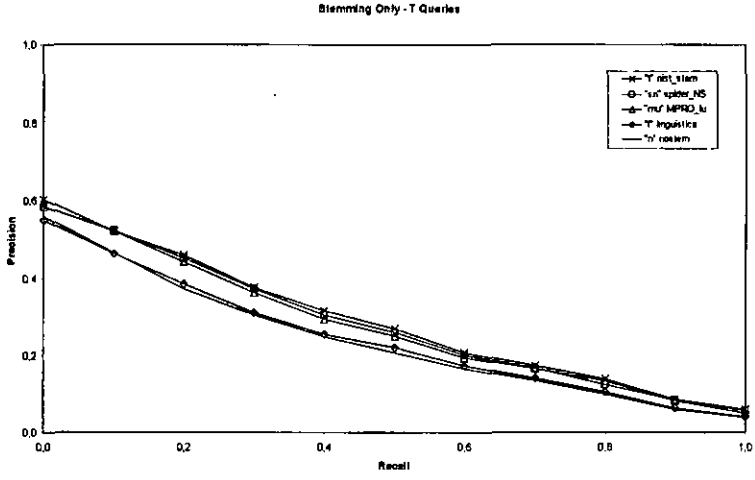

Figure 1. Recall/precision curve for "stemming only" runs (short T queries). The top three methods show very similar performance

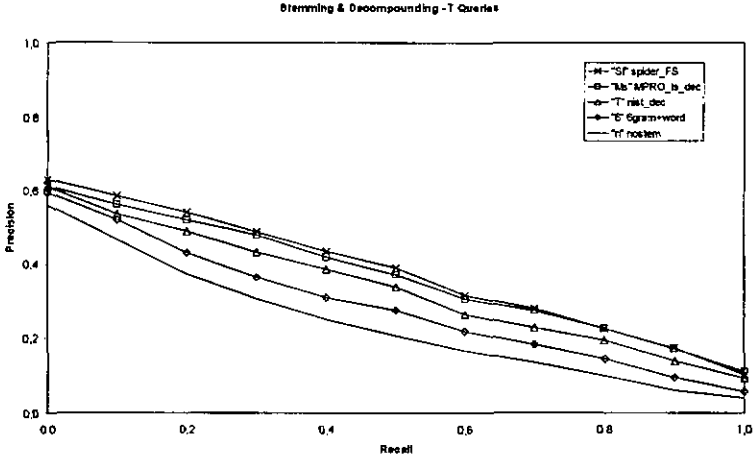

Figure 2. Recall/precision curve for runs using stemming and decompounding (short T queries). The choice of decompounding method produces much larger performance differences than those observed for "stemming only" experiments

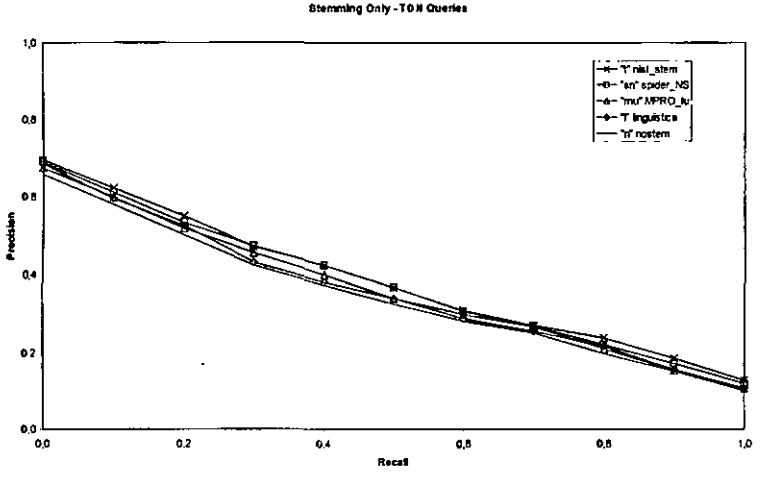

Figure 3. Recall/precision curve for "stemming only" runs (long TDN queries). As for short queries, the performance of the top methods is very similar

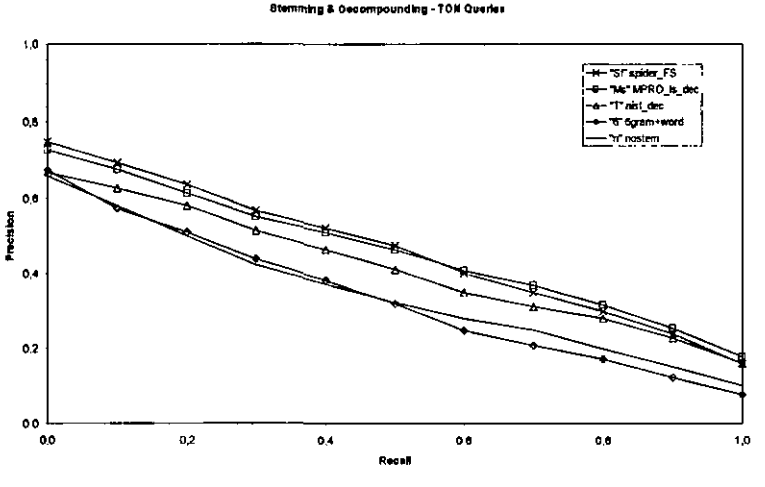

Figure 4. Recall/precision curve for runs using stemming and decompounding (long TDN queries). As for short queries, the performance differences are more evident than for the *"stemming only"* experiments

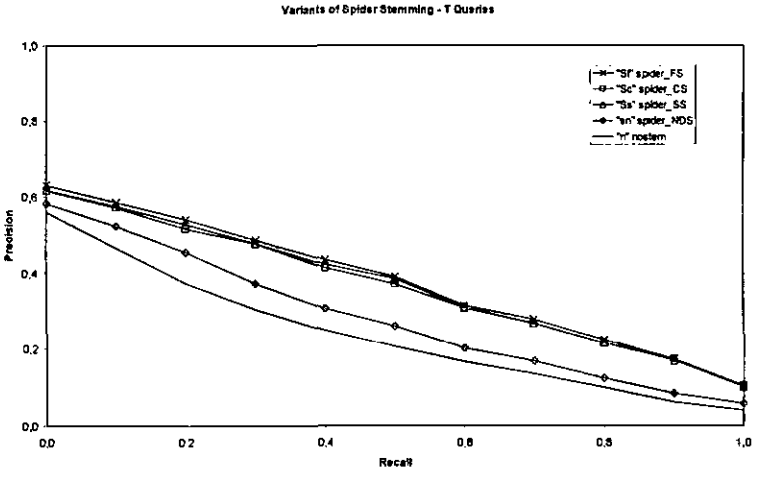

Figure 5. Variants of Spider stemming (short T queries). Given are the three decompounding variants, the stemming only variant, and the baseline (no stemming)

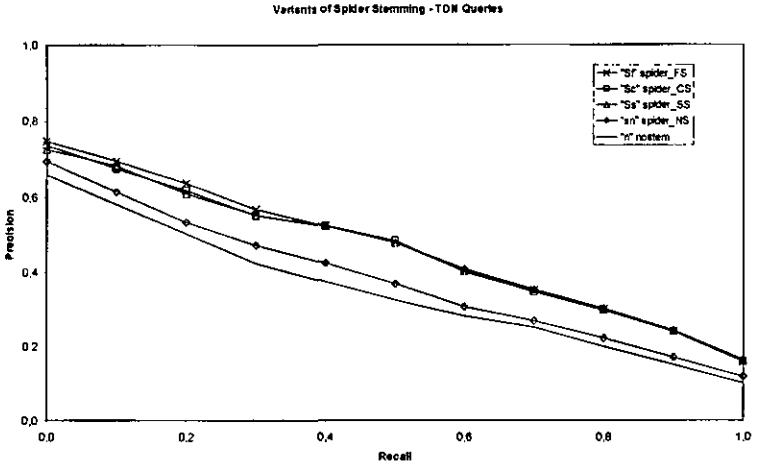

Figure 6. Variants of Spider stemming (long TDN queries). Given are the three decompounding variants, the stemming only variant, and the baseline (no stemming)

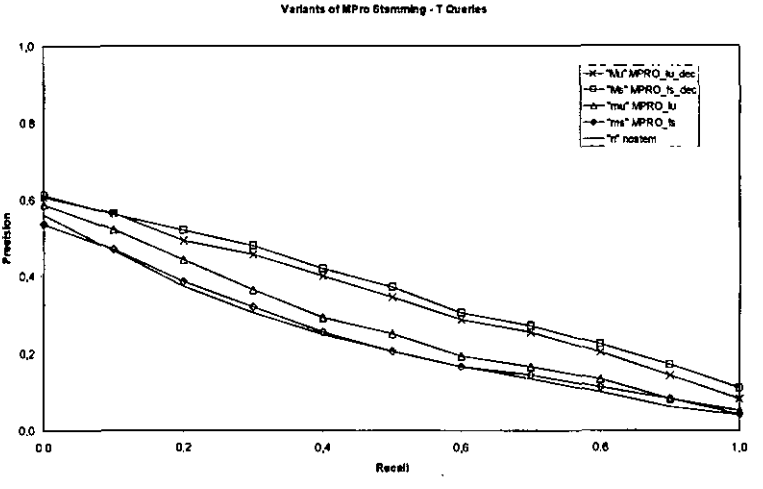

Figure 7. Variants of MPRO stemming (short T queries). Given are the two decompounding variants, the two stemming only variants, and the baseline (no stemming)

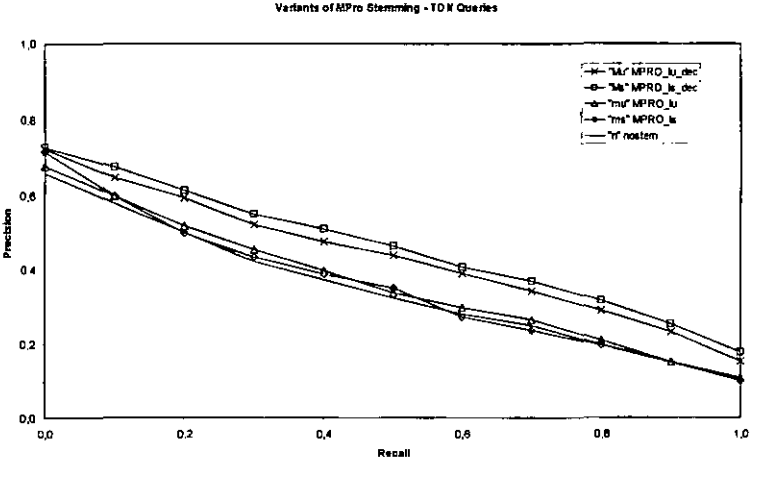

Figure 8. Variants of MPRO stemming (long TDN queries). Given are the two decompounding variants, the two stemming only variants, and the baseline (no stemming)

The top three non-decompounding methods, "t" "nist_stem", "sn" "spider_NS", and "mu" "MPRO_lu", have similar performanee for both query lengths, as does the 6-gram/word-based run, but for T queries only (see also figures 1 and 3). The latter performs clearly worse for TDN queries.

Looking at the total number of relevant dneuments retrieved (Table 3), all methods outperform no stemming for hoth query lengths. Again, the decompounding methods, such as the Spider and MPRO variants and "T" "nist_dec" (upper half of Table 3) perform hest, with Spider and MPRO being very close. In this statistic, the combined 6-gram/word-based run favorably compares to the stemming methods that do not use decompounding (lower half of Table 3). Of the latter methods, "sn" "spider_NS" performs best.

Table 3. Number of relevant doeuments retrieved (total number of relevant documents: 1790)

Run (T)	Recall (T)	Run (TDN)	Recall (TDN)
"Sf" spider_FS	1669 (+30.3%)	"Ms" MRPO_ls_dec	1704 (+11.0%)
"Sc" spider_CS	1665 (+30.0%)	"Mu" MPRO_lu_dec	1696 (+10.5%)
"Ss" spider_SS	1657 (+29.4%)	"Ss" spider_SS	1694 (+10.4%)
"Ms" MPRO_ls_dec	1625 (+26.9%)	"Sc" spider_CS	1693 (+10.3%)
"Mu" MPRO_lu_dec	1624 (+26.8%)	"Sf" spider_FS	1690 (+10.1%)
"T" nist_dec	1577 (+23.1%)	"T" nist_dec	1632 (+6.3%)
"6" 6gram+word	1551 (+21.1%)	"6" 6gram+word	1594 (+3.8%)
"sn" spider_NS	1434 (+11.9%)	"sn" spider_NS	1595 (+3.9%)
"mu" MPRO_lu	1433 (+11.9%)	"t" nist_stem	1595 (+3.9%)
"t" nist_stem	1427 (+11.4%)	"ms" MPRO_ls	1568 (+2.1%)
"ms" MPRO_ls	1398 (+9.1%)	"mu" MPRO_lu	1547 (+0.8%)
"l" linguistica	1300 (+1.5%)	"l" linguistica	1538 (+0.2%)
"n" nostem	1281	"n" nostem	1535

Recently, "high precision measures", i.e. indieations of how well retrieval systems perform if only a small number of highly ranked documents are returned, have become increasingly popular. One main reason for this popularity is the claim that typical web users rarely look past the initial 10 or 20 retrieved items. We use the "Precision @ 10 Documents" measure to investigate if stemming helps in such scenarios. As we indicated in the introduction, stemming is traditionally often seen as a vehicle to increase recall. However, recall plays no important role in a scenario restricted to such few retrieved items. By adopting this precision-centered measure we can observe if stemming approaches help to obtain a better weighting of search terms and therefore a hetter ranking for the top retrieved items. When using the "Precision @ 10 Documents" measure, differences will become most prominent if queries have more than 10 relevant documents. In ease there are very few relevant documents for a query, two systems may obtain the same score even though they retrieve those documents at different ranks. We feel that this is not necessarily a

drawback of the measure when only 10 documents are used, as long as these implications are kept in mind. These considerations may become more important, however, when similar scores for larger document sets (e.g. "Precision @ 100 documents") are calculated.

When we look at the results (see Table 4), it becomes clear that a positive effect is most easily seen for the short T queries. There, stemming seems to provide a large number of extra term matches that indeed positively influences ranking. As for average precision, decompounding approaches exhibit the best performance, with the Spider and MPRO variants producing almost identical scores. The effect is less pronounced for long TDN queries, where no stemming outperforms a number of stemming only approaches (plus the NIST decompounding run and the 6-gram/word-based run). The fact that the longer queries often already contain various different forms of the key search terms means that stemming adds less novel information to be used by the weighting algorithm. Only the more sophisticated Spider and MPRO decompounding can apparently provide good extra matches.

Table 4. Precision @ 10

Run (T)	Prec@10 (T)	Run (TDN)	Prec@10 (TDN)
"Sf" spider_FS	0.3835 (+30.9%)	"Ss" spider_SS	0.4435 (+13.9%)
"Ss" spider_SS	0.3824 (+30.6%)	"Sf" spider_FS	0.4388 (+12.7%)
"Sc" spider_CS	0.3824 (+30.6%)	"Sc" spider_CS	0.4341 (+11.5%)
"Ms" MPRO_ls_dec	0.3800 (+29.7%)	"Ms" MPRO_ls_dec	0.4306 (+10.6%)
"Mu" MPRO_lu_dec	0.3706 (+26.5%)	"Mu" MPRO_lu_dec	0.4247 (+9.1%)
"T" nist_dec	0.3553 (+21.3%)	"T" nist_dec	0.3800 (-2.4%)
"6" 6gram+word	0.3047 (+4.0%)	"6" 6gram+word	0.3412 (-12.4%)
"sn" spider_NS	0.3400 (+16.1%)	"sn" spider_NS	0.3965 (+1.8%)
"mu" MPRO_lu	0.3271 (+11.7%)	"l" linguistica	0.3847 (-1.2%)
"t" nist_stem	0.3259 (+11.3%)	"mu" MPRO_lu	0.3835 (-1.5%)
"l" linguistica	0.2929 (+0.0%)	"t" nist_stem	0.3835 (-1.5%)
"ms" MPRO_ls	0.2906 (-0.7%)	"ms" MPRO_ls	0.3777 (-3.0%)
"n" nostem	0.2929	"n" nostem	0.3894

In the following, we will investigate how well the observations we made in this section hold up when we perform a careful statistical analysis.

8 Statistical Analysis

We used the IR-STAT-PAK tool by Blustein (1998) for a statistical analysis of the results in terms of average precision, which has been tailored for evaluating multiple "TREC-style" experimental results, supporting a wide range of different popular effectiveness measures. This tool provides an Analysis of Variance (ANOVA),

which is the parametric test of choice in such situations but requires checking some assumptions concerning the data. Hull (1993) provides details of these; in particular, the scores in question should be approximately normally distributed and their variance has to be approximately the same for all runs. IR-STAT-PAK uses the Hartley test to verify the equality of variances, and in our case, indicates that the assumption is satisfied, which prompted us to continue with applying the ANOVA test for our data. The program also offers an arcsine transformation, $f(x) = \arcsin\left(\sqrt{x}\right)$, which Tague-Sutcliffe (1997) suggests for use with Precision/Recall measures to better meet the demand for normally distributed data. We used both raw and transformed scores to assure the suitability of our data for an ANOVA test (see Table 6). Alternatively, non-parametric tests, such as the Friedman test, can be used in case the equality of variance assumption is not satisfied. IR-STAT-PAK does not support non-parametric tests.

Investigating the average precision scores we obtain for T queries, the following results for average precision (after Tukey T test grouping, see Table 5 for ANOVA details) are obtained:

Raw data:

"Sf", "Ss", "Ms", "Sc" $\succ$ "t", "6", "sn", "mu", "ms", "l", "n"

"Mu" $\succ$ "6", "sn", "mu", "ms", "l", "n"

"T" $\succ$ "ms", "l", "n"

Arcsine-transformed data:

"Sf", "Ss", "Sc", "Ms" $\succ$ "6", "t", "sn", "mu", "ms", "l", "n"

"Mu" $\succ$ "sn", "mu", "ms", "l", "n"

"T" $\succ$ "ls", "l", "n"

Table 5. ANOVA of average precision (T queries).

Experiment		**T-raw**		**T-arcsine**		
Source	DF	Mean sq	F	Mean sq	F	Critical
Runs	12	0.232	17.577	0.433	17.393	1.7618
Query	84	0.857	64.880	1.428	57.382	1.2818
Error	1008	0.013		0.025		
Total	1104					

The " $\succ$ " symbol denotes a statistically significant difference between two runs with probability p=0.95.

The analysis thus indicates that the all Spider decompounding variants and the respective "Ms" "MPRO_ls_dec" variant significantly outperform all methods without decompounding. They also outperform "Mu" "MPRO_lu_dec" and "T" "nist_dec", but that difference is not statistically significant.

For TDN queries, the findings remain constant. There is a slight shift in ranking of the individual methods, but essentially the same significant differences are found (see also Table 6 for ANOVA details):

TDN queries, raw:

"Sf", "Ms", "Ss", "Sc" $\succ$ "t", "sn", "mu", "l", "ms", "n", "6"

"Mu" $\succ$ "mu", "l", "ms", "n", "6"

"T" $\succ$ "n", "6"

TDN queries, arcsine:

"Ms", "Sf", "Ss", "Sc" $\succ$ "t", "sn", "mu", "ms", "l", "n", "6"

"Mu" $\succ$ "mu", "ms", "l", "n", "6"

"T" $\succ$ "l", "n", "6"

Table 6. ANOVA of average precision (TDN queries).

Experiment		**TDN-raw**		**TDN-arcsine**		
Source	DF	Mean sq	F	Mean sq	F	Critical
Runs	12	0.202	12.801	0.365	13.749	1.7618
Query	84	0.895	56.628	1.454	54.852	1.2818
Error	1008	0.016		0.027		
Total	1104					

The results for the high-precision scores are very interesting, since we detected less pronounced performance differences than for the average precision scores.

For short T queries, the Spider and MPRO variants using decompounding outperform most stemming only variants (and the baseline of not using stemming at all) to a degree that is statistically significant (see also Table 7 for ANOVA details). This is an important finding, given that stemming is still often seen mainly as a

recall-enhancing device for languages such as English. Stemming can thus be valuable even in cases where recall is of secondary importance, such as often encountered by Web-based search services.

Raw data:

"Sf" ≻ "mu", "t", "6", "n", "l", "ms"

"Sc", "Ss" ≻ "t", "6", "n", "l", "ms"

"Ms", "Mu" ≻ "6", "n", "l", "ms"

"T" ≻ "n", "l", "ms"

Arcsine-transformed data:

"Ss", "Sc" ≻ "t", "mu", "6", "l", "ms", "n"

"Sf", "Ms", "Mu" ≻ "6", "l", "ms", "n"

"T" ≻ "l", "ms", "n"

Table 7. ANOVA of Precision @ 10 (T queries).

Experiment		**T-raw**		**T-arcsine**		
Source	DF	Mean sq	F	Mean sq	F	Critical
Runs	12	0.119	10.168	0.281	10.640	1.7618
Query	84	1.308	112.200	2.515	95.228	1.2818
Error	1008	0.012		0.026		
Total	1104					

For long TDN queries, the benefit of stemming and decompounding is less pronounced due to the "richer" representation of the original, unstemmed query when compared to its shorter counterpart (see also Table 8 for ANOVA details). Consequently, fewer significant differences are detected, and no approach or variant outperforms the baseline of not using stemming at all in this scenario.

Raw data:

"Ss" ≻ "ms", "6"

"Sf", "Sc", "Ms", "Mu" ≻ "6"

Arcsine-transformed data:

"Ss", "Ms", "Sf", "Sc", "Mu" ≻ "6"

Table 8. ANOVA of Precision @ 10 (TDN queries).

Experiment		**TDN-raw**		**TDN-arcsine**		
Source	DF	Mean Sq	F	Mean Sq	F	Critical
Runs	12	0.081	5.209	0.161	5.343	1.7618
Query	84	1.215	78.563	2.268	75.044	1.2818
Error	1008	0.015		0.030		
Total	1104					

9 Query-by-Query Analysis

In addition to the overall analysis presented in Sections 7 and 8, we also analyzed the performance of individual queries in more detail. A set of nine queries (10% of the overall query set) were selected for showing conspicuous behavior, either in terms of overall performance or because of outliers by individual methods. Various measures, such as mean average precision, high precision measures (mean of precision at 5, 10, 15 documents retrieved, mean of precision at 10, 20, 30 documents retrieved, and precision at 100 documents retrieved) and uniquely retrieved relevant documents, as well as their respective standard deviations, were used to identify this set of interesting queries. For each query in this set, we determined the most important keywords, and examined all methods with regard to the stems that they produce, and the constituents of the compounds as identified by the various decompounding approaches.

We continue by giving a short analysis of each of the nine selected queries. The title fields of the nine queries are given as heading to each subsection. Key search terms mentioned in the analysis can be contained in any of the fields of the query (title, description, narrative).

9.1. CLEF 2000 Campaign, Topic 1:

"Architektur in Berlin (Architecture in Berlin)"

Approaches performing well: "Sf", "Sc", "Ss", "Ms", "I"

The key search terms in this query are "Architektur" (architecture) and "Berlin". The best results regarding average precision, early precision, and number of relevant documents retrieved come from the Spider decompounding runs ("Sf", "Sc", "Ss"), the "Ms" "MPRO_ls_dec" decompounding run, and Linguistica. Common characteristic of these runs is that they conflate "Architektur"

(architecture) and "Architekt" (architect), which pulls in approximately 10% more relevant documents than the methods that omit this conflation.

9.2. CLEF 2000 Campaign, Topic 5:

"Mitgliedschaft in der Europäischen Union (European Union Membership)"

Approaches performing well: "n", "l", "t", "sn", "mu", "Mu"

Key search terms in this query are "Mitgliedschaft/Mitgliedstaaten/-Mitgliedsstaaten" (membership/member state) and "Europäischen" (European).

A peculiarity of this query is the two different spellings for "Mitglied(s)staaten", once with binding-s (in the "description" part), and once without (in the "narrative" part). While probably not intended by the topic creator, this is not strictly speaking a typo, since both forms are valid. The "Sf", "Sc", "Ss", "ms", "Ms" and "Mu" methods conflate the two variations, whereas the other methods produce two distinct indexing features. However, conflation of the two different spellings for "Mitglied(s)staaten" does not seem to be essential for good retrieval.

Generally speaking, decompounding seems to be counterproductive for this query. Apart from the exception of "Mu", all decompounding methods produce results that are substantially worse (in the case of the Spider stemmer, between 48% and 66%) than the ones returned by methods without decompounding (or even without stemming at all). It seems that the core concept of "member state" gets diluted too much when its constituents are added as independent search terms. From a linguistic point of view, "Mitgliedschaft" should not be split, contrary to the behavior of some of the approaches, since "schaft" is in this case a derivational suffix.

9.3. CLEF 2000 Campaign, Topic 26:

"Nutzung von Windenergie (Use of Wind Power)"

Approaches performing well: all decompounding approaches.

Key search term is "Windenergie" (wind power).

Methods using decompounding outperform other methods substantially. Decomposition of "Windenergie" into "Wind" and "Energie" results in the retrieval of nearly twice as many relevant documents. Stemming proper makes little difference, since the keywords occur in their base forms in the query statement and in the relevant documents.

9.4. CLEF 2000 Campaign, Topic 34:

"Alkoholkonsum in Europa (Alcohol Consumption in Europe)"

Approaches performing well: all decompounding approaches.

Key search terms are "Alkoholkonsum/Alkoholmisshrauch" (alcohol consumption/alcohol abuse) and "Europa" (Europe). This is another query where decompounding is crucial. Methods with decompounding (most important to find other compounds consisting of "alkohol" as constituent) retrieve 50% additional relevant documents on average.

9.5. CLEF 2000 Campaign, Topic 35:

"Wölfe in Italien (Wolves in Italy)"

Approaches performing well: no clear picture

Keywords are "Wölfe" (wolves) and "Italien/Italiens" (Italy, of Italy). This query is a special case: it has only one relevant document in the collection. Stemming differences across methods are minor, mainly in terms of some less frequent word forms being conflated or not. All methods find the one relevant document, but average precision numbers fluctuate wildly depending on the rank at which the document is retrieved. These large fluctuations are mainly due to the properties of the average precision measure, and we believe them to obscure the performance of the different stemming methods in this case.

9.6. CLEF 2001 Campaign, Topic 6:

"Embargo gegen den Irak (Embargo on Iraq)"

Approaches performing well: "n", "l"

Key search terms are "Embargo" (embargo), "Irak/irakischen" (Iraq, Iraqi) and "Bevölkerung" (population). The best two methods for this query are "no stemming" and "Linguistica", which both do no or very little stemming. This is apparently due to the fact that all keywords are present in their base forms in the query statement as well as in the documents, with stemming only introducing further noise.

9.7. CLEF 2001 Campaign, Topic 21:

"Ölkatastrophe in Sibirien (Siberian Oil Catastrophe)"

Approaches performing well: "Sf", "Se", "Ss"

Keywords are "Ölkatastrophe/Ölpipeline/Ölleitung" (oil catastrophe/oil pipeline/oil pipeline), "Umweltaspekte" (environmental aspects), "Ökosystem" (ecological system) and "Sibiren/Sibiriens" (Siberia/of Siberia). The Spider

decompounding methods are the most aggressive in producing conflations: they are the only methods that decompound all the compound keywords except "Ökosystem", which is not decompounded by any method. MPRO splits "Oelleitung" but not "Ölleitung" because of the umlaut treatment in the version we used.

In general, this query demonstrates how decompounding helps to absorb differences in writing styles. For instance, the query uses "Ölpipeline" in the description part and "Ölleitung" in the narrative part of the query, which are essentially synonyms.

9.8. CLEF 2002 Campaign, Topic 25:

"Schatzsucher (Treasure Hunting)"

Approaches performing well: decompounding methods and "t"

Key search terms are "Schatzsucher/Schatzsuche" (treasure hunter/treasure hunting), "Reliquien" (relics), "Schiffen" (ships).

To achieve high performance, it seems to be essential to establish a relation between "Schatzsucher" and "Schatzsuche". The decompounding methods do this via the component "Schatz" (treasure), whereas "t" "NIST_stem" conflates the two words to a common stem ("Schatzsu"). All other methods that fail to establish such a relation perform substantially worse. Another nicety of this query is the accidental conflation of "finde" (look for) and "Fund" (a find) to "find" by the Spider variations which positively affects the result.

9.9. CLEF 2001 Campaign, Topic 27:

"Schiffskollisionen (Ship Collisions)"

Approaches performing well: "Sf", "Sc", "Ss"

Key search terms are "Schiffskollisionen" (ship collisions), together with its constituents "Kollisionen" (collisions), "Schiffen" (ships) and "Zusammenstößen" (collisions) in the narrative part of the query. The best results come from the Spider variations that use decompounding. These are the only methods that split "Schiffskollisionen", which seems to be essential for good retrieval performance for the short T version of this query. In the TDN run the constituents occur as separate words resulting in much better performance for all systems except for no stemming and Linguistica.

In summary, we found that stemming is in most cases beneficial independent of the method applied. Approaches conducting a balanced conflation, i.e. avoiding both understemming and overstemming as best as possible, seem to be superior.

Cases where no stemming outperforms most or all of the stemming methods exist when all the keywords are already present in their base forms (and occur rarely in inflected forms) and represent simple concepts (no compound nouns) (e.g. Topic 6, 2001 Campaign). Furthermore, queries consisting of proper names show such effects, especially in short queries. For long queries, in these cases, the noise introduced by stemming cancels out the benefits from matching additional word forms.

Not surprisingly, we found a number of queries where the key to successful retrieval is clearly the ability to split compounds, since the constituents are important to the respective topic statements, and the compound commonly is transcribed in phrasal form in the documents.

However, there are also cases where decompounding is counterproductive (such as topic 5, 2000 campaign). Thus, it is important to avoid the decomposition of "false" compounds such as "Mitgliedschaft" (membership) and of compounds for which the splitting causes a shift of meaning such as "Frühstück" (breakfast), composed of "Früh" (early) and "Stück" (piece). Also, decompounding of words that are rarely written in phrasal form e.g. "Mitgliedstaaten" (member states) may negatively impact weighting.

There may be potential in development of corpus-based decompounding methods in this regard, as it may well depend on the document collection itself if a key concept of the query gets diluted by compound splitting, or if additional, good matches are found. An interesting line of research may be to consider how to automatically infer from corpus statistics how well compounds are represented by their constituents, and use this as a factor in the decision of whether to apply compound splitting or not. Similar considerations apply for stemming proper, where especially the effect of handling derivational suffixes may differ depending on the terms or even the query and collection used for searching.

Lastly, we found some oddities, which were either due to the evaluation measure used (e.g. mean average precision in case that there are very few relevant documents for a query) or to some erroneous conflations, which turned out to produce good matches by chance.

10 Conclusions

In our experiments, we demonstrated that stemming is useful for German text retrieval in most cases. Compared to a system without stemming, we obtained performance gains measured in mean average precision of up to 23% for short (T) and up to 11% for long queries (TDN). For recall we observed improvements of 12% for T, and 4% for TDN. Exceptions where stemming is detrimental are queries comprising proper names or other invariant words (where no inflection exists, or the inflection is rarely used). The different stemming methods show only weak significant differences, mainly in the high recall range.

An important finding of this study is that decompounding contributes more to performance improvement than stemming, 16% to 34% for short and 9% to 28% for long queries. In contrast to stemming, we observed larger differences between individual methods implying that it is important to carefully choose the right degree of decompounding. There are two main goals with respect to decompounding: Firstly to produce as many extra matches for retrieval as possible, and secondly to avoid inappropriate splittings. The two best runs, "Sf" "spider_FS" and "Ms" "MPRO_ls_dec", put different emphasis on these goals. "Sf" "spider_FS" splits aggressively, and "Ms" "MPRO_ls_dec" avoids linguistically incorrect splittings. However, they show almost the same performance.

Looking beyond the popular average precision measure, we have investigated the impact of stemming in a "high precision scenario", such as often assumed in Web settings. Traditionally, stemming is often seen mainly as a vehicle to enhance recall. However, our experiments indicate the best stemming and decompounding approaches lead to better precision at 10 documents retrieved than using no stemming. Some of the differences are statistically significant.

This report makes a major contribution to the discussion on the benefits of stemming that goes beyond the focus on the German language by employing a complete spectrum of stemming and decompounding approaches on a large reliable corpus, which allowed us to conduct a thorough analysis of the results. One outcome of such a broad analysis is that stemming can be done using comparatively simple approaches such as the NIST German stemmer, which showed competitive performance. In contrast, decompounding requires a more sophisticated approach, either providing a sufficient lexical coverage for aggressive splitting or linguistic knowledge to achieve correct handling of compounds.

Furthermore, we found that for German as a morphological rich language, purely language independent methods do not give competitive performance. The exception is the 6-gram/word-based run, but only for short queries. Even so, the results are still significantly worse than for other methods using decompounding, and the increased index size from using *n*-grams is a substantial drawback.

Future work could validate these findings that we obtained for unrestricted text for settings where domain specific terminology is frequent. Such terminology can have properties that may necessitate adaptations especially related to decompounding (e.g. medical terms of Latin origin, or names of chemical substances).

Acknowledgements

We would like to thank IAI for the opportunity to use MPRO. Thanks go to John Goldsmith for providing the Linguistica software, and the CLEF consortium and its data providers for the construction of the test collection. Decompounding for the NIST stemmer is based joint work with Paul Over from NIST. Jacques Savoy and three anonymous referees provided detailed comments and suggestions that helped improve the paper.

References

Blustein J (1998) IR STAT PAK. URL: http://www.csd.uwo.ca/~jamie/IRSP-overview.html (last visit 11/19/2002).

Braschler M and Schäuble P (2001) Experiments with the Eurospider Retrieval System for CLEF 2000. In Peters C. (Ed.): Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, pp. 140-148.

Choueka Y (1992) Responsa: An Operational Full-Text Retrieval System With Linguistic Components for Large Corpora. In: Computational Lexicology and Lexicography: a Volume in Honor of B. Quemada.

Frakes WB (1992) Stemming Algorithms. In: Frakes, W. B. and Baeza-Yates, R. (Eds.): Information Retrieval, Data Structures & Algorithms. Prentice Hall, Eaglewood Cliffs, NJ, USA. pp. 131-160.

Frisch E and Kluck M (1997) Pretest zum Projekt German Indexing and Retrieval Testdatahase (GIRT) unter Anwendung der Retrievalsysteme Messenger und freeWAISsf. IZ Arbeitsbericht Nr. 10, GESIS IZ Soz., Bonn, Germany (in German).

Goldsmith J (2001) Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics, 27(2):153-198. MIT Press. Available at: URL: http://humanities.uchicago.edu/faculty/goldsmith/Linguistica2000/ (last visit 11/19/2002)

Harman D (1991) How Effective is Suffixing?. In Journal of the American Society for Information Science, 42(1):7-15.

Harman D (1997) The TREC Conferences. In Sparck-Jones, K. and Willett, P. (Eds.): Readings in Information Retrieval, Morgan Kaufmann Publishers, San Francisco, CA. USA.

Hull DA (1996) Stemming Algorithms - A Case Study for Detailed Evaluation. In Journal of the American Society for Information Science 47(1):70-84.

Hull DA (1993) Using Statistical Testing in the Evaluation of Retrieval Experiments. In Proceedings of the 16th ACM SIGIR Conference, Pittshurg. USA.

Hull DA, Grefenstette G, Schultze BM, Gaussier E, Schütze H, Pedersen O (1996) Xerox TREC-5 Site Report: Routing, Filtering, NLP and Spanish Tracks. In Proceedings of the Fifth Text Retrieval Conference (TREC 5), Gaithersburg. USA.

Kraaij W and Pohlmann R (1996) Using Linguistic Knowledge in Information Retrieval. OTS Working Paper OTS-WP-CL-96-001, University of Utrecht, The Netherlands.

Kraaij W and Pohlmann R (1996) Viewing Stemming as Recall Enhancement. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Zurich. Switzerland.

Krause J. and Womser-Hacker C (1990) Das Deutsche Patent-informationssystem. Entwicklungstendenzen, Retrievaltests und Bewertungen. Carl Heymanns (in German).

Lovins JB (1968) Development of a Stemming Algorithm. In Mechanical Translation and Computational Linguistics, 11(1-2):22-31.

Maas D (1996) MPRO – Ein System zur Analyse und Synthese deutscher Wörter. in R. Hauser (ed.): Linguistische Verifikation, Max Niemeyer Verlag, Tübingen (in German).

Mayfield J, McNamee P and Piatko C (1999) The JHU/APL HAIRCUT System at TREC-8. In Proceedings of the Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-246, pp. 445-451

Monz C, and de Rijke M (2002) Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German and Italian. In Peters, C., Braschler, M., Gonzalo, J. and Kluck, M. (Eds): Evaluation of Cross-Language Information Retrieval Systems. CLEF 2001, Lecture Notes in Computer Science, LNCS 2406, pp. 262-277.

Moulinier I, McCulloh JA, Lund E: West Group at CLEF 2000 (2001) Non-English Monolingual Retrieval. In Peters C. (Ed.): Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, pp. 253-260.

Popovic M and Willet P (1992) The effectiveness of stemming for natural-language access to Slovene textual data. In Journal of the American Society for Information Science, 3(5):384-390.

Porter MF (1980) An Algorithm for Suffix Stripping. In Program, 14(3):130-137. Reprint in: Sparck Jones. K. and Willett, P. (Eds.): Readings in Information Retrieval. Morgan Kaufmann Publishers, San Francisco. CA, USA, pp. 313-316.

Ripplinger B (2002) Linguistic Knowledge in Cross-Language Information Retrieval, PhD Thesis, Herbert Utz Verlag. Munich, Germany.

Savoy J (1999) A stemming procedure and stopword list for general French corpora. Journal of the American Society for Information Science, 50(10):944-952.

Savoy J (2003) Cross-Language Information Retrieval: Experiments Based on CLEF 2000 Corpora. Information Processing & Management, 39(1):75-115.

Singhal A, Buckley C and Mitra M (1996) Pivoted Document Length Normalization. In Proceedings of of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland.

Sheridan P and Ballerini JP (1996) Experiments in Multilingual Information Retrieval using the SPIDER System. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland.

Tague-Sutcliffe J (1997) The Pragmatics of Information Retrieval Experimentation, Revisited. In Sparck-Jones, K. and Willett, P. (Eds.): Readings in Information Retrieval, Morgan Kaufmann Publishers, San Francisco, CA, USA.

Tomlinson S (2002) Stemming Evaluated in 6 Languages by Hummingbird SearchServer™ at CLEF 2001. In Peters, C., Braschler, M., Gonzalo, J. and Kluck, M. (Eds): Evaluation of Cross-Language Information Retrieval Systems. CLEF 2001, Lecture Notes in Computer Science, LNCS 2406, pp. 278-287.

Wechsler M, Sheridan P and Schäuble P (1997) Multi-language text indexing for internet retrieval. In Proceedings of the 5th RIAO Conference. Computer-Assisted Information Searching on the Internet, Montreal, Canada, pp. 217-232.

Womser-Hacker C (1989) Der PADOK-Retrievaltest. In "Sprache und Computer" Band 10, Georg Olms Verlag (in German).

Combination Approaches for Multilingual Text Retrieval

Martin Braschler

Eurospider Information Technology AG
Schaffhauserstrasse 18
CH-8006 Zurich, Switzerland
martin.braschler@eurospider.com

Université de Neuchâtel
Institut Interfacultaire d'Informatique
Pierre-à-Mazel 7
CH-2001 Neuchâtel, Switzerland

Abstract. We describe the Eurospider component for Cross-Language Information Retrieval (CLIR) that has been employed for experiments at all three CLEF campaigns to date. The central aspect of our efforts is the use of combination approaches, effectively combining multiple language pairs, translation resources and translation methods into one multilingual retrieval system. We discuss the implications of building a system that allows flexible combination, give details of the various translation resources and methods, and investigate the impact of merging intermediate results generated by the individual steps. An analysis of the resulting combination system is given which also takes into account additional requirements when deploying the system as a component in an operational, commercial setting.

Keywords: Cross-Language Information Retrieval, combination methods, merging, operational systems

1 Introduction

The Eurospider components for Cross-Language Information Retrieval (CLIR) rely heavily on the combination of multiple simpler CLIR approaches. We discuss the implications of combining different translation methods, and present the supporting framework that we have implemented. Eurospider has participated with considerable success in all three CLEF campaigns held to date, and we analyze the results that have been obtained.

Definitions

The CLEF campaign offers both multilingual and bilingual cross-language information retrieval (CLIR) evaluation. Multilingual retrieval in CLEF is defined as the task of retrieving documents from a multilingual collection containing articles from newspapers, newsmagazines and news agencies written in any of four (English, French, German, Italian in CLEF 2000) or five (as before, plus Spanish in CLEF 2001 and CLEF 2002) languages. Bilingual retrieval is a "simpler" task, retrieving documents from a monolingual collection written in a language different from the one used for query formulation.

Overview

While initial experiments are much older (see e.g. Salton 1970), interest in CLIR has substantially increased after 1996. The CLIR approaches that have subsequently been proposed can be classified in different ways. Braschler et al. (1998a) suggested a classification in terms of what information is translated to "bridge the language gap" in CLIR: the queries (query translation: QT), the documents (document translation: DT), or both[1]. Oard (1997) divided approaches into those using controlled vocabulary vs. those working on free text, then further refining this classification by the type of resources used for classification: corpus-based or knowledge-based.

Current state-of-the-art translation approaches to free-text CLIR, such as those used by the vast majority of CLEF participants, fall broadly speaking into three categories: approaches using machine translation (MT) systems, approaches using machine-readable dictionaries (MRD) and approaches using corpus-based resources derived from suitable training data. Eurospider has used methods from all three fields in the TREC and CLEF campaigns.

Currently, query translation (QT) seems to be more popular than document translation (DT), mainly because of perceived limitations in scalability in the latter approach. We have attempted the combination of both methods.

Combination of Multiple Simpler Approaches

Much of the early work on CLIR after 1996 obtained an effectiveness that was far below what was observed for monolingual retrieval (for an overview of this type of work, see e.g. Grefenstette 1998). This substantial drop-off in effectiveness was found to be fairly consistent across different approaches, such as using MT or MRDs. As a consequence, we started early on to search for combination translation approaches in the hope of being able to substantially narrow the gap between CLIR

[1] Incidentially, a small subfield of CLIR has since sprung up that tries to build systems that use no translation at all, adding a fourth dimension to this kind of classification, see e.g. McNamee and Mayfield (2002a).

and monolingual retrieval (Braschler and Schäuble 1998, Braschler et al. 1998b). Using a combination approach is analogous to efforts in monolingual IR to combine multiple sources of evidence (Belkin et al. 1993b, Fox and Shaw 1993).

The combination of multiple translation methods also decreases the risk of retrieval failure due to missing or incorrect translations. A substantial part of the difference in average retrieval effectiveness between monolingual and cross-language IR can be attributed to a few (negative) outliers. Such outliers are less likely if multiple sources of translations are used, and retrieval performance becomes more robust. A combination of multiple translation resources makes it possible to greatly increase the lexical coverage of a CLIR system, which is essential if a wide range of queries are to be processed, including technical terminology, fashionable terms and names.

Merging of Different Languages

When going from bilingual CLIR to truly multilingual CLIR, involving more than one target language, the additional question arises of whether to handle the multiple target languages simultaneously or each language separately and then later merge the individual results. Systems that handle all languages simultaneously often have to translate all items into some form of common interlingua in order to build an unified index. By handling the languages in pairs this added difficulty can be avoided at the expense of having to deal with a number of intermediate results. However, combination approaches lend themselves naturally to this type of architecture, as they typically already require merging of intermediate results coming out of the different types of translation resources.

In both cases, serious issues arise with respect to weighting between the different languages, an aspect that is as such absent in the monolingual case, although there are parallels to the monolingual problem of merging databases indexed by different, autonomous systems (see e.g. Du and Callan 1998). If attempting to merge results from several bilingual retrieval runs on documents in different languages, the estimates of relevance obtained through the weighting scheme will typically not be comparable, since the different vocabularies of the languages yield different term frequencies. Similarly, when using a single, unified index to permit parallel retrieval, different numbers of documents across the languages will produce incompatible frequency counts.

Key Requirements

In summary, a combination approach to multilingual CLIR in CLEF has to address

1. how to combine different translation approaches, and which types of resources to use (MRD, MT, corpus-based)
2. what to translate (query, documents, both, none)
3. how to handle the multiple languages (simultaneously or separately)

Related work

Overviews on different approaches to cross-language information retrieval can be found e.g. in Braschler et al. (1998) and Oard (1997). Good examples of the use of all three main types of translation resources commonly used for CLIR (machine translation, machine-readable dictionaries and corpus-based data structures) can be found in experiments by CLEF participants. For machine translation, see e.g. Jones and Lam-Adesina (2002) and Figuerola et al. (2001), for machine-readable dictionaries and thesauri, see e.g. Hedlund et al. (2001) and Gey et al. (2002)., and for corpus-based CLIR see e.g. Hiemstra et al. (2001) and Nie and Simard (2002).

Combination approaches, such as those used in our system, make use of two or more of these types of translation resources. University of California at Berkeley has also used combination approaches since CLEF 2000 with considerable success (Gey et al. 2001, Chen 2002a). Further experiments with combination approaches were also carried out by Université de Neuchâtel (Savoy 2002a, 2002b) and Thomson Legal (Moulinier and Molina-Salgado 2002). All these papers also address the question of how to merge the intermediate results obtained by the individual components of the respective systems. Today, a majority of participants in the multilingual track of CLEF use combination approaches, including the groups scoring in the top three in terms of effectiveness. Our work differs from the work of most other CLEF participants in that we have used translation resources of all three types: MT, MRD and corpus-based, and that we have combined both query translation and document translation, resulting in a broad understanding of the properties of such combinations and in a system and a framework that can handle the widest possible range of translation resources. While Chen (2002a) reports on the combined use of bilingual dictionaries, parallel corpora and mining of search engine results for query translation, he does not use document translation. Other CLEF participants to try a combination of document and query translation are McNamee and Mayfield (2002b), but this was in unofficial experiments. They used a bilingual dictionary derived from a parallel corpus for translation.

Furthermore, our system is built with applications in operational, commercial settings in mind, therefore emphasizing some issues that we identified as being important in such situations, such as good lexical coverage, the avoidance of negative outliers and applicability to a broad range of domains. These criteria are usually not in the main focus of the CLEF evaluation.

In monolingual information retrieval, there is also earlier work which has relevance to our experiments on combination of approaches and on merging of

languages. In these studies, the corresponding problems are also referred to as "data fusion" (merging of multiple result lists calculated on the same document collection), "collection fusion" (merging of multiple result lists calculated on disjoint document collections), and "query combination" (merging of multiple query representations before retrieval). Much of the work on combining multiple result lists coming from the same document collection stems from the observation that the estimation of relevance of documents improves when multiple sources of evidence are available. Saracevic and Kantor (1988) observed in their studies with multiple manually generated query formulations that "[..] the more often an item was retrieved the more the odds shifted in favor of relevance". Similar work by Belkin et al. (1993a) with multiple Boolean query formulations showed that "progressive combination of query formulations leads to progressively improving retrieval performance". The same report also gives a good overview of early research on this phenomenon (see also Belkin et al. 1993b, Bartell et al. 1994). Early experiments on combining the output of different weighting schemes in different ways were conducted by Fox et. al (1992), Fox and Shaw (1993) and Lee (1995), generally with encouraging results. The problem of merging result lists from disjoint document collections, namely those formed by the different target languages, is closely related to the problem of merging output from multiple, separate document collections in the monolingual case. Earlier work on the monolingual problem can be found e.g. in Callan et al. (1995) and Voorhees et al. (1995), both of which also provide good introductions to the problem.

2 Individual Translation Resources and Approaches

2.1 Machine Translation (MT)

Machine translation systems, i.e. systems that attempt to automatically translate texts from one language to another, seem like an obvious choice for CLIR. However, there are several obstacles to using MT for CLIR. Firstly, MT systems attempt to determine the correct word sense for translation by using context analysis. If an MT system is applied to query translation, there may be little or no context available that the can be exploited. Furthermore, MT systems depend on correct, grammatical input. Queries are often only a string of search terms enumerated by the user. Lastly, the diversity of concepts entered by a user is nearly endless – users search for names, current affairs, rare items, "fashionable" terminology, etc. – all things that are hard to translate for an MT system that usually uses some form of dictionary as resource. MT would appear to be a more attractive choice for document translation – there is more context available to be exploited in a document, and grammar is typically cleaner. However, the main problem with document translation remains – if the MT system cannot translate a certain term, later searches will come up without result.

Eurospider has used MT systems both for document translation (DT) and query translation (QT) in CLEF experiments. For DT, all documents were processed

offline, and then inserted in their translated form into the system. QT would require a direct integration of the MT and retrieval systems. We have opted for a looser coupling by externally translating the query and then feeding the result to the retrieval system. No manual modifications to the query were made of any kind.

MT systems are only available for a limited number of "popular" language pairs. For our CLEF experiments, we were not able to locate a suitable German/Spanish MT system, so German/Spanish MT was simulated by first using MT to translate to English, and then having the system translate the resulting English output to Spanish. We feel this illustrates well the kind of difficulties that will be faced when attempting to deploy a CLIR system for lesser used language pairs. Clearly, MT has deficits in such situations.

2.2 Similarity Thesaurus (ST)

We have used corpus-based approaches with considerable success for CLIR. Corpus-based resources are automatically derived from suitable training data. Our experiments with corpus-based methods in CLEF use a so-called "similarity thesaurus" (ST) (Qiu and Frei 1993).

To build a multilingual similarity thesaurus, the training data is constructed to consist of pairs of documents with related content in the desired languages. For example, one pair of documents may consist of one document each for both languages about election results (Braschler and Schäuble 1998). Such document pairs can be found by obtaining collections of similar type, e.g. news articles in several languages, and then "aligning" individual documents. The alignment process works by picking out "indicators", i.e. special characteristics of a document, and comparing them with the indicators of documents in the other language. Indicators can be of different type, e.g.

1. Date, author, source of a document
2. Classifiers assigned to a document, if available
3. Proper names in documents that are spelled (nearly) identically across languages
4. Numbers and acronyms
5. Words contained in a small core vocabulary for which a dictionary is available before the alignment process ("seed dictionary").

We have shown that it is possible to align documents from independent sources based only on indicators such as those described above (see figure 1) as long as the documents are from the same domain (Braschler and Schäuble 1998). The similarity

thesaurus construction process that relies on the training data produced through this form of document alignment is relatively robust with regard to the quality of the alignments. There is no need for alignment on paragraph or even sentence level. It is sufficient if documents that are paired treat loosely the same subject. This is an essential advantage compared to some other corpus-based approaches that rely on training on so-called "parallel corpora", which contain every item in all languages that are to be covered. While it is hard and usually expensive to obtain this form of parallel training collection, it is substantially easier to find related collections in the necessary languages, and convert them to training data through the alignment process. Interesting developments that may help increase the availability of suitable parallel corpora for training comes from work on mining the World Wide Web for suitable documents (Hiemstra et al. 2001, Nie and Simard 2002). We will explore training the similarity thesaurus on such corpora in the future.

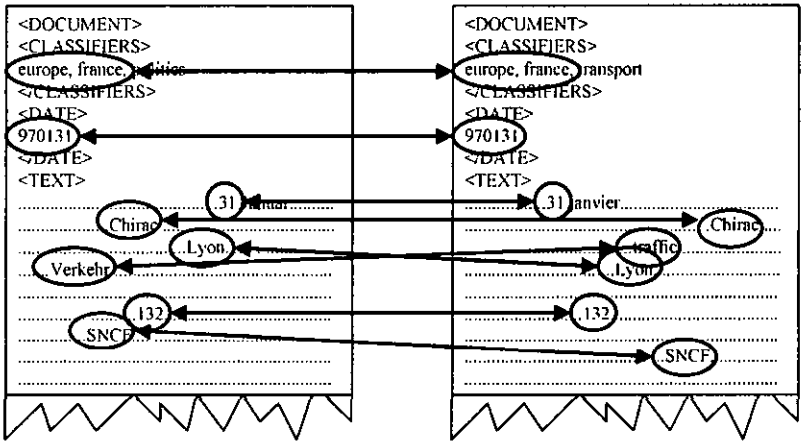

Figure 1. Alignment of documents. The two documents are aligned based on matching classifiers, dates, numbers, names, and a few terms coming from a seed dictionary.

The similarity thesaurus is constructed by calculating the similarity of every term in one language with all terms in the other language based on their occurrences in pairs of aligned documents. This process is similar to classical document retrieval. Where for document retrieval terms are used to find the documents they are contained in, we find those terms that are represented by the same documents, essentially using the documents to retrieve the terms. Indeed, it is possible to use the concept of a "dual space", where the roles of documents and terms are reversed for retrieval, and then use classical IR weighting schemes in their adapted form for similarity calculation. We use the standard *tf.idf* weighting adapted for the dual space ("*ff.iif*") to calculate the similarities between terms (see Formula 1). For every

term. the top thirty most similar terms in the target language are then saved to the resulting thesaurus.

$$sim(\varphi_h, \varphi_i) := \frac{\sum_{d_j \in \varphi_h \cap \varphi_i} a_{j,h} * a_{j,i}}{\sqrt{\sum_{d_j \in \varphi_h} (a_{j,h}^2)} * \sqrt{\sum_{d_j \in \varphi_i} (a_{j,i}^2)}},$$

$$a_{j,i} := ff(d_j, \varphi_i) * iif(d_j), \quad iif(d_j) := \log\left(\frac{1+\Phi}{1+|d_j|}\right)$$

Formula 1. The *ff.iif* weighting used for similarity thesaurus calculation. The similarity between two terms is the angle between two vectors composed of *ff.iif* weights, i.e. the product of the feature frequency and of a value that is analogous to the *idf* used in classical information retrieval.

Since these "similarities" are derived on a purely statistical basis, there is no guarantee that actual translations are found. Typically, however, terms are found that describe a meaning identical or very closely related to that of the original term, but in the target language. Clearly, such terms can then be used to search for the concept expressed by the query. An added benefit of the statistical approach underlying similarity thesaurus construction is that it is not only possible to calculate similarities on a term/term basis, but also for multiple terms in one go – capturing the "concept" of the entire query as a whole. In this way, some of the ambiguities of individual terms can be avoided when the query contains more than one search term.

Probably the most exciting aspect of corpus-based resources is that when deriving suitably powerful resources, such as the similarity thesaurus, the coverage is only limited by the scope of the training data. By feeding large amounts of appropriate training data, the resource can cover vocabulary that goes far beyond the core vocabulary usually covered by even the largest manually constructed translation resources. Even more interesting, this extra vocabulary can contain items such as names, technical terminology and "fashionable expressions", which are usually very important for operational search systems.

2.3 Machine-readable dictionaries (MRD)

Machine-readable dictionaries (MRDs) are close relatives to the usual, well-known printed bilingual dictionaries. They contain for each entry in the source language one or more possible translations. Before use in retrieval systems, extra annotations

present in printed dictionaries, such as case or gender information, must be removed, and abbreviations have to be expanded.

The main problem in using machine-readable dictionaries for CLIR is the ambiguity of many search terms. A CLIR system using MRDs can either use all possible translations, relying on a robust weighting mechanism during subsequent retrieval, or attempt to chose the most suitable translation, either by using frequency information if available, or by employing word-sense disambiguation. Results regarding the performance of these alternatives are inconclusive, and seem to heavily depend on the subsequent retrieval process.

We have used MRDs only in CLEF 2001, where we found no additional benefit with respect to using a combination of machine translation and similarity thesauri. This is probably due to the fact that an MT system tends to provide a similar lexical coverage to that of a typical MRD covering general terminology. We will not investigate MRDs in-depth in this paper. However, the integration of MRDs is essential for operational settings, where customer-specific translation resources in the form of dictionaries and thesauri have to be employed.

3 Combination Approaches

The Eurospider CLIR components allow flexible combination of multiple translation resources and retrieval approaches. By allowing such combinations, we aim to

1. Maximize lexical coverage for translation

   Lexical coverage remains a central problem in CLIR systems, since a query that cannot be translated results in what is essentially worthless retrieval results (if retrieval results are returned at all). Not only will average performance of the system suffer in CLEF-style evaluations, but the acceptance of real-world users for systems that cannot handle a substantial portion of queries is low – users are very sensitive to the worst experiences they have with a system (especially those made early in their use of the system). Furthermore, a free-text retrieval system is typically faced with a wide range of different search requests, from well-formed grammatical sentences to a few hastily entered terms mixed from names and popular expressions. Not all forms of translation resources are equally suited to handle these different requests.

2. Use multiple translation approaches/resources to avoid negative outliers

   Lexical coverage is central to avoiding negative outliers. However, such outliers can also stem from inappropriate translation. Translating the query using multiple different translation resources and approaches minimizes the

chance of very pronounced outliers. There is a danger, however, of also minimizing the effect of positive outliers.

3. Maximize the use of all available resources, allowing broad applicability to a range of operational settings

Beyond some of the most popular languages pairs, usually involving English, resources are still typically scarce. A system should be able to work with whatever is available. Similarly, there are few resources covering special terminology and fashionable terms. Again, a system should be ready to use whatever is available.

The last point is especially challenging, in that it requires a system that can potentially work with only a fraction of the resources that are available for popular language pairs.

The Eurospider system allows flexible combination of translation resources and approaches either through direct combination of translation resources, or through the merging of intermediate retrieval results.

Combination approaches have shown to be effective for multilingual retrieval in the past CLEF campaigns, with the best entries for the 2002 multilingual track coming from groups using such approaches (Savoy 2002b, Chen 2002b, Braschler et al. 2002b).

3.1 Direct Combination of Translation Resources

Two or more translation resources are directly combined into one "meta-resource" that comprises the union of the information contained in the individual resources. To make a combination in this sense possible, the resources must be "compatible", i.e. they should use matching types of input and output. Secondly, the resources must be available in a format that allows combination.

We have so far experimented with the direct combination of multiple machine-readable dictionaries, of multiple similarity thesauri, and a combination of machine-readable dictionaries and similarity thesauri. Only the combined MRDs were used in our CLEF experiments.

The direct combination is attractive, because it eliminates duplicate information, and therefore duplicate translation steps, and it avoids the need for later merging of intermediate results. When directly combining dissimilar translation resources, such as similarity thesauri and MRDs, weighting of the individual parts is a problem. This weighting problem is not dissimilar from the problem incurred when using the translation resources separately, and merging their individual

outputs – in both cases the weighting should reflect the potential of the corresponding translation resources to contribute to a good overall result. We therefore avoid the direct combination of incompatible translation resources, using the merging of their individual outputs instead.

In order to allow direct combination of resources, Eurospider has developed a framework that operates on a common XML format. We use ISO Latin-1 for encoding. A good framework for direct combination is especially helpful in operational settings, where customer-specific translation resources are available and must be integrated into the system's larger, general-purpose resources.

3.2 Combination of Intermediate Results

Alternatively to building one large multilingual search index, a multilingual text retrieval system may handle languages in pairs, where the intermediate results from each language pair are later merged into one unified result. This is the approach chosen by most participants in the CLEF multilingual track.

In an analogous way, it is also possible to use the different translation resources individually, and then later merge the resulting intermediate results into one combined result that contains the results obtained by using all of the resources. The resources proper need not be combined directly in this approach, making it possible to use resources of very different types that are not compatible for direct combination.

Eurospider has combined outputs from machine translation, from similarity thesaurus translation and from MRDs in this way. The merging approaches used for such combinations are the same as for the combination of different language pairs , and are detailed in Section 4.

3.3 Architecture of a Combination Approach

The Eurospider CLIR component allows the following combinations:

- Multiple language pairs
- Document translation and query translation
- Multiple query translation approaches and resources

The systems we used for CLEF experiments work by first running each individual query translation method for each individual language pair. At this stage, the system mirrors a collection of classical bilingual CLIR systems using different translation resources. It is therefore possible to use all refinements developed for

such bilingual CLIR, an important advantage of this form of combination approach. Specifically, we use post-translation blind relevance feedback, which was shown to be effective in Ballesteros and Croft (1996). Based on our experience in all CLEF campaigns, we currently use the 10 best terms from the 10 top-ranked documents and Smart ltc weights for expansion. For the actual retrieval, we use the Smart Lnu.ltn weighting scheme, which has repeatedly been shown as effective for this type of document collections (Singhal et al. 1996).

When the intermediate results for all language pair/query translation method combinations are available, the first merging step is executed. We have experimented with two options at this stage: merging all outputs of one translation method into a multilingual result (see figure 2), or merging all outputs for one target language into a monolingual result combining the different translation approaches (see figure 3).

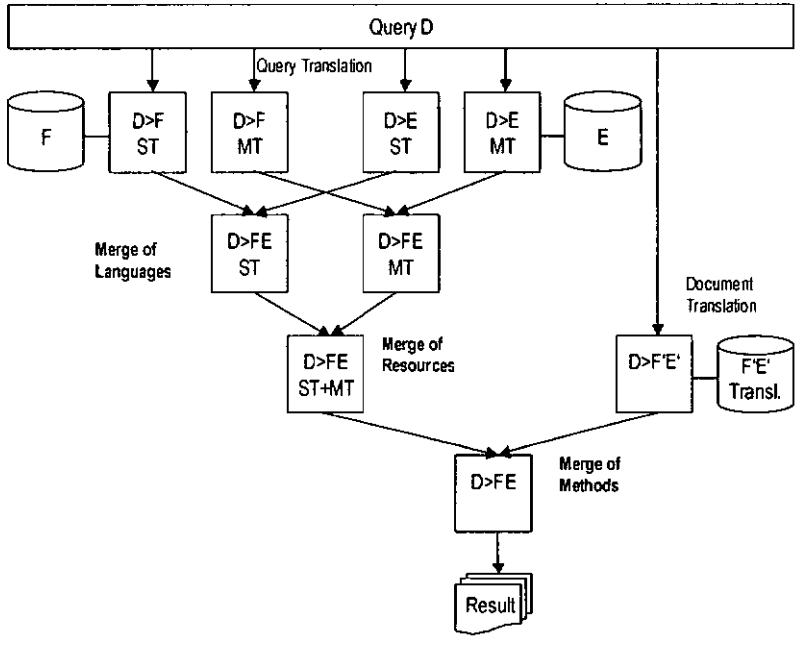

Figure 2. Combination system for CLIR, in this example for a system translating from German (D) to French (F) and English (E), combining different language pairs (merge of languages), different translation resources (merge of resources; machine translation "MT" and similarity thesauri "ST") and lastly different translation methods (merge of methods; query translation "QT" and document translation "DT")

In a second operation, the system then either merges the multilingual intermediate results into one multilingual result that covers all translation methods, or it merges the different monolingual results into a multilingual result. In both cases, the result covers all query translation methods and all language pairs.

We have found only small performance differences in our experiments with both alternatives, with first merging all languages, and then the methods in a second step being slightly beneficial for CLEF 2002.

In a final merging step, the result from query translation is then merged with a result obtained from document translation. The document translation result is produced by inserting all translated documents into the system and then doing simply monolingual retrieval on this index.

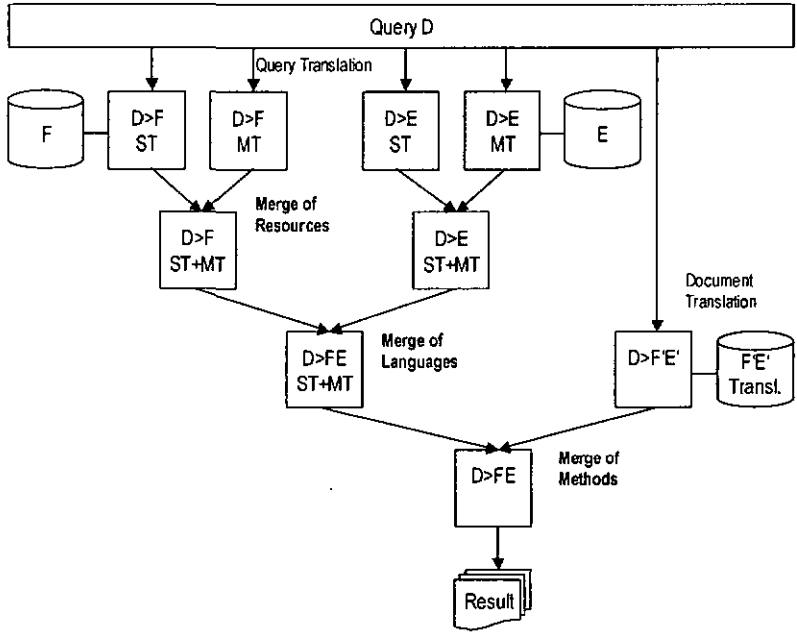

Figure 3. Combination system for CLIR, in this example for a system translating from German (D) to French (F) and English (E), combining different translation resources (merge of resources; machine translation "MT" and similarity thesauri "ST"), different language pairs (merge of languages) and lastly different translation methods (merge of methods; query translation "QT" and document translation "DT")

4 Merging

Clearly, merging is an integral part in our combination approach to CLIR. In earlier CLEF participations, we have used two simple merging strategies: rank-based merging and interleaving.

For merging, we essentially distinguish two scenarios:

1. Both retrieval results were calculated on the same search space. The two result lists will essentially be a reordering of each other (with some extra items appearing at the bottom of the lists). This is the case if the output from experiments using the same language pair and different translation resources is merged (analogous to "data fusion" in monolingual retrieval, see e.g. Belkin et al. 1993a, Fox and Shaw, 1993).
2. The sets of documents in the result lists are disjoint. This is the case if the runs were produced through retrieval on disjoint search spaces, e.g. one search on the English part of the multilingual CLEF collection, and another search on the French part (analogous to "collection fusion" in monolingual retrieval, see e.g. Callan et al. 1995, Voorhees et al. 1995).

Some merging strategies apply to both scenarios, whereas some strategies can only be used for the first scenario. Rank-based merging can only be applied if the search spaces are shared among the lists to be merged. Interleaving is a more general strategy and can be used for both scenarios.

The main difficulty in merging is the lack of comparability of scores across different result lists. Result lists obtained from different collections, or through different weightings, are commonly not directly comparable. The retrieval status value RSV that the weighting scheme attaches to every document is only used for sorting the list, and is only valid in the context of the query, weighting and collection used.

4.1 Simple Merging Strategies

The two merging strategies described below address the problem of incompatible RSVs by not using the original RSV scores at all in determining the new rank of a document in the merged result list.

Rank-based Merging

For rank-based merging, calculation of a new RSV value for the merged list is based on the ranks of the documents in the original result lists. To calculate the new

RSV of a document, its ranks in all the result lists are added. There are variations in this way of doing rank-based merging, such as the proposal by Belkin et al. (1993b) to use the median of all the ranks that a document obtained in the various result lists.

Clearly, the strategy only applies if the search space of all runs is shared, and therefore a substantial "overlap" in the documents retrieved for the individual runs exists. Since we feel that there is more importance in a rank difference between highly ranked documents than in a similar difference among lower ranked documents, we introduced a logarithmic dampening of the rank value, thus boosting the influence of highly ranked documents.

As mentioned previously (Section 3), we want to minimize the number of negative outliers. By using the dampening function, a document that receives one very good and one very poor rank will be ranked higher in the merged result list than a document receiving two mediocre ranks. Good documents are therefore not excessively penalized if one of the two retrieval runs performs very poorly. Of course, this effect has two sides: on the other hand, highly ranked documents from the poor run are also not downweighted drastically. In consequence, positive outliers are also blunted.

Interleaving

As an alternative that applies in both merging scenarios (same search space and disjoint search space), we have in the past used interleaving (also known as "round robin"): the merged result list is produced by taking one document in turn from all individual lists. If the collections are not disjoint, duplicates will have to be eliminated after merging, for example by keeping the most highly ranked instance of a document. Interleaving is a fairly old strategy, and has frequently been used as a baseline for comparison with newer schemes before (see e.g. Callan et al. 1995, Voorhees et al. 1995).

4.2 Improved Merging Strategies

We started to investigate improved merging strategies for the CLEF 2002 campaign. Two new methods were investigated. The first is a slight update of the interleaving strategy. The second is more elaborate, and presents an attempt to guess how well a query "hit" a language-specific subcollection.

Collection Size Based Interleaving

One main deficiency of interleaving as described above is that all result lists are handled equally, taking the same number of documents from each. It is extremely difficult to determine the number of relevant documents to be expected in the individual subcollections, but we have observed large collections in CLEF contribute on average more relevant documents per query than the smaller ones. Consequently, we have used a simple update to the straight interleaving method: we set the portion of documents taken from any one result list to be proportional to the size of the corresponding subcollection.

Feedback Merging

The second new strategy aims to predict the amount of relevant information contributed by each subcollection for a specific search request. It does this by carrying out an initial retrieval step, and then analyzing the top ranked documents from the result set, building an "ideal" query to retrieve that set of documents. This query is then compared to the original query, determining the overlap as an indication of the degree to which the concepts of the original query are represented in the retrieval result. The better such representation, the higher is the estimate of relevant documents (see figure 4). The result lists are then finally merged in proportion to these estimates. The biggest advantage of this method is its query dependence: whereas all the other methods described above use fixed ratios for merging the different result sets, this method determines an "optimal" ratio per query.

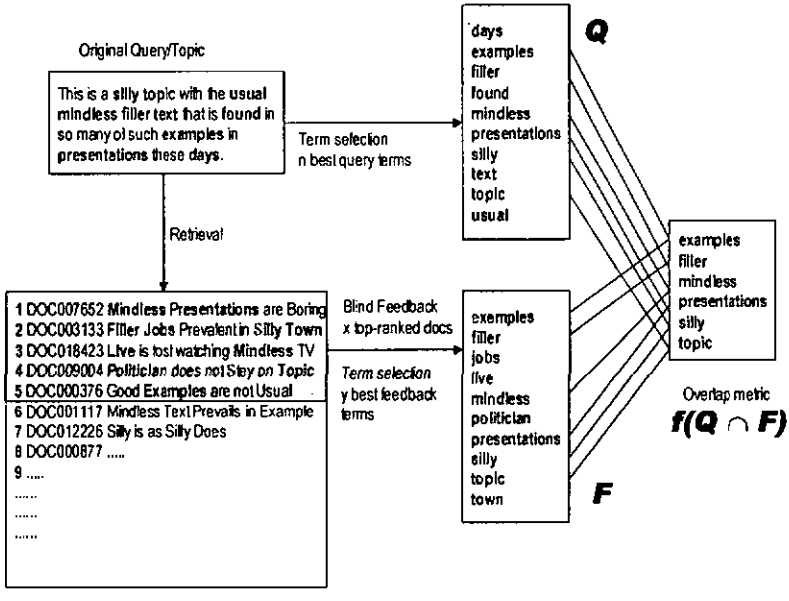

Figure 4. Simplified diagram of the feedback merging method. The query is used to obtain a list of highly ranked documents, from which representative terms are selected. High overlap of these terms with query terms indicates a high density of relevant items in the retrieval result.

5 Analysis

Eurospider has participated with combination systems in all three CLEF campaigns to date. The results in all cases compare favorably to other participants, with the systems placing among the top groups in all cases, and delivering the top performance for CLEF 2000.

Since the components are developed with integration into Eurospider's commercial retrieval products in mind, there are some design decisions that may contradict the objective of getting optimal performance on the CLEF corpus. In particular, we did not want to tune the system specifically with regard to the CLEF test collections in the following aspects:

1. Training of translation resources on the corpus itself. It was considered to be unrealistic for most operational cases that the collection itself can be used as training data.
2. Asymmetric handling of language pairs. There is a potential benefit in fine-tuning the system for every language pair. However, it is often unclear whether such tuning actually addresses peculiarities of a language pair or rather those of the underlying training collection. Again, it was considered unrealistic that extensive corpus-specific language tuning would be possible in operational settings.
3. Training on a previous year's relevance assessments. It is possible (and within the scope of CLEF permitted) to use the previous year's relevance assessments for fine-tuning of the system. However, by doing so, a situation is simulated where a lot of evaluation data is available for a corpus. We consider it to be unrealistic that corpus-specific evaluation data of characteristics similar to a set of CLEF relevance assessments would be available in most operational settings.

All these three optimizations were reported to be potentially beneficial for evaluation at the last CLEF workshops (Rogati and Yang 2002, Brand and Brünner 2002, Savoy 2002b).

By taking the former three considerations into account, it follows that all our CLEF experimental systems were handling languages symmetrically to the extent possible, translation resources were trained on similar but disjoint training data, and relevance assessments from the previous years were not taken into account. We think the good performance of our systems in CLEF is all the more remarkable under these circumstances.

Besides the objective inherent to the CLEF evaluation of getting good performance with respect to the measures calculated from the relevance assessments,

such as precision and recall, we were also interested in additional criteria which we consider essential for success in operational settings, such as:

1. Lexical coverage
2. Robustness with regard to negative outliers
3. Broad applicability of the system to a range of operational settings

By respecting these criteria, it may become necessary to add components to the combined approach that do not positively influence recall/precision measures or potentially even have a slightly negative impact.

In the next sections, we will comment on these three criteria with regard to the results observed at the three CLEF campaigns. Analysis of the individual results with respect to a single year's CLEF evaluation and their comparative performance with those of that year's participants is not the focus of this paper, and is treated in detail in previously published work (Braschler and Schäuble 2001, Braschler et al. 2002a, 2002b).

5.1 Lexical Coverage

Typically, for CLEF topics, modern MT systems work well. This is especially true for long queries, which are essentially complete sentences, allowing MT systems to make use of their facilities for word sense disambiguation. The situation is a bit trickier for short queries, but these also work reasonably well as long as the terms come either from general vocabulary, or contain names that are not mistakenly translated. In the remaining cases, a corpus-based approach that, if appropriately trained, covers domain-specific terminology and fashionable terms, can help.

We have boosted lexical coverage of our experiments by using the corpus-based similarity thesaurus in conjunction with machine translation. As training data, we used additional news wire data that did not overlap with the newspaper and news wire data in the CLEF collections. Clearly, to be able to beneficially use a similarity thesaurus for our combination system, the training data must be sufficiently comparable in terms of terminology to the target document collection. In the optimal case, we could use the target collection for training data. As we mentioned above (Section 5), we think this is normally not possible in operational settings. We consider the requirement of obtaining additional training documents from the same domain a much smaller obstacle and much easier to satisfy in operational settings. When looking at the results presented in Tables 1 and 2, a mixed picture emerges. Only in CLEF 2000 have we seen a benefit in terms of average precision when merging the output from the similarity thesaurus. However, this is mainly due to inconsistent quality of the thesauri we have used for the individual language pairs.

We have not had access to suitable training data for English and Spanish, which led to similarity thesauri that did not perform as well as desired. A detailed analysis we performed for CLEF 2000 (Braschler and Schäuble 2001) shows that additionally to providing broader lexical coverage, a combination of machine translation and similarity thesauri can also boost average precision, when good-quality thesauri are available, as was the case for German/French and German/Italian.

Table 1. Comparison of experiments using machine translation (MT), similarity thesauri (ST) and a combination of both. Shown are average precision figures. We only observed a clear benefit from combination for CLEF 2000.

	CLEF 2000	CLEF 2001	CLEF 2002
MT	0.2557	0.2929	0.3220
ST	0.1656	0.1778	0.1689
MT+ST	0.2622	0.2903	0.2876

Table 2. Direct per-query comparison of machine translation (MT), similarity thesauri (ST) and a combination of both.

	CLEF 2000	CLEF 2001	CLEF 2002
MT+ST vs. MT	23:17	19:31	18:32
MT+ST vs. ST	39:1	49:1	49:1
MT vs. ST	36:3	43:7	45:5

5.2 Robustness with respect to negative outliers

Measures such as average precision, while very valuable for objective evaluation of systems in a setting such as CLEF, emphasize a goal of tuning the system for optimal performance on an average over a number of queries. In operational settings, even single negative outliers can impact the user's perception of a system substantially. We have observed that combination approaches very effectively address the desire to avoid negative outliers. This is especially true for the combination of query translation (QT) and document translation (DT), which helps to boost queries that perform poorly when only using one method alone. Tables 3 and 4 show an analysis of the benefits gained by combining DT+QT as opposed to using either method alone. In all three campaigns, the combination experiment outperforms both DT and QT experiments. In CLEF 2002, however, we have observed that the run using only DT via MT performed practically identically when considering average precision to the more complicated run combination run. However, looking at the recall, the combined run retrieves nearly 10% more relevant items. While not superior in terms of the popular average precision measure, the combined DT+QT run seems therefore preferable in operational situations.

In both CLEF 2000 and CLEF 2001, the picture is even more favorable for the combined run, which clearly outperforms both DT and QT alone. When looking at individual queries, one can see that combination is almost always beneficial, with the vast majority of queries showing improvement. Even more important, practically all those queries that show decreased performance suffer only slightly.

As a side note, the document translation runs consistently outperformed query translation runs for all three campaigns. However, when the merging process used for fusing the outputs from the different language pairs in query translation is investigated, it becomes apparent that much of this difference stems from merging. We will return to this aspect at the end of this section.

One of the measures published by CLEF is the performance compared to a theoretical median of all participants for every individual query. Our goal is therefore to outperform at least this median for as many cases as possible. This again allows us to conclude that we avoid producing negative outliers. In CLEF 2001, we observed that the number of queries with performance below the median decreased by as much as half when comparing a system combining both QT and DT with a system employing either of the two translation options alone (see Table 5).

Table 3. Average precision results for document translation (DT), query translation (QT) and a combination of both. Combination outperforms the simpler methods in all cases.

	CLEF 2000	CLEF 2001	CLEF 2002
DT	0.2816	0.3099	0.3539
QT	0.2500	0.2773	0.2876
DT+QT	0.3107	0.3416	0.3554

Table 4. Direct per-query comparison of document translation (DT), query translation (QT) and a combination of both. Shown is the number of queries performing better in terms of average precision for the given alternatives. For example, the combined run for the 2002 campaign outperforms the respective document translation run for 28 out of 50 queries.

	CLEF 2000	CLEF 2001	CLEF 2002
DT+QT vs. DT	32:8	41:9	28:22
DT+QT vs. QT	31:9	36:14	41:9
DT vs. QT	24:16	25:25	30:20

Table 5. Comparison of the number of queries above and below the median performance of all CLEF participants in terms of average precision, shown for systems using document translation (DT), query translation (QT) and a combination of both. For example, our document translation run for the 2000 campaign outperformed the median performance for 30 out of 40 queries.

	CLEF 2000	CLEF 2001	CLEF 2002
DT vs. median	30:10	24:26	38:10
QT vs. median	23:15	22:22	39:11
DT+QT vs. med.	30:9	37:12	39:8

5.3 Broad applicability of the system to a range of operational settings

We try to ensure the applicability of the system to different settings by avoiding any undue "overtuning" to the collection. The choice of a symmetrical handling of all language pairs stems from this desire. We have observed in several cases that it would have been beneficial to chose a different approach for individual language pairs when focusing strictly on performance within the CLEF campaign. In CLEF 2000, for example, we used a dictionary-type word list to combine with machine translation for the German/English language pair. This additional resource hurt the overall result by 25% when compared to machine translation alone. Other examples include the English similarity thesaurus we used in both CLEF 2001 and CLEF 2002. This component hurt overall performance due to noisy training data, but still provided valuable improvements for some individual queries.

While we were therefore "ill-served" by the choice of including some of these resources for within-CLEF evaluation, we feel that we can conclude that the combination system was robust to such design. We experienced only slight performance degradation while still benefiting from the expanded lexical coverage that additional resources provide.

5.4 Evaluation of Merging

The process of merging intermediate results is central to our combination approach, and merits evaluation outside the scope of the criteria outlined above. As observed in our discussion of negative outliers, document translation consistently outperforms query translation in all three campaigns (Table 3). However, when the performance of the query translation run is subdivided into individual retrieval performance and subsequent merging, we can in retrospect investigate through the use of the relevance assessments provided by CLEF the performance loss incurred by imperfect merging. The knowledge of which items are relevant to a query allows us to produce a "perfect" merging result by taking the optimal ratio of items from all intermediate results. By doing this, we can see that there remains a lot of potential in developing better merging strategies.

Table 6. Performance of actual and theoretical merging strategies.

Merging strategy	CLEF 2002 MT query translation
Interleaving	0.3249
Collection size-based interleaving	0.3369
Feedback merging	0.3220
Relevance ratio	0.4067
Optimal	0.4876

An optimally merged query translation run (average precision of 0.4876, see Table 6) clearly outperforms document translation (0.3539, see Table 3). However, all our merging strategies miss the optimal performance by a wide margin. Thus, the avoidance of the merging problem by the document translation run turns out to be a real advantage.

When looking at the individual methods, we see that taking collection sizes into account when interleaving gives a small benefit compared to straight-forward interleaving. This result is consistent for the various query lengths we have tried.

The merging strategy based on feedback shows little improvement compared to the simpler interleaving methods, even though it allows different merging ratios for different queries. This is disappointing, since the use of variable merging ratios makes it possible to lift one of the main restrictions of the simpler merging methods. When looking at the ratios which were actually used for merging, we see that this method does indeed chose fairly different ratios for individual queries. However, we believe that these differences were obscured by the relatively large number of relevant items in the CLEF collections, and by problems with the estimation of the ratios from the feedback terms. While we saw encouraging improvements for some queries, feedback merging therefore did not bring any sustained improvement in retrieval effectiveness, even being slightly disadvantageous in several cases. We are working on improvements to this method.

The last two methods listed in Table 6 give theoretical upper bounds for merging performance. The "relevance ratio" experiment was obtained by determining the correct merging ratio per query based on the relevance assessments. This run shows the theoretical optimum which could be obtained by a method such as the feedback merging strategy. An even better performance is obtainable if different merging ratios are adopted not only per query, but also for different levels of recall within a query (the necessary merging strategy was demonstrated by Chen (2002b) during a presentation at the CLEF 2002 workshop). This upper bound is listed under the label "optimal" in Table 6. The huge performance difference between the actual merging strategies available and this optimum validates our belief that merging remains one of the big challenges in the design of better combination systems.

6 Conclusions

Eurospider has participated in all three CLEF evaluation campaigns to date. Our main objective was to test our CLIR component developed on the basis of a combination approach to CLIR. Our focus was on keeping the experiments applicable to operational settings. The goals of the combination approach in this respect were twofold: improved retrieval effectiveness, and improved system robustness with regard to negative outliers, lexical coverage and applicability to a broad range of domains. However, the second goal has not been explored extensively inside the CLEF campaign.

We have demonstrated how our component can combine a wide range of different aspects: different language pairs, different types of translation resources, and different levels of translation (document translation vs. query translation). While problems remain with merging, we have found definite benefits from a combination of multiple simpler approaches.

The combination of document translation and query translation improves performance for all three campaigns, in two cases substantially. The combined runs also perform better when compared query-by-query to the individual runs or the median CLEF performance.

Combination of multiple translation resources improves lexical coverage, and in some cases also retrieval performance. It also helps to avoid negative outliers, which is very important with respect to operational settings.

We have observed in all three campaigns that our experiments using only document translation outperform those using query translation alone. However, when investigating more closely, we find that this behavior is mainly due to a big drop in performance attributable to imperfect merging. Since merging is a central aspect of a combination system such as ours, with intermediate results merged across languages, translation resources and translation methods, this indicates that more work is needed on the merging problem, and that the combination approach has a lot of potential remaining for future improvements.

Acknowledgments

Peter Schäuble, Bärbel Ripplinger and Anne Göhring were involved in planning and conducting the Eurospider participation in the three CLEF campaigns. The paper has benefited from the work of three anonymous reviewers, who provided detailed and helpful comments. Their input is greatly appreciated.

References

Ballesteros L and Croft B (1996) Dictionary Methods for Cross-Lingual Information Retrieval. In: Proceedings of the 7th International DEXA Conference on Database and Expert Systems, September 9-13. Zurich, Switzerland. Springer. pp. 791-801

Bartell BT, Cottrell GW, Belew RK (1994) Automatic Combination of Multiple Ranked Retrieval Systems. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 173-181.

Belkin NJ, Cool C, Croft WB, Callan JP (1993a) The Effect of Multiple Query Representations on Information Retrieval System Performance. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 339-346

Belkin NJ, Kantor P, Cool C, Quatrain R (1993b) Combining Evidence for Information Retrieval. In: The Second Text REtrieval Conference (TREC-2), NIST Special Publication 500-215, pp. 35-43.

Brand R and Brünner M (2002) Océ at CLEF 2002. In: Working Notes for the CLEF 2002 Workshop. pp. 21-30

Braschler M, Krause J, Peters C, Schäuble P (1998a) Cross-Language Information Retrieval (CLIR) Track Overview. In: The Seventh Text REtrieval Conference (TREC-7), NIST Special Publication 500-242. pp. 1-8.

Braschler M, Mateev B, Mittendorf E, Schäuble P, Wechsler M (1998b) SPIDER Retrieval System at TREC7. In: The Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242. pp. 446-454.

Braschler M and Schäuble P (1998) Multilingual Information Retrieval Based on Document Alignment Techniques. In: Research and Advanced Technology for Digital Libraries, Second European Conference. ECDL '98, Lecture Notes in Computer Science, Vol. 1513, Springer, pp. 183-197.

Braschler M and Schäuble P (2001) Experiments with the Eurospider Retrieval System for CLEF 2000, CLEF 2000. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000. Lecture Notes in Computer Science. Vol. 2069. Springer, pp. 140-148.

Braschler M, Ripplinger B, Schäuble P (2002a) Experiments with the Eurospider Retrieval System for CLEF 2001. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001. Lecture Notes in Computer Science, Vol. 2406, Springer, 2002, pp. 102-110.

Braschler M, Göhring A, Schäuble P (2002b) Eurospider at CLEF 2002. In: Working Notes for the CLEF 2002 Workshop. pp. 127-132.

Callan JP, Lu Z, Croft WB (1995) Searching Distributed Collections with Inference Networks. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 21-28.

Chen A (2002a) Multilingual Information Retrieval Using English and Chinese Queries. In: Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 44-58.

Chen A (2002b) Cross-Language Retrieval Experiments at CLEF 2002. In: Working Notes for the CLEF 2002 Workshop, pp. 5-20.

Du A and Callan J (1998) Probing a Collection to Discover Its Language Model. Technical Report 98-29, Department of Computer Science, University of Massachusetts.

Figuerola CG, Berrocal JLA, Zazo AF, Díaz RG (2001) A Simple Approach to the Spanish-English Bilingual Retrieval Task. In: Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069, Springer, pp. 224-229.

Fox EA, Koushik MP, Shaw J, Modlin R, Rao D (1992) Combining Evidence from Multiple Searches. In: The First Text REtrieval Conference (TREC-1), NIST Special Publication 500-207, pp. 319-328.

Fox EA, Shaw JA (1993) Combination of Multiple Searches. In: The Second Text REtrieval Conference (TREC-2), NIST Special Publication 500-215, pp. 243-252

Gey F, Jiang H, Petras V, Chen A (2001) Cross-Language Retrieval for the CLEF Collections – Comparing Multiple Methods of Retrieval. In: Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069, Springer, pp. 116-128.

Gey FC, Jiang H, Perelman N (2002) Working with Russian Queries for the GIRT, Bilingual and Multilingual CLEF Tasks, In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum. CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, 2002, pp. 235-243.

Grefenstette G (1998) Cross-Language Information Retrieval, Kluwer Academic Publishers.

Hedlund T, Keskustalo H, Pirkola A, Sepponen M, Järvelin K (2001) Bilingual Tests with Swedish, Finnish, and German Queries: Dealing with Morphology, Compound Words, and Query Structure. In: Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069, Springer, pp. 210-223.

Hiemstra D, Kraaij W, Pohlmann R (2001) Translation Resources, Merging Strategies, and Relevance Feedback for Cross-Language Information Retrieval. In: Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069, Springer, pp. 102-115.

Jones GJF, Lam-Adesina AM (2002) Exeter at CLEF 2001: Experiments with Machine Translation for Bilingual Retrieval. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 59-77.

Lee JH (1995) Combining Multiple Evidence from Different Properties of Weighting Schemes. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 180-188.

McNamee P and Mayfield J (2002a) JHU/APL Experiments at CLEF: Translation Resources and Score Normalization. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 193-208.

McNamee P and Mayfield J (2002b) Scalable Multilingual Information Access. In: Working Notes for the CLEF 2002 Workshop, pp. 133-140.

Moulinier I and Molina-Salgado H (2002) Thomson Legal and Regulatory Experiments for CLEF 2002. In: Working Notes for the CLEF 2002 Workshop, pp. 91-96.

Nie JY, Simard M (2002) Using Statistical Translation Models for Bilingual IR. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 137-150.

Oard DW (1997) Alternative approaches for crosslanguage text retrieval. In: AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence.

Qiu Y and Frei HP (1993) Concept Based Query Expansion. In: Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 160-169.

Rogati M and Yang Y (2002) Cross-Lingual Pseudo-Relevance Feedback Using a Comparable Corpus. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum. CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 151-157.

Salton G (1970) Automatic Processing of Foreign Language Documents. Journal of the American Society for Information Science, 21:187-194.

Saracevic, T., Kantor, P. (1988) A study of information seeking and retrieving. III. Searchers, searches and overlap. Journal of the American Society for Information Science, 39:197-216.

Savoy J (2002a) Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 27-43.

Savoy J (2002b) Report on CLEF-2002 Experiments: Combining Multiple Sources of Evidence. In: Working Notes for the CLEF 2002 Workshop, pp. 31-46.

Sheridan P, Braschler M, Schäuble P (1997) Cross-language information retrieval in a multilingual legal domain. In: Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, pp. 253 - 268.

Singhal, A, Buckley, C, Mitra, M (1996) Pivoted Document Length Normalization. In: Proeeedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 21-29.

Voorhees EM. Gupta NK, Johnson-Laird B (1995) Learning Collection Fusion Strategies. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Researeh and Development in Information Retrieval, pp. 172-179.

Original published in C.A. Peters (Ed.): CLEF 2002, LNCS 2785, pp. 164-174, 2003.

Eurospider at CLEF 2002

Martin Braschler*, Anne Göhring, Peter Schäuble

Eurospider Information Technology AG
Schaffhauserstrasse 18, CH-8006 Zürich, Switzerland
{martin.braschler|peter.schauble}@eurospider.com

Abstract. For the CLEF 2002 campaign, Eurospider participated in the multilingual and German monolingual tasks. Our main focus was on trying new merging strategies for our multilingual experiments. In this paper, we describe the setup of our system, the characteristics of our experiments, and give an analysis of the results.

1 Introduction

In the following, we describe our experiments carried out for CLEF 2002. Much of the work for this year's campaign builds again on the foundation laid in the previous two CLEF campaigns [2] [3]. Eurospider participated both in the main multilingual track and in the German monolingual track. The main focus of the work was on the multilingual experiments, continuing our successful practice of combining multiple approaches to cross-language retrieval into one system. Using a combination approach makes our system more robust against deficiencies of any single CLIR approach. The main effort for this year was spent on the problem of merging multiple result lists, such as obtained either from searches on different subcollections (representing the different languages of the multilingual collection) or from searches using different translation resources and retrieval methods. We feel that merging remains an unsolved problem after the first two CLEF campaigns. We introduced a new merging method based on term selection after retrieval, and we also implemented a slightly updated version of our simple interleaving merging strategy.

The remainder of this paper is structured as follows: we first discuss the system setup that we used for all experiments. This is followed by a closer look at the multilingual experiments, including details of this year's new merging strategies. The next section describes our submissions for the German monolingual track. The paper closes with conclusions and an outlook.

---

* also affiliated with Université de Neuchâtel, Institut Interfacultaire d'Informatique, Pierre-à-Mazel 7, CH-2001 Neuchâtel, Switzerland.

2 System Setup

2.1 Retrieval System

All experiments used a system consisting of a version of the core software that is included in the "relevancy" system developed by Eurospider Information Technology AG. This core software is used in all commercial products of Eurospider. Prototypical enhancements included in the CLEF system are mainly in the components for multilingual information access (MLIA).

As an additional experiment, we collaborated this year with Université de Neuchâtel in order to produce a special multilingual run. The results for this run are based both on output from the Eurospider system and on output produced by the Neuchâtel group. More details will be given later in this paper.

Indexing: Table 1 gives an overview of indexing methods used for the respective languages:

Table 1. Overview of indexing setup for all languages

Language	Stopword removal	Stemming	Convert diacritical chars
English	yes	English "Porter-like" stemmer	-
French	yes	French Spider stemmer	no
German	yes	German Spider stemmer, incl. decompounding	yes
Italian	yes	Italian Spider stemmer	no
Spanish	yes	Spanish Spider stemmer	yes

Ranking: The system was configured to use straight Lnu.ltn weighting, as described in [16]. An exception was made for some of the monolingual runs, which either used BM25 weighting [12] or a mix of both Lnu.ltn and BM25. All runs used a blind feedback loop, expanding the query by the top 10 terms from the 10 best ranked documents.

Individual Language Handling: We decided to make indexing and weighting of all languages as symmetric as possible. We did not use different weighting schemes for different languages, or different policies with regard to query expansion. By choosing this approach, we wanted to both mirror a realistic, scalable approach as well as avoid overtraining to characteristics of the CLEF multilingual collection.

2.2 Submitted Runs:

Table 2 summarizes the main characteristics of our official experiments:

Table 2. Main characteristics of official experiments.

Run tag	Track Topic lang Topic fields Run type	Translation	Merging strategy	Expansion	Weighting
EIT2MNU1	Multilingual DE TDN Auto	Documents (MT)	-	Blind Feedback 10/10	Lnu.ltn
EIT2MNF3	Multilingual DE TDN Auto	Queries (ST+MT), Documents (MT)	Term selection	Blind Feedback 10/10	Lnu.ltn
EIT2MDF3	Multilingual DE TD Auto	Queries (ST+MT), Documents (MT)	Term selection	Blind Feedback 10/10	Lnu.ltn
EIT2MDC3	Multilingual DE TD Auto	Queries (ST+MT), Documents (MT)	Interleave collection	Blind Feedback 10/10	Lnu.ltn
EAN2MDF4	Multilingual EN TD Auto	Queries (MT+Uni Neuchâtel)	Logrank	Blind Feedback 10/10+ Uni Neuchâtel	Lnu.ltn + Uni NE
EIT2GDB1	Ger. Monol.. DE TD Auto	-	-	Blind Feedback 10/10	BM25
EIT2GDL1	Ger. Monol. DE TD Auto	-	-	Blind Feedback 10/10	Lnu.ltn
EITGDM1	Ger. Monol. DE TD Auto	-	Logrank	Blind Feedback 10/10	Lnu.ltn + BM25
EITGNM1	Ger. Monol. DE TDN Auto	-	Logrank	Blind Feedback 10/10	Lnu.ltn + BM25

3 Multilingual Retrieval

3.1 Approach

As for last year, we again spent our main effort on the multilingual track. The goal of this track in CLEF is to select a topic language, and use the queries in that language to retrieve documents in five different languages from a multilingual collection. A single result list has to be returned, potentially containing documents in all languages. For details on the track definition see [1].

A system working on such a task needs to bridge the gap between the language of the search requests and the languages used in the multilingual document collection. For translation, we have successfully used combination approaches, integrating more than one translation method, in the past two campaigns. In 2001, we attempted our most ambitious combination yet [2], combining three forms of query translation (similarity thesauri [11] [15], machine translation and machine-readable dictionaries) with document translation (through machine translation). This year, we shifted the focus of our experiments somewhat, and used a slightly simpler combination approach, concentrating on the pieces that performed well in 2001 and discarding the machine-readable dictionaries from the system. The experiments used either a combination of document translation (DT) and two forms of query translation (QT) - machine translation (MT) and similarity thesaurus (ST) - or document translation only.

More than last year, we concentrated on the problem of producing the final, multilingual result list to be returned by the system in answer to a search request. Our combination approach, as indeed most successful approaches to multilingual retrieval used in the CLEF 2001 campaign, produces several intermediate search results, either due to using different translation methods, or due to handling only a subset of the five languages at a time. These intermediate search results need then be "merged" to produce the multilingual result list necessary for submission to CLEF.

It seems to be generally agreed among the community of active participants in CLEF that merging is an important problem in designing truly multilingual retrieval systems. It was pointed out in the discussions at the CLEF 2001 workshop that not much progress had been made on this topic up to that campaign. This year, several groups have put emphasis either on this issue (e.g. [4], [6], [7], [9], [14]) or alternatively on how to avoid merging altogether (e.g. [8], [10]). The simple merging approaches employed in our experiments are very similar to some of the work described by these groups.

3.2 Merging

In our previous participations in CLEF, we have used two simple merging strategies: rank-based merging, and interleaving. There are two fundamentally different merging scenarios, depending on whether the runs to be merged share the same search space or not. In the case that we merge runs obtained from different translation methods, translation resources or weighting schemes, but using the same target language, there will usually be a substantial overlap in the sets of documents in the respective result lists (this problem is also referred to as "data fusion"). This is not the case when we merge runs that use disjoint search spaces (sometimes referred to as "database merging"), e.g. different target languages. Some merging strategies apply to both these scenarios (e.g. interleaving), while others only operate on one of the two alternatives (e.g. rank-based merging).

The main difficulty in merging is the lack of comparability of scores across different result lists. Result lists obtained from different collections, or through different weightings, are normally not directly comparable. The retrieval status value RSV is only meaningful for sorting the list, and is often only valid in the context of the query, weighting and collection used.

Some merging strategies make use of training data, typically based on relevance assessments for some prior queries. We have decided to avoid reliance on availability of relevance assessments, since we envision many scenarios where it is unfeasible to tune the system to such an extent. We have however used the relevance assessments of this year's collection to evaluate our results against "optimal" merging strategies.

3.3 Simple Merging Strategies

One way of avoiding the problem of non-comparable RSVs is by using a merging strategy that does not make direct use of retrieval scores. The following two simple merging strategies use the position of documents in the result lists to determine a new, merged score.

3.3.1 Rank-based Merging

For rank-based merging, calculation of a new RSV value for the merged list is based on the ranks of the documents in the original result lists. To calculate the new RSV, the document's ranks in all the result lists are added up. Clearly, the strategy is limited to situations with a shared search space, since a substantial "overlap" in the documents retrieved is necessary to obtain meaningful new scores. We feel that a rank difference between highly ranked documents should be emphasized more than a similar difference among lower ranked documents. As a consequence, we introduced a logarithmic dampening of the rank value that boosts the influence of highly ranked documents. The underlying assumption of this strategy is based on observations made a while back, e.g. by Saracevic and Kantor [13], who report that

the probability of relevance of a document increases if that document is found by multiple query representations.

3.3.2 Interleaving

As an alternative that applies for merging scenarios with both a shared search space or disjoint search spaces, we have in the past used interleaving. The merged result list is produced by taking one document in turn from all individual lists. If the collections are not disjoint, duplicates will have to be eliminated after merging, for example by keeping the most highly ranked instance of a document. This very simple strategy, and variants of it, also go by the name of "round-robin" or "uniform merging" (see e.g. [17]).

3.4 New Merging Strategies

We introduced two new merging strategies for this year's experiments. The first one is a slight update of the interleaving strategy. The second one is more elaborate, and presents an attempt at guessing how well the concept of the query is represented in a specific subcollection.

3.4.1 Collection Size-Based Interleaving

One main limitation of interleaving as described above is that all result lists are given equal weight, taking the same number of documents from each. There are many situations, especially in case of disjoint search spaces, where it is unsatisfactory to assume that all result lists will contribute an equal number of relevant items. Clearly, it is extremely difficult to determine the amount of contribution to be expected in the individual subcollections covered by the runs. We have however observed that the average ratio of relevant items is quite stable across the different languages in CLEF when averaged over all queries for a given year. Consequently, we have used a simple update to the straight interleaving method, making use of the fact that the subcollections of the CLEF multilingual collection vary considerably in size, and setting the portion of documents taken from any one result list to be proportional to the size of the corresponding subcollection.

3.4.2 Feedback Merging

The second new strategy aims to predict the amount of relevant information contributed by each subcollection specifically for a given search request. The amount of relevant information contributed by a subcollection can vary widely from query to query. Lifting the limit of using the same merging ratios for every query should help to avoid picking documents from poor retrieval runs too early. In order to achieve this objective, results from an initial retrieval step for every subcollection are collected, and the respective top-ranked documents are then "analyzed" by the system, selecting the best terms to build an "ideal" query to retrieve that set of documents. The system then compares this query to the original query, determining the overlap as an estimate of the degree to which the concepts of the original query

are represented in the retrieval result. The underlying assumption is that the better such representation, the higher the likelihood that relevant documents are found. The result lists are then finally merged in proportion to these estimates. No manual intervention with respect to any of the intermediate results takes place. This is the only method we applied that is query-dependent.

3.5 Optimal Retrospective Merging Strategies

We evaluated the effectiveness of our merging strategies by comparing their performance to "baselines" obtained from merging strategies that make retrospective use of the relevance assessments that became available after the evaluation campaign.

3.5.1 Relevance-Ratio Merging

The feedback merging strategy attempts to estimate for a given query how many relevant items are contributed by each of the subcollections. This information can be retrospectively extracted from the relevance assessments, and the merging can be repeated analogous to the interleaving strategy, but with the exact ratios in place of the estimates. While this strategy is query-dependent, it merges using equal ratios for all recall levels.

3.5.2 Optimal Merging

It is possible to improve even on the relevance-ratio merging strategy by lifting the limitation that merging ratios are equal for all recall levels. The idea for optimal merging was presented by Chen at the CLEF 2002 workshop [5]. The strategy works by taking in turn items from those result lists that provide the most relevant items while keeping the number of irrelevant items taken at a minimum. I.e., in every step, the strategy looks in the remainders of the lists to be merged for the shortest uninterrupted sequences of zero or more irrelevant items followed by the longest uninterrupted sequences of relevant items. This sequence is then attached to the merged list, and the corresponding result list is shortened. The resulting final list has different merging ratios both for each query and each recall level.

3.6 Results

The results for our officially submitted multilingual experiments are shown below.

Table 3. Key performance figures for the multilingual experiments

Run tag	Average Precision	Precision @ 10 docs	Relevant retrieved
EIT2MNU1	0.3539	0.6560	5188
EIT2MNF3	0.3554	0.6520	5620
EIT2MDC3	0.3400	0.5860	5368
EIT2MDF3	0.3409	0.6040	5411
EAN2MDF4	0.3480	0.6260	5698

Total number of relevant documents: 8068
"Virtual best performance": 0.4834
"Virtual median performance": 0.2193

Table 4. Comparison of performance against median for multilingual experiments

Run Tag	Best	Above Median	Median	Below Median	Worst
EIT2MNU1	4	34	2	9	1
EIT2MNF3	5	34	3	8	0
EIT2MDC3	1	36	3	10	0
EIT2MDF3	2	37	1	10	0
EAN2MDF4	1	45	0	4	0

As can be seen from Table 3, there is very little difference between runs EIT2MDC3 and EIT2MDF3, which differ only in the merging strategy employed for the combination of the individual language pairs. Additionally to having essentially equivalent average precision values, the two runs also differ only very slightly when compared on a query-by-query basis. Clearly, the merging strategy based on feedback, which we newly introduced for this year, had little impact, even though it allows different merging ratios for different queries. When looking at the ratios which were actually used for merging, we see that the new method indeed chooses fairly different ratios for individual queries. However, we believe that these differences were obscured by later merging steps that combined the various query translation and document translation methods. These further combinations blunted the effect of negative, but unfortunately also of positive "outliers". When concentrating the analysis on merging of a single, specific translation method, larger differences between the simple merging strategies and the feedback method become apparent. Even so, feedback merging did not bring any sustained improvement in retrieval effectiveness, instead even being slightly disadvantageous in several cases. The fact that feedback merging allows different merging ratios for individual queries seems to have resulted in a slight improvement in recall, however.

The run based exclusively on document translation (EIT2MNU1) compares favorably when compared to the equivalent combined DT and QT run (EIT2MNF3). This is analogous to what we observed in earlier campaigns, where DT-based runs outperformed QT-based runs. Indeed, a run based solely on query translation (not

officially submitted) performs considerably worse than either the DT-only or the combined run (0.2876 vs. 0.3539 and 0.3554 for TDN queries). However, we think this is probably more due to the imperfect merging strategies we used than due to any superiority of DT over QT. When using the retrospective merging strategies described earlier in the paper it becomes apparent that our best merging results are approximately 20% below what is obtained with the "relevance ratio" method and a sobering 45%-55% below what an optimal merging strategy making full use of the relevance assessments would produce. Since our DT run needs no merging, it is not affected by the substantial loss of performance incurred in this step. There remains clearly much to do with regard to the merging problem.

Even though EIT2MNU1 and EIT2MNF3 show little variation in performance as measured by average precision, closer analysis shows some striking differences. EIT2MNF3, which is based on a combination of query translation and document translation, outperforms EIT2MNU1 in terms of average precision by more than 20% for 14 queries, whereas the reverse is only true for 4 of the queries. Also, EIT2MNF3 retrieves roughly 8% more relevant documents than run EIT2MNU1. We believe that this shows that the combination approach boosts reliability of the system by retrieving extra items, but that we have not found the ideal combination strategy this year that would have maximized this potential.

The potential to retrieve more relevant items by using combination approaches is also demonstrated by our final multilingual entry, EAN2MDF4, which was produced by merging output from the Eurospider system with results kindly provided to us by Université de Neuchâtel (J. Savoy). This run gave the best performance of all our multilingual experiments.

Compared to median performance, all five multilingual runs perform very well. All runs have around 80% of the queries performing on or above the median, with the Eurospider/Neuchâtel combined run outperforming the median in more than 90% of cases (Table 4). The "virtual median performance", obtained by an artificial run that mirrors the median performance among all submissions for every query, is outperformed by more than 50% for all experiments. The best run obtains slightly below 75% of the "virtual best performance", which is obtained by an artificial run combining the best entries for every query.

In absolute terms, our results are among the best reported for the multilingual track in this year's campaign, within 8-10% of the top result submitted by Université de Neuchâtel. Given the sample size of 50 queries, this difference is too small to be statistically significant.

4 Monolingual Retrieval

For monolingual retrieval, we restricted ourselves to the German document collection. We fine-tuned our submissions compared to last year, and experimented with two different weighting schemes.

In contrast to last year, we used blind feedback, which was found to be beneficial in CLEF 2001 by several groups. Our German runs used the Spider German stemmer which includes the splitting of German compound nouns (decompounding). We chose the most aggressive splitting method available in the system, in order to split a maximum number of compound nouns.

Table 5. Key performance figures for the monolingual experiments

Run tag	Average precision	Precision @ 10	Relevant retrieved
EIT2GDB1	0.4482	0.5160	1692
EIT2GDL1	0.4561	0.5420	1704
EIT2GDM1	0.4577	0.5320	1708
EIT2GNM1	0.5148	0.5940	1843

Total number of relevant documents: 1938
Virtual best performance: 0.6587
Virtual median performance: 0.4244

Table 6. Comparison of performance against median for monolingual experiments

Run Tag	Best	Above Median	Median	Below Median	Worst
EIT2GDB1	1	28	2	19	0
EIT2GDL1	3	27	5	15	0
EIT2GDM1	1	30	5	14	0
EIT2GNM1	6	32	4	8	0

The results show very little difference between EIT2GDB1 and EIT2GDL1, which used the BM25 and Lnu.ltn weighting scheme, respectively (Table 5). Query-by-query analysis confirms little impact from choosing between the two alternatives. Not surprisingly, merging the two runs (into EIT2GDM1) leads again to very similar performance.

On an absolute basis, the runs all perform well, with all runs having around 60%-75% of queries with performance above the median (Table 6). All runs also outperform the "virtual median performance". For German monolingual, this seems to be a harder benchmark than for multilingual, since "virtual median performance" is around 65% of "virtual best performance", compared to roughly 45% for multilingual.

5 Conclusions and Outlook

Our main focus this year was on experiments on the problem of merging multiple result lists coming from different translation resources, weighting schemes, and the different language-specific subcollections in the multilingual task. We introduced a new method based on feedback merging, which however showed little impact in practice. Post-hoc experiments with merging strategies that make use of the relevance assessments show that our merging methods still perform far below what is theoretically possible. We will try to use our new experiences to improve on our merging strategies for next year.

As last year, we again used a combination translation approach for the multilingual experiments. The results confirm last year's good performance. We can again conclude that combination approaches are robust; 80-90% of all queries performed on or above median performance in our systems. This good performance was achieved even though we used no special configuration for individual languages, in order to more accurately reflect situations where only few details about the collections to be searched are known in advance, and to avoid potential overtuning to characteristics of the CLEF collection.

For the monolingual experiments, we compared the impact of choosing between two of the most popular high-performance weighting schemes: BM25 and Lnu.ltn. We could not detect a meaningful difference, either in overall performance or when comparing individual queries. Consequently, combining the two weightings gives no clear advantage over using either one individually.

Acknowledgements

We thank Jacques Savoy from Université de Neuchâtel for providing us with his runs that we used for merging in the experiment labeled "EAN2MDF4".

References

1. Braschler, M. & Peters, C.: CLEF2002: Methodology and Metrics. This volume.
2. Braschler, M., Ripplinger, B., Schäuble, P.: Experiments with the Eurospider Retrieval System for CLEF 2001. In: Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, Lecture Notes in Computer Science 2406, Springer 2002, pages 102-110.
3. Braschler, M., Schäuble, P.: Experiments with the Eurospider Retrieval System for CLEF 2000. In: Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, Lecture Notes in Computer Science 2069, Springer 2001, pages 140-148.
4. Chen, A.: Cross-Language Retrieval Experiments at CLEF-2002. This volume.
5. Chen, A.: Cross-Language Retrieval Experiments at CLEF-2002. In: Results of the CLEF 2002 Cross-Language Systems Evaluation Campaign. Working Notes for the CLEF 2002 Workshop, pages5-20, 2002.
6. Lin, W.-C., Chen, H.-H.: Merging Mechanisms in Multilingual Information Retrieval. This volume.
7. Martínez, F., Ureña, L. A., Martín, M. T.: SINAI at CLEF 2002: Experiments with Merging Strategies. This volume.
8. McNamee, P., Mayfield, J.: Scalable Multilingual Information Access. In: Results of the CLEF 2002 Cross-Language Systems Evaluation Campaign, Working Notes for the CLEF 2002 Workshop, pages 133-140, 2002.
9. Moulinier, I., Molina-Salgado, H.: Thomson Legal and Regulatory Experiments for CLEF 2002. This volume.
10. Nie, J.-Y., Jin, F.: Merging Different Languages in a Single Document Collection. In: Results of the CLEF 2002 Cross-Language Systems Evaluation Campaign, Working Notes for the CLEF 2002 Workshop, pages 59-62, 2002.
11. Qiu, Y., Frei, H.: Concept Based Query Expansion. In: Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA, pages 160 - 169, 1993.
12. Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M.: Okapi at TREC-3. In: Proceedings of the Third Text REtrieval Conference (TREC.3), NIST Special Publication 500-226, pages 109-126, 1994.
13. Saracevic, T., Kantor, P.: A study of information seeking and retrieving. III. Searchers, searches, overlap. Journal of the American Society for Information Science, 39:3, pages 197-216.
14. Savoy, J.: Combining Multiple Sources of Evidence. This volume.
15. Sheridan, P., Braschler, M., Schäuble, P.: Cross-language information retrieval in a multilingual legal domain. In: Proceedings of the First European Conference

on Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science 1324, Springer 1997. pages 253 - 268.

16. Singhal, A., Buckley, C., Mitra, M.: Pivoted Document Length Normalization. In: Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, pages 21-29, 1996.
17. Voorhees, E. M., Gupta, N. K., Johnson-Laird, B.: Learning Collection Fusion Strategies. In: Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval, pages 172-179, 1995.