Cross-Language Information Retrieval (CLIR) Track Overview

Peter Schäuble, Páraic Sheridan

Swiss Federal Institute of Technology (ETH)
CH-8092 Zürich, Switzerland

1 Introduction

Cross-Language Information Retrieval (CLIR) was a new task in the TREC-6 evaluation. In contrast to the *multilingual* track included in previous TREC evaluations, which was concerned with information retrieval in Spanish or Chinese, the *cross-language* retrieval track focuses on the retrieval situation where the documents are written in a language which is *different* than the language used to specify the queries. The TREC-6 track used documents in English, French and German and queries in English, French, German, Spanish and Dutch.

There are many applications or scenarios in which a user of a retrieval system may be interested in finding information written in a language other than the user's native or preferred language. In some applications, a user may want to discover all possible relevant information in a multilingual textbase, irrespective of the language of the relevant information. This may be the case when searching certain collections of legal or patent information for example. In other cases a user may even have some language comprehension ability in the languages of the documents (passive vocabulary) but may not have a sufficiently rich active vocabulary in the document languages to confidently specify queries in those languages. In this case a cross-language search which permits the user to specify native language queries but retrieves documents in their original language is useful. Cross-language retrieval also has the added advantage of requiring only one query to a multi-lingual text collection, rather than having a user submit individual queries in each of the languages of interest.

Situations where a retrieval system user is faced with the task of querying a multilingual document collection are becoming increasingly common. These range across document collections made up of documents from local offices of multinational companies, collections composed of documents from different regions of multilingual countries such as Switzerland or Canada, or the document collections of large organisations such as the United Nations or European Commission. It is however, the global information infrastructure of the internet that has been largely responsible for the growing awareness of a need for cross-language information retrieval systems. This has in turn led to a growing body of research into the problems of cross-language retrieval and the development of several different approaches for CLIR.

1.1 Approaches to CLIR

Central to the problem of cross-language information retrieval is the matching of user queries specified in one language against documents written in a different language - *crossing the language boundary*. This may be achieved by translating the user queries, translating the documents, or by translating both the queries and documents, perhaps into some intermediary or interlingual representation. Both queries and documents might be translated into a common controlled indexing vocabulary for example, either manually or automatically. More common however, are approaches which work with the full text of documents and queries for matching. These can usefully be classified according to what resources are used to aid in crossing the language boundary; machine translation, machine-readable dictionaries, or corpus resources.

NEC in Japan (Yamabana et al., 1998) have used machine translation technology for cross-language retrieval by translating users' queries in an interactive process using both dictionaries and statistical information derived from bilingual corpora. The machine translation company Systran have also reported work addressing the cross-language retrieval problem with a system that includes machine translation technology at all stages of the retrieval process and which allows users to include detailed linguistic information in queries (Gachot et al., 1998). Researchers at Carnegie Mellon University (CMU) (Carbonell et al., 1997) have investigated the use of translation techniques from the study of Example-based Machine Translation (EBMT) for cross-language retrieval. The evaluation of machine translation as an approach to CLIR has been facilitated by a collaboration between the University of Maryland and the LOGOS corporation in the TREC-6 CLIR track.

Approaches to cross-language information retrieval which rely on corpus resources include the use of Latent Semantic Indexing (LSI) by researchers at Bellcore and elsewhere (Rehder et al., 1998) (Littman et al., 1998) (Landauer and Littman, 1991), the Generalised Vector Space Model proposed by CMU (Carbonell et al., 1997), a corpus-based approach from CNR in Pisa (Peters and Picchi, 1997), and our work at ETH which uses *similarity thesauri* for query translation (Mateev et al., 1996) (Sheridan et al., 1997) (Sheridan and Schäuble, 1997) (Sheridan and Ballerini, 1996). A common thread in these approaches is the use of corpus resources to train the cross-language retrieval mechanism or to build the information structures used for retrieval. Several of these approaches have also been evaluated in this TREC-6 track.

The use of existing linguistic resources, especially machine-readable bilingual dictionaries, is a natural approach to cross-language retrieval. Researchers at the Xerox Research Centre Europe (Hull and Grefenstette, 1996) (Grefenstette et al., 1998), the University of Massachusetts (Ballesteros and Croft, 1996) (Ballesteros and Croft, 1998), and the Computing Research Laboratory (CRL) of New Mexico State University (NMSU) (Davis, 1998) (Davis and Ogden, 1997), have extensively investigated the use of machine-readable dictionaries for cross-language retrieval and have variously addressed ways of overcoming some of the problems with dictionary-based translation. Many groups have submitted results based on machine-readable dictionaries for the TREC-6 evaluation.

Apart from the use of controlled vocabulary indexing or the use of manually-constructed multilingual thesauri, the TREC-6 CLIR evaluation has covered all classes of approaches to cross-language retrieval.

2 CLIR-Track Task Description

The Cross-Language Information Retrieval (CLIR) track requires the retrieval of either English, German or French documents that are relevant to topics formated in different languages. Participating groups were to choose any cross-language combination, for example English queries against German documents or French queries against English documents. In order have a baseline retrieval performance measurement for each group, the results of a monolingual retrieval experiment in the document language were also to be submitted, corresponding to each cross-language experiment run. For instance, if a cross-language experiment was run with English queries retrieving German documents then the result of the equivalent experiment where German queries retrieve German documents must also be submitted. These results are considered comparable since the queries are equivalent across the languages.

The document collections for the CLIR track were not however equivalent in English, French and German. The different document collections used in each language are outlined in Table 2. The Associated Press collection consists of newswire stories in English. The French SDA collection is a similar collection of newswire stories from the Swiss news agency (Schweizerische Depeschen Agentur). The German document collection has two parts. The first part is composed of further newswire stories from the Swiss SDA while the second part consists of newspaper articles from a Swiss newspaper, the 'Neue Zuercher Zeitung' (NZZ). The newswire collections in English, French and German were chosen to overlap in timeframe (1988 to 1990) for two reasons. First, since a single set of topics was going to be formulated to cover all three document languages, having the same timeframe for newswire stories increased the likelihood of finding a greater number of relevant documents in all languages. The second reason for the overlapping timeframe was to allow groups who use corpus-based approaches to cross-language retrieval to investigate what useful corpus information they could extract from the document collections being used.

Document Collections			
Doc. Language	**Source**	**No. Documents**	**Size**
English	AP news, 1988-1990	242,918	760MB
German	SDA news, 1988-1990	185,099	330MB
	NZZ articles, 1994	66,741	200MB
French	SDA news, 1988-1990	141,656	250MB

Table 1: Document Collections used in the CLIR track.

The use of corpus-based approaches was further facilitated by the comparable nature of the SDA collections in German and French. These are newswire stories prepared by the same Swiss news agency, though the stories do not overlap perfectly and are not translated between the two languages. The stories are produced independently in each language, but there is in fact a high overlap of stories (e.g. international events) which are of interest in both the German-speaking and French-speaking parts of Switzerland. One of the resources provided to CLIR track participants was a list of 83,698 news documents in the French and German SDA collections which were likely to be comparable based on an alignment of stories using news descriptors assigned manually by the SDA reporters, the dates of the stories, and common cognates in the texts of the stories.

The 25 test topic (query) descriptions were provided by NIST in English, French and German. The 25 topics were equivalent across the languages. Participating groups who wished to test other query languages were permitted to create translations of the topics in their own language and use these in their tests, as long as the translated topics were made publicly available to the rest of the track participants. The final query set therefore also had translations of the 25 topics in Spanish, provided by the University of Massachusetts, and Dutch, provided by TNO in the Netherlands.

Although not strictly within the definition of the cross-language task, participation by groups who wanted to run mono-lingual retrieval experiments in either French or German using the CLIR data was also permitted. Since the CLIR track was run for the first time this year, this was intended to encourage new IR groups working with either German or French to participate. The participation of these groups also helped to ensure that there would be a sufficient number of different system submissions to provide the pool of results needed for relevance judgements.

The evaluation of CLIR track results was based on the standard TREC-adhoc evaluation measures. Participating groups were free to experiment with different query length and with both automatic and manual experiments according to the definitions used for the main TREC adhoc task.

3 Results

A total of thirteen different groups, representing seven different countries, participated in the TREC-6 CLIR track. The first important result from this track therefore, was the participation of new groups in TREC, especially new groups from Europe. Participating groups were encouraged to run as many different experiments as possible, both with different kinds of approach to CLIR and with different language combinations. An overview of the submitted runs is given in Table 3. This shows that the main query languages were used equally, each used in 29 experiments, whereas English was somewhat more popular than German or French as the choice for the document language to be retrieved. This is in part because the groups who used the query translations in Spanish and Dutch only evaluated those queries against English documents. A total of 95 result sets were submitted for evaluation in the CLIR track.

Language Combinations						
	Query Language					
Doc. Language	*English*	*German*	*French*	*Spanish*	*Dutch*	**Total**
English	7	15	10	2	6	**40**
German	12	10	4	-	-	**26**
French	10	4	15	-	-	**29**
Total	**29**	**29**	**29**	**2**	**6**	**95**

Table 2: Overview of submissions to CLIR track.

An important contribution to the track was made by a collaboration between the University of Maryland and the LOGOS corporation, who provided a machine translation of German documents into English. Only the German SDA documents were prepared and translated in time for the submission deadline, but machine translation output of the NZZ document collection is now also available. This MT output was provided to all participants as a resource, and was used to support experiments run at ETH (Mateev et al., 1996), Duke University (Rehder et al., 1998), Cornell University (Buckley et al., 1998), Berkeley (Gey and Chen, 1998), and the University of Maryland (Oard and Hackett, 1998).

Cross-Language retrieval using dictionary resources was the approach taken in experiments submitted by groups at New Mexico State University (Davis and Ogden, 1998), University of Massachusetts (Allan et al., 1998), the Commissariat à l'Energie Atomique of France (Elkateb and Fluhr, 1998), the Xerox Research Centre Europe (Grefenstette et al., 1998), and at TNO in the Netherlands. Machine readable dictionaries were obtained from various sources, including the internet, for different combinations of languages, and used in different ways by the various groups.

Corpus-based approach to CLIR were evaluated by ETH, using similarity thesauri, and the collaborative group of Duke University, the University of Colorado, and Bellcore, who used latent semantic indexing (LSI). An innovative approach for cross-language retrieval between English and French was also tested at Cornell University. This approach was based on the assumption that there are many similar-looking words (near cognates) between English and French and that, with some simple matching rules, relevant documents could be found without a full translation of queries or documents.

The offer allowing groups to participate in a monolingual capacity using the document collections for French or German was accepted by groups at Dublin City University (Smeaton et al., 1998), University of Montreal (Nie and Chevallet, 1998), and IRIT France (Boughanem and Soulé-Dupuy, 1998).

An overview of results for each participating group are presented in Figure 1. This figure represents the results based on only 21 of the total 25 test queries. The remaining 4 queries have not yet been fully judged for relevant documents. The figure shows results for each group and each document language for which experiments were submitted. The y axis represents the average precision achieved over the set of 21 queries for the *best* experiment submitted by each group and each document language. Cross-language experiments are denoted by, for example, *'X to French'*, whereas the corresponding monolingual experiments are denoted, *'French'*. For example, the figure shows that the best experiment submitted by Cornell University performing cross-language retrieval of French documents achieved average precision of 0.2.

Note that the presentation of results in Figure 1 does not distinguish between fully automatic cross-language retrieval, and those groups who included some interactive aspect and user involvement in their experiments. The groups at Xerox, Berkeley and Dublin submitted experiments which involved manual interaction.

Although Figure 1 does not provide a sound basis for between-group comparisons, we can make some general comments on the overall results. Comparing cross-language results to the corresponding monolingual experiments, it seems that cross-language retrieval is performing in a range of roughly 50% to 75% of the equivalent monolingual case. This is consistent with previous evaluations of cross-language retrieval. Groups at ETH, Cornell, Xerox, TNO, Berkeley and Maryland have all achieved cross-language results in this range, compared to good levels of monolingual retrieval performance. While the LSI group, CEA and NMSU have also achieved good relative cross-language performance, their baseline monolingual retrieval performance is somewhat lower than the rest.

A slightly more detailed analysis of results can be achieved by plotting some

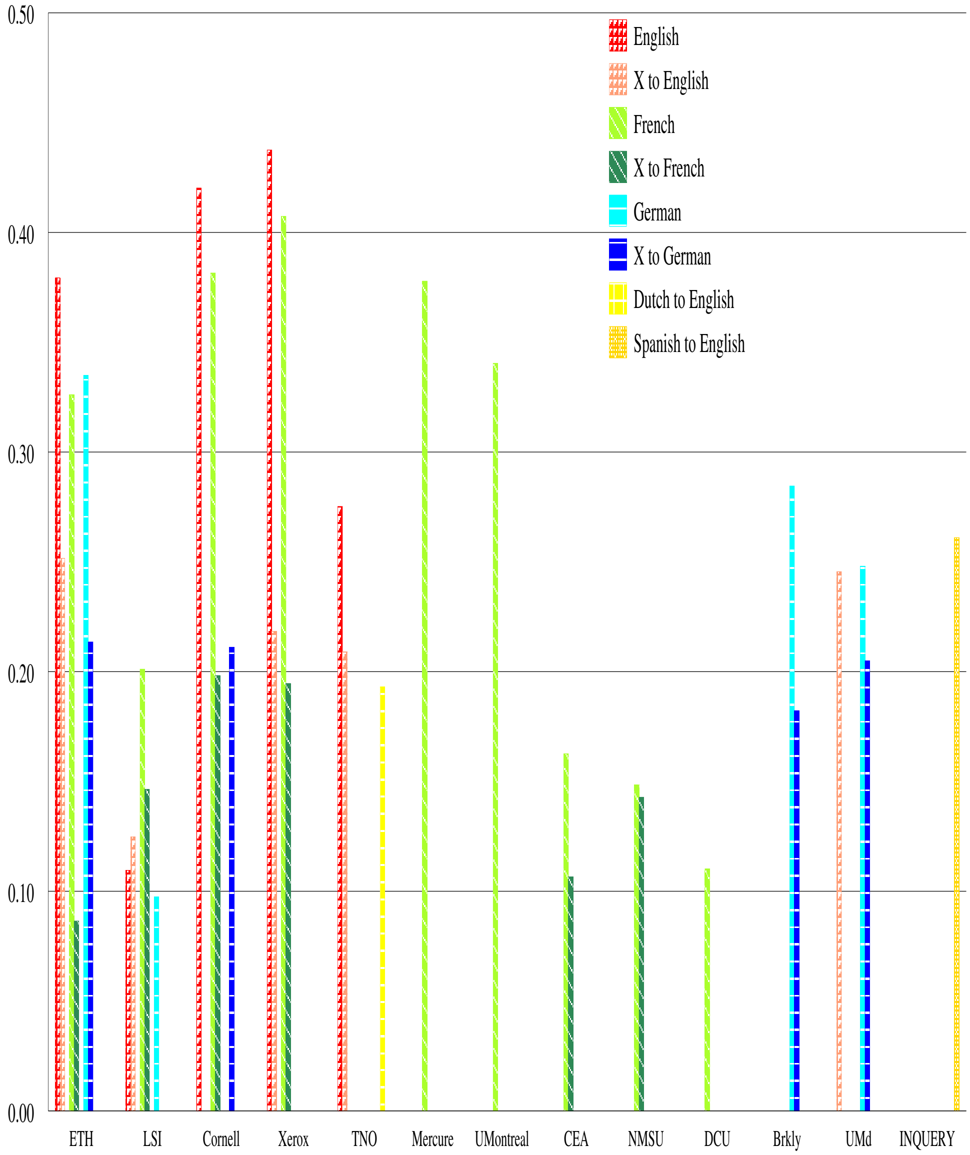

Figure 1: CLIR Track Results (Average Precision, best run)

of the experiments on standard recall/precision graphs, although the data for this is taken from an earlier point when only *thirteen* of the twenty five queries had been completely judged for relevance. We do not provide either a complete or detailed view of all the CLIR track results, but present in Figure 2 a selection of the results achieved for cross-language retrieval of French documents with English queries, and in Figure 3 some of the results of cross-language retrieval to German documents from English queries.

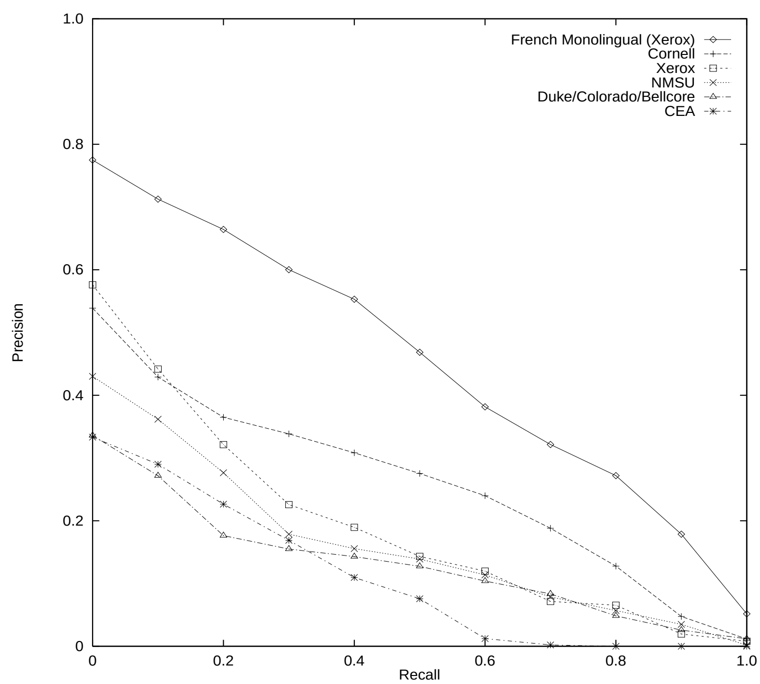

Figure 2: English-to-French CLIR: 13 Queries

The English-to-French CLIR experiments illustrated in Figure 2 include Cornell's cognate-matching approach, machine-readable dictionary approaches by Xerox, NMSU, and CEA, and the LSI run of the Duke/Colorado/Bellcore group which used corpus information derived from an alignment of the French SDA collection with English machine-translation output of the comparable German SDA collection. In each case we have presented the best automatic run submitted by these groups. The baseline for comparison is the monolingual French-French retrieval experiment submitted by the Xerox Research Centre Europe (Grenoble). The success of Cornell's approach is testament to the high degree of cognate overlap in English and French.

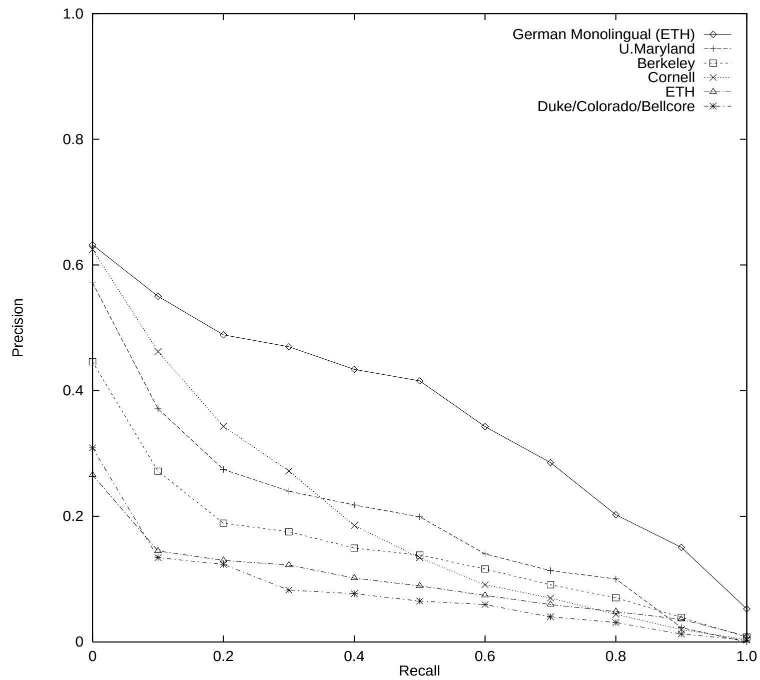

Figure 3: English-to-German CLIR: 13 Queries

Figure 3 includes English-to-German experiments using the LOGOS Machine Translation output (i.e. matching the English queries directly against the English MT output) from Maryland and Cornell, a dictionary-based experiment from Berkeley, and corpus-based approaches from ETH and the LSI group. The ETH experiment was based on a similarity thesaurus constructed using document *summaries* ('Title' and 'Lead' fields) of the German collection aligned with English MT output of the document summaries. The LSI approach was based on an alignment between documents and translations of the German SDA documents. The upper baseline in Figure 3 is provided by the monolingual German experiment run at ETH Zurich. Comparison of the cross-language results shows that for the 13 queries included, retrieval based on the MT output of the documents has done relatively well. The results also suggest that further investigation of the use of corpus material is necessary, though the corpus used in these experiments must be considered to be of poor quality with respect to the alignment of documents.

On the whole, the CLIR results are encouraging, especially given that this is the first year this track has been run. Ten groups submitted cross-language experiments. Many different approaches to cross-language retrieval have been tried and evaluated, and groups using each of the different approaches have achieved good results. As the track grows in the future, the challenge will be to find a concise way of presenting and comparing the performance of the different groups with different approaches running experiments with different language combinations.

4 Outlook

The outlook for the cross-language retrieval track is good. We believe that this track has shown already this year that cross-language retrieval is feasible. Groups now have a foundation on which to develop their approaches and a baseline against which to compare future performance. The existence now of a substantial test collection with documents in three languages and parallel queries in five languages, together with the relevance judgements provided by NIST, is an important asset to the CLIR community.

The TREC test collection is just one resource for CLIR however. It seems from the results presented in the previous section that the availability of resources, whether they be machine translation, machine-readable dictionaries, or high-quality corpora, is a key determinant of the level of cross-language retrieval performance that can be achieved. The acquisition of better dictionary and corpus resources is likely to be an important source of improved CLIR performance in the future.

The CLIR track has again illustrated some of the problems of automatically translating natural language text long known in the field of machine translation. The importance of translating the correct senses of words (e.g. *'logging'*) and of correctly translating multi-word units (e.g. *'fast food'*) as a single entity rather than word-by-word has been illustrated in several of the test queries. Groups which have used automatic MT, either for document translation or query translation, have demonstrated however that MT has an important contribution to make to cross-language information retrieval.

Plans for the TREC-7 CLIR track include the acquisition of more document data in each of German, French and English. There are no plans to add new document languages in the immediate future, although this has been discussed as a long-term goal of the track. The formulation of topic descriptions for TREC-7 CLIR will be spread out over three different sites, with native French, German and English speakers responsible for the formulation of the topics in each language. The three sites, at NIST, Bonn, Germany, and Lausanne, Switzerland, will then also be responsible for completing the relevance judgements of documents in the retrieved pool. The involvement of these sites in the process is aimed at ensuring the fluency of topic descriptions and the speed and reliability of relevance judgements for our future test collections.

Acknowledgements: We would like to express our appreciation to the Neue Zuercher Zeitung (NZZ), the Schweizerische Depeschen Agentur (SDA), and the Associated Press (AP) for making their data available to the TREC community. The authors would also like to thank NIST, on behalf of the CLIR track participants, for their work in converting this data into a valuable CLIR test collection.

References

Allan, J., Callan, J., Croft, W. B., Ballesteros, L., Byrd, D., Swan, R., and Xu, J. (1998). INQUERY Does Battle with TREC-6. In *Proceedings of the Sixth Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.*

Ballesteros, L. and Croft, W. B. (1996). Dictionary-based Methods for Cross-lingual Information Retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications.*

Ballesteros, L. and Croft, W. B. (1998). Statistical methods for cross-language information retrieval. In Grefenstette, G., editor, *Cross-Language Information Retrieval*, chapter 3. Kluwer Academic Publishers, Boston.

Boughanem, M. and Soulé-Dupuy, C. (1998). Mercure at Trec-6. In *Proceedings of the Sixth Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.*

Buckley, C., Mitra, M., Walz, J., and Cardie, C. (1998). Using Clustering and SuperConcepts Within SMART: TREC 6. In *Proceedings of the Sixth Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.*

Carbonell, J., Yang, Y., Frederking, R., Brown, R. D., Geng, Y., and Lee, D. (1997). Translingual information retrieval: A comparative evaluation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence.*

Davis, M. and Ogden, W. (1997). QUILT: Implementing a Large-Scale Cross-Language Text Retrieval System. In *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA*, pages 27–31.

Davis, M. and Ogden, W. (1998). Free Resources and Advanced Alignment for Cross-Language Text Retrieval. In *Proceedings of the Sixth Text Re-*

trieval Conference (TREC6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.

Davis, M. W. (1998). On the effective use of large parallel corpora in cross-language text retrieval. In Grefenstette, G., editor, *Cross-Language Information Retrieval*, chapter 2. Kluwer Academic Publishers, Boston.

Elkateb, F. and Fluhr, C. (1998). EMIR at the CLIR Track of TREC6. In *Proceedings of the Sixth Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.*

Gachot, D. A., Lange, E., and Yang, J. (1998). The SYSTRAN NLP browser: An application of machine translation technology in multilingual information retrieval. In Grefenstette, G., editor, *Cross-Language Information Retrieval*, chapter 9. Kluwer Academic Publishers, Boston.

Gey, F. and Chen, A. (1998). Phrase Discovery for English and Cross-Language Retrieval at TREC6. In *Proceedings of the Sixth Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.*

Grefenstette, G., Hull, D., Gaussier, E., and Schulze, B. (1998). Xerox TREC-6 Site Report: Cross-Language Text Retrieval. In *Proceedings of the Sixth Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.*

Hull, D. and Grefenstette, G. (1996). Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retreval. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland*, pages 49–57.

Landauer, T. and Littman, M. (1991). Fully Automatic Cross-Language Document Retrieval Using Latent Semantic Indexing. In *Proceedings of the 11th International Conference on Expert Systems and their Applications.*

Littman, M. L., Dumais, S., and Landauer, T. K. (1998). Automatic cross-language information retrieval using latent semantic indexing. In Grefenstette, G., editor, *Cross-Language Information Retrieval*, chapter 5. Kluwer Academic Publishers, Boston.

Mateev, B., Munteanu, E., Sheridan, P., Wechsler, M., and Schäuble, P. (1996). ETH TREC-6: Routing, Chinese, Cross-Language and Spoken Document Retrieval. In *Proceedings of the Sixth Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.*

Nie, J. and Chevallet, J. (1998). Using Terms or Words for French Information Retrieval? In *Proceedings of the Sixth Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.*

Oard, D. and Hackett, P. (1998). Document Translation for Cross-Language Text Retrieval at the University of Maryland. In *Proceedings of the Sixth Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.*

Peters, C. and Picchi, E. (1997). Using Linguistic Tools and Resources in Cross-Language Retrieval. In *Proceedings of the 3rd DELOS workshop; Cross-Language Information Retrieval, ERCIM Workshop Proceedings No. 97-W003, (ISBN: 2-912335-02-7), Zurich, Switzerland,*, pages 75–84.

Rehder, B., Littman, M., Dumais, S., and Landauer, T. (1998). Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing. In *Proceedings of the Sixth Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.*

Sheridan, P. and Ballerini, J. P. (1996). Experiments in Multilingual Information Retrieval using the SPIDER System. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland*, pages 58–65.

Sheridan, P., Braschler, M., and 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, Pisa, Italy*, pages 253–268.

Sheridan, P. and Schäuble, P. (1997). Cross-Language Multi-Media Information Retrieval. In *Proceedings of the 3rd DELOS workshop; Cross-Language Information Retrieval, ERCIM Workshop Proceedings No. 97-W003, (ISBN: 2-912335-02-7), Zurich, Switzerland,*, pages 65–74.

Smeaton, A., Kelledy, F., and Quinn, G. (1998). Ad hoc Retrieval Using Thresholds, WSTs for French Monolingual Retrieval, Document-at-a-Glance for High Precision and Triphone Windows for Spoken Documents. In *Proceedings of the Sixth Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.*

Yamabana, K., Muraki, K., Doi, S., and Kamei, S. (1998). A language conversion front-end for cross-language information retrieval. In Grefenstette, G., editor, *Cross-Language Information Retrieval*, chapter 8. Kluwer Academic Publishers, Boston.