Cross-Language Text Retrieval Research in the USA

Douglas W. Oard
College of Library and Information Services
University of Maryland, College Park, MD 20742
oard@glue.umd.edu

Abstract

The increasing availability of networked access to multilingual text collections has generated increased interest in the development of effective and efficient cross-language text retrieval technology. Examples of cross-language text retrieval applications are discussed and a classification of known approaches is introduced. This is used to structure a comprehensive discussion of published research and known commercial practice in the United States on the topic. The paper concludes by describing the structure of sponsored research on cross-language text retrieval in the United States and some brief observations of the potential for collaboration with European researchers on aspects of the problem which are of mutual interest.

1 Introduction

The explosive growth of the Internet and other sources of networked information have made automatic mediation of access to networked information sources an increasingly important problem. Much of this information is expressed as electronic text, and it is becoming practical to automatically convert some printed documents and recorded speech to electronic text as well. Thus, automated systems capable of detecting useful documents are finding widespread application.

With even a small number of languages it can be inconvenient to issue the same query repeatedly in every language, so users who are able to read more than one language will likely prefer a multilingual text retrieval system over a collection of monolingual systems. And since reading ability in a language does not always imply fluent writing ability in that language, such users will likely find cross-language text retrieval particularly useful for languages in which they are less confident of their ability to express their information needs effectively.

The use of such systems can be also be beneficial if the user is able to read only a single language. This is a particularly important consideration in the United States, where monolingual users are quite common. For example, when only a small portion of the document collection will ever be examined by the user, performing retrieval before translation can be significantly more economical than performing translation before retrieval. So when the application is sufficiently important to justify the time and effort required for translation, those costs can be minimized if an effective cross-language text retrieval system is available. Even when translation is not available, there are circumstances in which cross-language text retrieval could be useful to a monolingual user. For example, a researcher might find a paper published in an unfamiliar language useful if that paper contains references to works by the same author that are in the researcher's native language.

Multilingual text retrieval can be defined as selection of useful documents from collections that may contain several languages (English, French, Chinese, etc.). This formulation allows for the possibility that individual documents might contain more than one language, a common occurrence in some applications. Both cross-language and within-language retrieval are included in this formulation, but it is the cross-language aspect of the problem which distinguishes multilingual text retrieval from its well studied monolingual counterpart. At the SIGIR 96 workshop on "Cross-Linguistic Information Retrieval" the participants discussed the proliferation of terminology being used to describe the field and settled on "Cross-Language" as the best single description of the salient aspect of the problem. "Multilingual" was felt to be too broad, since that term has also been used to describe systems able to perform within-language retrieval in more than one language but that lack any cross-language capability. "Cross-lingual" and "cross-linguistic" were felt to be equally good descriptions of the field, but

"cross-language" was selected as the preferred term in the interest of standardization. Unfortunately, at about the same time the U.S. Defense Advanced Research Projects Agency (DARPA) introduced "translingual" as their preferred term, so we are still some distance from reaching consensus on this matter.

A couple of preliminary remarks are in order to establish the scope of this survey. My goal is to review for you the present state of cross-language text retrieval research in the United States of America (USA). Although that focus has led me to mention only in passing a good deal of timely and exciting work from Europe and the Pacific Rim, we are fortunate to have excellent speakers from both regions scheduled for this workshop. While it is not possible to describe the technical details of each approach in this brief paper, I have endeavored to cite at least one widely available reference for each system and I have included URL's for preprints of those papers where available.[1] If this paper helps DELOS members identify research groups in the USA that are conducting research related to their own then it will have served its purpose.

I have been careful to refer here to "text retrieval" because I do not plan to discuss the emerging work on cross-language speech retrieval in this survey. One goal of the Symposium on Cross-Language Text and Speech Retrieval sponsored by the American Association for Artificial Intelligence that is meeting at Stanford University later this month is to explore the state of the art on that topic as well. Traditionally "text retrieval" and "information retrieval" have been used interchangeably, but as retrieval from other modalities (e.g., speech or images) has become more practical it is becoming more common to be explicit about the sort of information being retrieved.

I will not attempt to draw a sharp distinction between retrieval and filtering in this survey. Although my own work on adaptive cross-language text filtering has led me to make this distinction fairly carefully in other presentations (c.f., [21]), such an approach does little to help understand the fundamental techniques which have been applied or the results that have been obtained in this case. Since it is still common to view filtering (detection of useful documents in dynamic document streams) as a kind of retrieval, I will simply adopt that perspective here.

2 Fundamental Approaches

Cross-language text retrieval has an extensive research heritage. The first practical approach to cross-language text retrieval required that the documents be manually indexed using a predetermined vocabulary and that the user express the query using terms drawn from that same vocabulary. This is referred to as a "controlled vocabulary" approach. In such systems a multilingual thesaurus is used to relate the selected terms from each language to a common set of language-independent concept identifiers, and document selection is based on concept identifier matching. In the hands of a skilled user who is familiar with controlled vocabulary search techniques, such systems can be remarkably effective. Of particular note, if well designed, controlled vocabulary cross-language text retrieval systems can be just as effective as monolingual applications of similar techniques. Controlled vocabulary cross-language text retrieval systems are presently widely used in commercial and government applications for which the number of concepts (and hence the size of the indexing vocabulary) is manageable. Unfortunately, the requirement to manually index the document collection makes controlled vocabulary text retrieval techniques unsuitable for high-volume applications in which the documents are generated from diverse sources that are not easily standardized.

This limitation has motivated the search for approaches which are amenable to less well structured situations. The alternative to the use of a controlled vocabulary is to use the words which appear in the documents themselves as the vocabulary. Such systems are referred to as free text (or sometimes full text) retrieval systems. Two basic approaches to cross-language free text retrieval have been emerged: dictionary-based approaches and corpus-based approaches.

Dictionary-based approaches essentially seek to extend the fundamental idea of a multilingual thesaurus by using bilingual dictionaries to translate the query into every language in which a document might be found. Two factors limit the performance of this approach. The first is that many words do not have a unique translation, and sometimes the alternate translations have very different meanings. Monolingual text retrieval systems face similar challenges from homonomy and polysemy (multiple meanings for a single term), but this translation ambiguity significantly exacerbates the problem. Use

[1] Links to every known internet-accessible cross-language text retrieval resource can be found at http://www.ee.umd.edu/medlab/mlir/

of every possible translation, for example, can greatly expand the set of possible meanings because some of those translations are likely to introduce additional homonomous or polysemous word senses in the second language. This problem is particularly severe in view of the observed tendency of untrained users to enter such short queries (often a single word) that it would not even be possible for a human to determine the intended meaning (and hence the proper query translation) from the available context.

The second problem with a dictionary-based approach is that the dictionary may lack some terms that are essential for a correct interpretation of the query. This may occur either because the query deals with a technical topic which is outside the scope of the dictionary or because the user has entered some form of abbreviation or slang which is not included in the dictionary. As dictionaries specifically designed for query translation are developed, the effect of this limitation may be reduced. But it is unlikely to be eliminated completely because language use is a creative activity, with new terms entering the lexicon all the time. There will naturally be a lag between the introduction of a term and its incorporation into a standard reference work such as a dictionary.

Corpus-based approaches seek to overcome these limitations by constructing query translation techniques which are appropriate for the way language is used in a specific application. Because it would be impractical to construct large tailored bilingual dictionaries manually, corpus-based approaches instead analyze large collections of existing text and automatically extract the information needed to construct these application-specific translation techniques. The collections which are analyzed may contain existing translations and the documents that were translated (a "parallel" collection), or they may be composed of documents on similar subjects which are written in different languages (a "comparable" collection).

Present corpus-based approaches are limited by two factors. The most significant limitation is that a parallel document collection which uses language in a manner similar to that found in the application may not be available in a suitable form. Techniques based on comparable document collections may eventually overcome this limitation, but research on the use of comparable document collections for text retrieval is presently at a very early stage [25]. While a translation technique developed from a parallel document collection can be used for unrelated applications, significant reductions in retrieval effectiveness should be expected.

The other limitation of corpus-based techniques is that even when a suitable document collection is available, the methods presently used to extract the information on which the translation technique will be based introduce errors as well. The field of "corpus linguistics" has explored the use of corpus-based techniques in to a variety of applications such as text retrieval, speech recognition, machine translation and ontology construction. Initial corpus-based experiments typically emphasize statistical analysis over linguistic theory, an approach which has led to some remarkable successes. In machine translation, for example, early statistical approaches demonstrated performance that was competitive with that achieved by contemporaneous linguistically motivated approaches [3]. But purely statistical approaches also introduce errors that no human would make because the techniques typically exploit term cooccurrence and some of the cooccurrence information can be misleading. One recurring theme in corpus linguistics is that significant performance improvements can be achieved when appropriate linguistically motivated constraints are effectively integrated with the statistical analysis. Since corpus-based techniques for cross-language text retrieval are for the most part still in the early "statistics only" phase, integration of linguistic constraints with these techniques appears to be a promising direction for future research.

As an example of how linguistically motivated techniques might be incorporated, consider the case of what has been called "phrase indexing." No corpus-based system that I know of has yet demonstrated cross-language text retrieval effectiveness on a par with the within-language effectiveness of the same underlying retrieval techniques in the absence of a perfectly matched parallel document collection. But three European research groups have reported dramatic improvements in performance when phrases are processed in addition to individual words, presumably because the use of phrases constrains translation ambiguity [13, 26, 31] and in some initial experiments with phrase indexing I have recently obtained similar results.

3 Research in the USA

Although there was cross-language text retrieval work reported in Europe as early as 1964, the earliest reported work in the USA was performed by Salton at Cornell University in 1969 [28]. Salton aug-

mented the SMART text retrieval system with a small cross language dictionary that was developed by translating some of the words in an existing English concept list (a simple type of thesaurus with only synonym links) into German. Although Salton used full-text indexing, the limited size of his thesaurus and the small scope of his test collection produced results similar to those achieved with controlled vocabulary systems. From these experiments Salton concluded that although retrieval effectiveness varied across document collections (a well known phenomenon in text retrieval), "cross-language processing ... is nearly as effective as processing within a single language." After examining the retrieval failures in more detail Salton concluded that "it would therefore seem essential that a more complete thesaurus be used under operational conditions for future experiments." For a 1973 paper Salton implemented an English-French multilingual concept list, this time achieving more complete coverage by independently developing the section for each language after establishing a common set of concepts [27], but interpretation of the results was hampered by the small size of the evaluation collection that was used.

Salton's later work moved away from thesaurus-based techniques, but interest in the use of multilingual thesauri as a basis for controlled vocabulary cross-language text retrieval flourished in both Western Europe and the former Soviet Union. Although my recent survey of cross-language text retrieval did not identify a single instance of experimental work on that topic in the USA [23], the National Library of Medicine (NLM) is presently developing the Unified Medical Language System (UMLS). One goal of that project is to integrate existing French, German, Spanish and Portuguese translations of the Medical Subject Headings (MeSH) controlled vocabulary into a single thesaurus [29].[2] The UMLS multilingual thesaurus will be used to provide subject access to NLM's extensive collection of bibliographic records. There also are at least two American companies that have significant experience and experience with controlled vocabulary cross-language text retrieval, although quite a bit of their business base for these products is comprised of overseas customers. VTLS provides multilingual library automation software and Access Innovations develops customized text retrieval applications [4, 12].

Cross-language free text retrieval research has received considerably more attention in the USA, but virtually all of the activity has occurred since 1990. While this has produced a vibrant and rapidly expanding research community, little if any of the work here has been informed by the extensive European experience with controlled vocabulary cross-language text retrieval. This is likely a result of the temporal and geographic separation between the present projects and the burst of activity on controlled vocabulary cross-language text retrieval that occurred in Europe in the 1970's which culminated in the development of ISO standard 5964 for multilingual thesaurus development in 1978 and its most recent revision in 1985. Only now are we beginning to ask what lessons can be learned from that research that would be useful for the applications presently being investigated.

In 1990, Landauer and Littman (then with Bellcore) developed a corpus-based cross-language free text retrieval technique which has come to be known as Cross-Language Latent Semantic Indexing (CL-LSI) [17, 18]. The remarkable thing about this work is that in addition to beginning the present development of cross-language text retrieval systems in the USA, it remains to this day the only technique that has demonstrated cross-language text retrieval effectiveness that is on a par with the within-language performance of that same technique [9]. This result is particularly significant because a monolingual text retrieval system based on Latent Semantic Indexing has achieved effectiveness measures nearly equal to those of the best participating systems at the third Text Retrieval Conference [8].

In CL-LSI a set of representative bilingual documents are first used to form a training collection by adjoining a translation of each document to the document itself. A rank revealing matrix decomposition (the singular value decomposition) is then used to compute a mapping from sparse term-based vectors (usually with weights base on both within-document and collection-wide term frequency) to short but dense vectors that appear to capture the conceptual content of each document while suppressing the effect of variations in term usage. CL-LSI appears to achieve it's effectiveness by suppressing cross-language variations in term choice as well. In principle this technique can be extended to multiple languages, although the retrieval effectiveness of such a configuration has not yet been determined experimentally. Berry and Young repeated this work using passages from the Bible in English and Greek [2]. They were able to demonstrate that fine-grained training data, using only the first verse of each passage to identify the principal components, improved retrieval performance over Landauer and Littman's coarser approach.

It is important to caveat the reported results for LSI by observing that both sets of experiments were conducted with experiment designs that matched the retrieval application to the characteristics of

[2] Information on the presence of non-English terminology in UMLS is available at http://www.nlm.nih.gov/publications/factsheets/umls_metathesaurus.html

the parallel document collection that was used to develop the translation technique. Our experiments with this technique show a significant reduction in performance when a parallel document collection that is more weakly related to the retrieval application is used [20]. This limitation is not unique to CL-LSI, however. It results from the fact that corpus-based techniques generally seek to balance the adverse effect of invalid inferences that result from misleading statistical cooccurrence observations with the beneficial effects of correctly recognizing that only a limited number of senses for words with several possible meanings are present in the training collection. As term use in the training and evaluation collections begins to diverge, this "beneficial" effect rapidly becomes a liability.

One particularly attractive feature of CL-LSI is that the short dense feature vectors that are computed for each document are inherently language-independent. Most other techniques require that language identification be performed so that appropriate processing can be applied to each document. This is not a significant obstacle, however, since language identification techniques with better than 95% accuracy are available [14].

Davis and Dunning of New Mexico State University have recently conducted the first large-scale evaluations of cross-language text retrieval techniques using material from the Text Retrieval Conferences (TREC-4 and TREC-5) [6, 5]. For the evaluation of Spanish text retrieval at TREC-4 they manually translated 25 Spanish queries into English and then used them to select documents from a collection of about 58,000 Spanish language newspaper articles with a modified version of the Inquiry text retrieval system developed at the University of Massachusetts. Although these first experiments were for the most part unsuccessful, Davis' recent TREC-5 experiments on another 25 queries translated manually from Spanish to English and 173,000 Spanish language newswire stories have produced about 75% of the average precision achieved in a monolingual evaluation using the same system and collection. In these later experiments Davis used a modified version of the Cornell University SMART system.

Davis' results indicate that when used alone, dictionary-based query expansion achieves about 50% of the average precision that would be achieved by a monolingual system, but that when translation ambiguity is limited this performance can be improved. This is quite consistent with similar results that have been obtained in Europe on smaller collections [13], suggesting that although the size of the collection may affect absolute performance measures, the effect on relative performance between monolingual and cross-language retrieval may be inconsequential. To improve over this baseline, Davis limited the dictionary-based query expansion using part-of-speech information that was determined statistically combined with additional constraints on the permissible translations that were determined using a large parallel corpus. This work is particularly interesting because it combines dictionary-based and corpus-based techniques in a single retrieval system. And because the content of the parallel corpus of United Nations documents that was used was not particularly closely related to the content of the newswire stories, these experiments offer some insight into the effect of a domain mismatch as well.

Davis' corpus-based approach for restricting translation ambiguity seeks to select translations which would select similar sentences from documents in the parallel document collection. The technique is based on similarity matching between a vector which represents the query and vectors which represent individual sentences in the document collection. Thus, Davis is exploring a technique based on sentence-level alignment in a parallel collection in contrast to the coarser document-level alignment on which CL-LSI is based.

At the University of Maryland, Dorr and I have developed a technique based on term-level alignment which also offers the potential for integration of dictionary-based and corpus-based techniques [24]. The basic idea is to estimate the domain-specific probability distribution on the possible translations of each term based on the observed frequency with which terms align in a parallel document collection. We then use this statistically enhanced bilingual dictionary as a linear operator to rapidly map the vectors which represent documents from one language into another. The effectiveness of this technique depends on a sort of "consensus translation effect" in which several terms in the source language can potentially contribute to the weight assigned to a single term in the target language. As a result, it is only practical to apply our vector translation technique to vectors which represent documents. Typical queries simply don't contain enough terms or enough variation in term usage to develop a useful consensus translation effect. This limitation fits well with our focus on cross-language text filtering because our adaptive information need representation is not amenable to query translation.

In our initial experiments we have used a purely corpus-based approach for developing our statistically enhanced dictionary. In an evaluation conducted using the same collections used by Davis (and one additional TREC collection), we found that that implementation of our technique achieves about half the effectiveness of CL-LSI. An examination of the transfer mapping developed from our term

alignment step reveals that many of the detected alignments do not represent valid translations. This is not a surprising result since term alignment is a challenging problem which is presently the focus of a good deal of research effort. In our next experiments we plan to constrain the allowable translations to those which occur in a broad-coverage bilingual dictionary, seeking to match or exceed the performance of CL-LSI.

In addition to demonstrating three techniques for adaptive cross-language text filtering, our work at Maryland has made two other contributions that may be of interest. The first is that we have developed a methodology for evaluating corpus-based adaptive cross-language text filtering effectiveness which does not depend on the development of an expensive specialized test collection [22]. A fair evaluation of such techniques requires two training collections, one of which must be a parallel bilingual corpus, and one evaluation collection. We have found a way to align TREC topic descriptions that were originally developed independently for each language and then to measure the quality of that alignment. Our approach is based on a very small number of topics, but until a suitable test collection is available for cross-language filtering evaluation it represents the best technique I know of for conducting such evaluations. A second useful result is that we have developed a technique to measure the degradation in effectiveness which results from the different domains of the UN collection and the Spanish documents used in TREC. This may be helpful when interpreting Davis' results, and more broadly it may offer some insight into the fundamental limits on the performance of corpus-based techniques when a well-suited parallel document collection is not available.

Dictionary-based cross-language text retrieval is being investigated by Ballesteros and Croft at the University of Massachusetts [1], but even this work has a significant corpus-based aspect to it. By exploiting a pseudo-relevance feedback technique that has been shown to be effective for within language retrieval, they have achieved significant performance improvements over unconstrained dictionary-based query translation. These techniques essentially seek to modify the query to more closely resemble the documents in the collection. They achieved their best results when performing this technique twice, once before the dictionary-based query translation and once before using the translated query to rank order the documents in the evaluation collection. This technique requires the availability of document collections in each language, but it is not necessary that the individual documents in these collections be related in any way. Thus this approach is similar to techniques being investigated in Europe that use comparable rather than parallel corpora to improve the performance of a dictionary-based technique [30], but it offers the potential to avoid the need for document alignment completely.

In addition to these major projects there have been a number of smaller efforts which I will review here briefly for completeness. Evans, *et al.* at Carnegie Mellon University investigated a different cross-language application of Latent Semantic Indexing, using it to suggest terms from a controlled vocabulary of 125 English medical terms based on natural language queries expressed in Spanish [10]. Their report presents a couple of examples in which the most highly ranked terms would be good choices for use in a controlled vocabulary search, but the focus of the paper was on automatic thesaurus construction, so no cross-language retrieval experiments using these data were reported.

Lin and Chen at the University of Arizona have also recently conducted a small-scale experiment on automatic multilingual thesaurus construction [19]. They used a Hopfield neural network to cluster about 1000 titles from Chinese technical papers, many of which contained a mixture of Chinese and English words. They found 36 clusters which together accounted for about two thirds of the titles and reported that manual inspection showed that the terms associated with "all concept descriptors appeared to be relevant and precise" and that some clusters contained both Chinese and English terms. Like Evans, *et al.* , Lin and Chen did not actually use the resulting thesaurus in a retrieval experiment. As a result, this work is more interesting as an application of neural networks than for the insight it provides into important issues in cross-language text retrieval.

Relatively simple cross-language text retrieval capabilities have been added to commercial text retrieval systems produced by two American companies. Paracel produces a text filtering system based on special purpose parallel processing hardware for which the information need must be explicitly given by the user.[3] Provisions are provided to translate these information need specifications between a limited number of languages. If expert assistance is available to refine the version of the specification for each language over time, the system offers the potential for fast and effective filtering of documents in multiple languages. But although the specification translation function produces some cross-language retrieval functionality, manual tuning in each language is necessary to achieve optimum performance. So although Paracel's approach provides an elegant solution in their particular application, it offers

[3] Paracel Inc., 80 South Lake Avenue, Suite 650, Pasadena, CA 91101-2616

relatively little insight into the potential performance of fully automatic cross-language text retrieval systems.

Gachot, *et al.* at SYSTRAN have recently described a cross-language text retrieval tool which is based on capabilities that were originally developed for their existing machine translation system [11]. Both exact-match queries and ranked retrieval based on similarity are supported by the design they describe, but their research is in such a preliminary stage that it is not yet clear how their sophisticated linguistic representations will affect retrieval performance. The results of the European Multilingual Information Retrieval (EMIR) project do indicate that such an approach has potential [26], so it will be interesting to see how this project develops.

Finally, there are two academic research efforts in early stages of their development which bear watching. Frederking, *et al.* at Carnegie Mellon University are beginning an ambitious project which seeks to integrate the cross-language text retrieval with some form of translation to assist the user with document selection. Although my focus in this survey has been on the performance of fully automated cross-language text retrieval systems, as experimental systems evolve towards practical applications, issues such as this will become increasingly important. Others have investigated the value of rough translations for document selection, but the Carnegie Mellon University project represents the first substantial effort in the USA to create a complete cross-language text retrieval system that includes this capability. It will be interesting to compare the results they achieve with those reported by Kikui, *et al.* from Japan [15].

The other new project is being pursued by Kwok at Queens College in New York. Building on his earlier work on Chinese text retrieval, Kwok has conducted some initial dictionary-based cross-language text retrieval experiments between English and Chinese [16]. This is an interesting language pair because written Chinese lacks indications of word boundaries. Since that characteristic is common to several Asian languages, the techniques Kwok develops may be of interest to researchers in Europe who are working on cross-language text retrieval applications for which one of the languages is difficult to segment into isolated words.

4 Research Sponsors

In the USA, both the Defense Advanced Research Projects Agency (DARPA) and the National Science Foundation (NSF) support cross-language text retrieval research. A very rough estimate of the total support might be a couple of million dollars each year, approximately evenly divided between the two agencies. DARPA supports cross-language text retrieval research through three programs, TIPSTER, a new Broad Agency Announcement (BAA) numbered 97-09, and Small Business Innovative Research (SBIR) grants. Together these programs address near-, mid- and long-term technology investment strategies.

In conjunction with several other government sponsors, the DARPA Information Technology Office (ITO) has supported an effort known as TIPSTER which investigates both text retrieval and text extraction since 1991. The TIPSTER Phase II effort included a "multilingual entity task" in which some of the participants worked extensively on cross-language text retrieval issues [7]. Other participants in that task investigated only monolingual issues, but did so in several languages. The American participants were New Mexico State University, SRI International, Systems Research and Applications (SRA), BBN Systems and Technologies Corporation, and the MITRE Corporation. Although the proceedings of the Phase II workshop is now in print, I have not yet had the opportunity to carefully review it to determine the scope and significance of the cross-language text retrieval research that was reported. The TIPSTER Phase III effort that began in 1996 is supporting multilingual entity task research at New Mexico State University, Queens College and the University of Southern California Information Sciences Institute.

DARPA ITO has recently issued a BAA 97-09, seeking innovative proposals for a variety of information management and collaboration support tasks for which the marketplace is unlikely to provide solutions in the next five years. Cross-language text (and speech) retrieval is an important aspect of some of these tasks, and during the proposers' brief the DARPA program managers indicated that there is a potential for significant support for such efforts.[4] BAA 97-09 proposals are due to DARPA on February 26, 1997.

[4] Additional information on BAA 97-09 is available at http://www.ito.darpa.mil/Solicitations.html

Small, near-term research efforts are sometimes sponsored through the SBIR program. In 1994, DARPA awarded a SBIR grant to Textwise Inc. for feasibility study of a cross-language text retrieval. The proposed system, known as CINDOR, was designed to automatically navigate a multilingual thesaurus.[5] Since SBIR awards seek to develop systems with commercial potential that would also be of interest to the sponsoring agency, they often do not result in academic publication of the research results. This appears to have been the case for the Textwise research.

In the USA, NSF funds basic and applied research which seeks to establish a technology base for future developments. NSF has recently initiated a five year multimedia retrieval research program known as STIMULATE which includes a substantial cross-language text (and speech) retrieval component. Proposals were due in September 1996 and the initial research grants are in the process of being awarded. Details on these awards are not yet available. Additional calls for proposals are expected in 1997 and 1998.

In conjunction with DARPA, the National Institutes of Standards and Technology (NIST) sponsors an annual Text REtrieval Conference (TREC) at which various aspects of text retrieval are explored. The distinguishing feature of TREC is that it includes a simultaneous blind evaluation of a fairly large number of participating systems, each of which uses an identical test collection. The number of participating systems offers significant economies of scale and makes cross-system comparisons more practical than it would be with isolated experiments. It is this venue in which Davis conducted the cross-language text retrieval experiments reported above. Those experiments were singletons, however, lacking a group of participating systems with which the results could be compared. In the 1997 TREC-6 evaluation, NIST has agreed to support test corpus acquisition and blind evaluation for a special interest "pre-track" in which the practicality of large scale cross-language text retrieval system evaluation will be explored. If this pre-track is successful, NIST may elect to include a cross-language text retrieval special interest track in subsequent Text Retrieval Conferences.

The cross-language track at TREC should be of particular interest to European research groups because NIST accepts TREC participants from outside the USA as well. The deadline has already passed for TREC-6 applications, but groups participating in any part of TREC-6 are generally able to participate in any of the special interest tracks simply by notifying the coordinator for that track of their interest. For groups wishing to join participate in the future, TREC-7 applications will likely be due in early January, 1998. If suitable licensing arrangements can be made with the corpus providers, it may also be possible for nonparticipating systems to obtain the test collection and relevance judgments that are developed after the TREC-6 conference meets in November 1997.

5 Conclusions

My goal in this presentation has been to help draw together the worldwide cross-language text retrieval research community. Concentrating on the achievements and limitations of the work in a single nation may seem like a strange way to approach such a task for what is clearly a transnational problem. But although I have read extensively on the work here in Europe, we have only recently begun to forge the international partnerships that I believe will be necessary if we are to meet the demands of the worldwide market for cross-language text retrieval systems. So I have come here to help clarify what it is that we in the USA can bring to such joint ventures and to learn more about the diverse research currently underway here in Europe. The timing of this workshop is indeed fortuitous, since we will be able to use what we learn here to help shape the symposium that David Hull and I are co-chairing at Stanford later this month. With nearly half of the 50 or so participants in that symposium coming from outside the USA, I am hopeful that we will be able to effectively use that opportunity to further explore the potential for international collaboration on this important problem.

6 Acknowledgments

The author would like to thank Bonnie Dorr and Dagobert Soergel their helpful comments.

[5] Additional details on this award are available at ftp://ftp.dtic.dla.mil/pub/sbir/sb94arpa.awd

References

[1] Lisa Ballesteros and Bruce Croft. Ductionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems*, pages 791–801, 1996. http://ciir.cs.umass.edu/info/psfiles/irpubs/ir.html.

[2] M. Berry and P. Young. Using latent semantic indexing for multilanguage information retrieval. *Computers and the Humanities*, 29(6):413–429, December 1995.

[3] Peter F. Brown, John Cocke, Steven A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June 1990.

[4] Vinod Chachra. Subject access in an automated multithesaurus and multilingual environment. In Sally McCallum and Monica Ertel, editors, *Automated Systems for Access to Multilingual and Multiscript Library Materials*, pages 63–76. International Federation of Library Associations and Institutions (IFLA), K. G. Saur, August 1993.

[5] Mark Davis. New experiments in cross-language text retrieval at NMSU's Computing Research Lab. In D. K. Harman, editor, *The Fifth Text REtrieval Conference (TREC-5)*. NIST, November 1996. To appear. http://crl.nmsu.edu/users/madavis/Site/Book2/trec5.ps.

[6] Mark Davis and Ted Dunning. A TREC evaluation of query translation methods for multi-lingual text retrieval. In D. K. Harman, editor, *The Fourth Text Retrieval Conference (TREC-4)*. NIST, November 1995. http://crl.nmsu.edu/users/madavis/Site/Book2/trec4.ps.

[7] Defense Advanced Research Projects Agency. *Tipster Text Program*. Morgan Kaufmann, 1996.

[8] S. T. Dumais. Latent Semantic Indexing (LSI): TREC-3 report. In Donna Harman, editor, *Overview of the Third Text REtrieval Conference*, pages 219–230. NIST, November 1994. http://www-nlpir.nist.gov/TREC/.

[9] Susan T. Dumais, Thomas K. Landauer, and Michael L. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In Gregory Grefenstette, editor, *Working Notes of the Workshop on Cross-Linguistic Information Retrieval*. ACM SIGIR, August 1996. http://superbook.bellcore.com/~std/papers/SIGIR96.ps.

[10] D. A. Evans, S. K. Handerson, I. A. Monarch, J. Pereiro, L. Delon, and W. R. Hersh. Mapping vocabularies using "latent semantics". Technical Report CMU-LCL-91-1, Carnegie Mellon University, Laboratory for Computational Linguistics, July 1991.

[11] Denis A. Gachot, Elke Lange, and Jin Yang. The SYSTRAN NLP browser: An application of machine translation technology in multilingual information retrieval. In *Workshop on Cross-Linguistic Information Retrieval*, pages 44–54. ACM SIGIR, August 1996.

[12] Marjorie M. K. Hlava, Richard Hainebach, Gerold Belonogov, and Boris Kuznetsov. Cross-language retrieval - English/Russian/French. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence, March 1997. To appear. http://www.ee.edu/medlab/filter/sss/papers/hlava.ps.

[13] David A. Hull and Gregory Grefenstette. Experiments in multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996. http://www.xerox.fr/people/grenoble/hull/papers/sigir96.ps.

[14] Gen-itiro Kikui. Identifying the coding system and language of on-line documents on the internet. In *Sixteenth International Conference of Computational Linguistics (COLING)*. International Committee on Computational Linguistics, August 1996. http://isserv.tas.ntt.jp/chisho/paper/9608KikuiCOLING.ps.Z.

[15] Genichiro Kikui, Yoshihiko Hayashi, and Seiji Suzaki. Cross-lingual information retrieval on the WWW. In *Multilinguality in Software Engineering: The AI Contribution*. European Coordinating Committee for Artificial Intelligence, August 1996. http://isserv.tas.ntt.jp/chisho/paper/9608KikuiMULSAIC.ps.Z.

[16] K. L. Kwok. Evaluation of english-chinese cross-lingual retrieval experiment. In *AAAI Symposium on Cross Language Text and Speech Retrieval*. American Association for Artificial Intelligence, March 1997. To appear. http://www.ee.umd.edu/medlab/filter/sss/papers/kwok.ps.

[17] Thomas K. Landauer and Michael L. Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 31–38. UW Centre for the New OED and Text Research, Waterloo Ontario, October 1990. http://www.cs.duke.edu/~mlittman/docs/x-lang.ps.

[18] Thomas K. Landauer and Michael L. Littman. A statistical method for language-independent representation of the topical content of text segments. In *Proceedings of the Eleventh International Conference: Expert Systems and Their Applications*, volume 8, pages 77–85, Avignon France, May 1991.

[19] Chung-hsin Lin and Hsinchun Chen. An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents. *IEEE Transactions on Systems, Man and Cybernetics*, 26(1):75–88, February 1996. http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html.

[20] Douglas W. Oard. Adaptive filtering of multilingual document streams. In *Submitted to RIAO 97*, June 1997.

[21] Douglas W. Oard. The state of the art in text filtering. *User Modeling and User Adapted Interaction*, 1997. To appear.

[22] Douglas W. Oard and Bonnie J. Dorr. Evaluating cross-language text filtering effectiveness. In Gregory Grefenstette, editor, *Proceedings of the Cross-Linguistic Multilingual Information Retrieval Workshop*. ACM SIGIR, August 1996. http://www.ee.umd.edu/medlab/filter/papers/sigir96.ps.

[23] Douglas W. Oard and Bonnie J. Dorr. A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies, April 1996. http://www.ee.umd.edu/medlab/filter/papers/mlir.ps.

[24] Douglas William Oard. *Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications*. PhD thesis, University of Maryland, College Park, August 1996. http://www.ee.umd.edu/medlab/filter/papers/thesis.ps.gz.

[25] Eugenio Picchi and Carol Peters. Cross language information retrieval: A system for comparable corpus querying. In *Workshop on Cross-Linguistic Information Retrieval*, pages 24–33. ACM SIGIR, August 1996.

[26] Khaled Radwan and Christian Fluhr. Textual database lexicon used as a filter to resolve semantic ambiguity application on multilingual information retrieval. In *Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 121–136, April 1995.

[27] G. Salton. Experiments in multi-lingual information retrieval. *Information Processing Letters*, 2(1):6–11, March 1973. TR 72-154 at http://cs-tr.cs.cornell.edu.

[28] Gerard Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21(3):187–194, May 1970.

[29] Peri L. Schuyler, William T. Hole, Mark S. Tuttle, and David D. Sheretz. The UMLS metathesaurus: representing different views of biomedical concepts. *Bulliten of the Medical Library Association*, 81(2):217–222, April 1993.

[30] Páraic Sheridan and Jean Paul Ballerini. 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*, August 1996. http://www-ir.inf.ethz.ch/Public-Web-Pages/sheridan/papers/SIGIR96.ps.

[31] Pim van der Eijk. Automating the acquisition of bilingual terminology. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, pages 113–119, April 1993.