Cross-Language Information Retrieval (CLIR) Track Overview

Martin Braschler[1], Jürgen Krause[2], Carol Peters[3], Peter Schäuble[4]

[1] Eurospider Information Tech. AG, Schaffhauserstr. 18, CH-8006 Zürich, Switzerland
[2] Informationszentrum Sozialwissenschaften, Lennéstr. 30, D-53113 Bonn, Germany
[3] Istituto di Elaborazione della Informazione (IEI-CNR), Via S. Maria 46, Pisa, Italy
[4] Swiss Federal Institute of Technology (ETH), CH-8092 Zürich, Switzerland

1 Introduction

This year, the TREC cross-language retrieval track took place for the second time. In TREC-7, we extended the task presented to the participants. The goal was for groups to use queries written in a single language in order to retrieve documents from a multilingual pool of documents written in many different languages. This is also a more comprehensive task than the usual definition of cross-language information retrieval, where systems work with a language pair, retrieving documents in a language L1 using queries in language L2.

The document languages used this year were English, German, French, and, newly introduced for TREC-7, Italian. The queries were available in all of these languages. Because it seemed unlikely that all interested parties can work with all four languages, it was agreed that there would be a secondary evaluation involving a smaller task. Consequently, groups were allowed to send in runs using the English queries to retrieve documents from a subset of the pool containing just the English and French documents. Coordination of the track took place at ETH in Zurich, as for last year.

The continued interest in the cross-language track showed the importance of this emerging area. There are many applications where information should be accessible to users regardless of its language. With the ever growing amount of information available to us all, situations when a user of an information retrieval system is faced with the task of querying a multilingual document collection are becoming increasingly common. Such collections can be made up of documents from multinational companies, from multilingual countries or from large international organizations such as the United Nations or the European Commission. Of course, the world wide web is also an example for such a document collection.

A lot of users of such multilingual data sources have some foreign language knowledge, but their proficiency may not be good enough to formulate queries to appropriately express their information need. Such users will benefit greatly if they can enter queries in their native language, because they will be able to inspect the documents even if they are untranslated. Monolingual users, on the other hand, can use translation aids, manual or automatic, to help them access the search results.

2 Overview of CLIR

Approaches to cross-language information retrieval can be categorized according to how they solve the problem of matching the query and documents across different languages – how they "*cross the language barrier*". This may be achieved by using query translation, document translation, or by using both query and document translation (see also Figure 1). One possibility would be to translate queries and documents into a controlled, language independent indexing vocabulary. TextWise (Diekema et al., 1999) uses WordNet synsets as such a vocabulary. More common in the framework of TREC are free text approaches. These methods can be further classified according to what resources are used to cross the language boundary: machine translation, machine-readable dictionaries, or corpus-based resources.

Machine translation (MT) seems an obvious choice for cross-language information retrieval systems. Groups that experimented with machine translation for CLIR include NEC in Japan (Yamabana et al., 1998). They used MT to translate users' queries in an interactive process involving both dictionaries and statistical information derived from bilingual corpora. Systran is also reporting work addressing cross-language retrieval (Gachot et al., 1998). In the context of this year's track, the Berkeley group carried out experiments with three different off-the-shelf MT systems (Gey et al., 1999). Note that CLIR is a difficult problem to solve based on MT alone: queries that users typically enter into a retrieval system are rarely complete sentences and provide little context for sense disambiguation.

Corpus-based approaches to CLIR include, among others, the use of Latent Semantic Indexing (LSI) by researchers at Bellcore and elsewhere (Littman et al., 1998), the Generalized Vector Space model proposed by CMU (Carbonell et al., 1997) and work from ETH/Eurospider using Similarity Thesauri and Pseudo-relevance feedback (Braschler and Schäuble, 1998). All these approaches use corpus resources as training data to adapt the CLIR mechanism or build information structures used for subsequent retrieval.

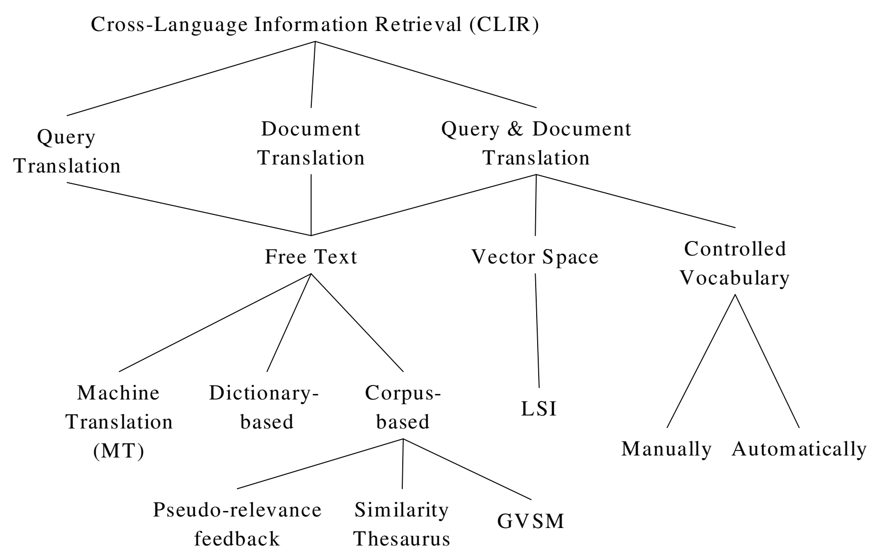

Figure 1: Approaches to cross-language information retrieval

Another natural approach to cross-language retrieval is the use of existing linguistic resources, mainly machine-readable bilingual dictionaries. Among groups that looked into the use of such dictionaries are researchers from Xerox Research Centre Europe (Hull and Grefenstette, 1996), the University of Massachusetts (Ballesteros and Croft, 1996) and the Computing Research Laboratory at the New Mexico State University (NMSU) (Davis and

Ogden, 1997). Various ideas have been proposed to address some of the problems associated with dictionary-based translations, such as ambiguities and vocabulary coverage.

Approaches from all three main classes have been used by the participants in the TREC-7 CLIR track.

3 CLIR-Track Task Description

This year, the participants were asked to retrieve documents from a multilingual pool containing documents in four different languages. They were able to chose the topic language, and then had to find relevant documents in the pool regardless of the languages the texts were formulated in. As a side effect, this meant that most groups had to solve the additional task of merging results from various bilingual runs.

The languages present in the pool were English, German, French and Italian, with Italian being a new language introduced for TREC-7. The 28 topics were distributed in each of these four languages. To allow for participation of groups that do not have the resources to work in all four languages, a secondary evaluation was provided that permitted such groups to send in runs using English queries to retrieve documents from a subset of the pool just containing texts in English and French. This year, we did not have monolingual runs as part of the cross-language track.

The TREC-7 task description also defined a subtask, working with a second data collection, containing documents from a structured data base from the field of social science. Unfortunately, the introduction of this data was probably premature, since no groups were working with this data this year. The data will however be used again for next year's CLIR track.

The document collection for the main task contained the same documents as used for TREC-6. The English texts were taken from the Associated Press, covering three years (1988 to 1990) worth of news stories. For German and French, news stories were taken from SDA, the "Schweizerische Depeschenagentur" (Swiss News Agency). They were chosen to cover the same time period. While the texts for German and French were produced by the same company, this does not mean that they contain translations. However, there is a sizeable topic overlap between the texts in these two languages. For German, additionally texts from the Swiss news paper "Neue Zürcher Zeitung" (NZZ) were used. We had one year of articles (from 1994) available to participants. As an extension to this document collection, Italian texts from SDA were introduced in TREC-7. Again, while produced by the same company as the French and some of the German texts, they are not direct translations from either of these languages. We had texts from 1989 and 1990 available in Italian. Table 1 gives more details of the document collections.

There have been significant changes in the way the topics were created for this year. The experience of the first CLIR track showed that it is difficult to produce topics in all languages in a single place. Therefore, a distributed approach to topic creation was chosen. We had four different sites, each located in an area where one of the topic languages is natively spoken.

Document collections			
Doc. Language	**Source**	**No. Documents**	**Size**
English	AP news, 1988-90	242,918	750 MB
German	SDA news, 1988-90 NZZ articles, 1994	185,099 66,741	330 MB 200 MB
French	SDA news, 1988-90	141,656	250 MB
Italian	SDA news, 1989-90	62,359	90 MB

Table 1: details for the document collections.

The topic creation sites were:

- English: NIST, Gaithersburg, MD, USA (Ellen Voorhees)
- French: EPFL Lausanne, Switzerland (Afzal Ballim)
- German: IZ Sozialwissenschaften, Germany (Jürgen Krause, Michael Kluck)
- Italian: CNR, Pisa, Italy (Carol Peters).

From each site seven topics were chosen to be included in the topic set. The other 21 queries were then translated. This ultimately led to a pool of 28 topics, each available in all four languages.

Participants were free to experiment with different topic fields, and with both automatic and manual runs, similar to the definitions of the TREC adhoc task.

4 Results

A total of nine groups from five different countries submitted results for the TREC-7 CLIR track (see Table 2). Because of the different task this year, the number of runs per group could be reduced, since the number of language combinations is much smaller with the document pool being fixed. The participants submitted 27 runs, 17 for the main task, and 10 for the secondary evaluation. Five groups (Berkeley, Eurospider, IBM, Twenty-One and Maryland) tackled the main task. English was, not surprisingly, the most popular topic language, with German coming in a strong second. Every language was used by at least one group.

Participant	Country
CEA (Comissariat à l'Energie Atomique)	France
Eurospider Information Technology AG	Switzerland
IBM T.J. Watson Research Center	USA
Los Alamos National Laboratory	USA
TextWise LLC	USA
Twenty-One (University of Twente/TNO-TPD)	Netherlands
University of California at Berkeley	USA
University of Maryland	USA
Université de Montréal	Canada

Table 2: Table of participants.

The relevance assessments used for evaluation of these runs were produced at the same four sites that were used for topic creation.

Remarkably, average precision numbers are generally much higher than last year. While it is hard to compare absolute levels across topic sets, this would be an indication that the level of the results has improved this year. Unfortunately, there is little mention by participants about experiments on last year's topics with the current systems that could substantiate such an assumption. Also, the Twenty-One group (Hiemstra et al., 1999) makes a case as to why absolute levels may be too high ("flattering") this year.

Figure 2 shows a comparison of runs for the main task. Shown are the best automatic runs against the full document pool for each of the five groups that solved the full task. As can be seen, most participants performed in a fairly narrow band. This is interesting given the very different approaches of the individual participants: IBM used translation models automatically trained on parallel and comparable corpora (McCarley, 1999). Twenty-One used sophisticated dictionary lookup and a "boolean-flavoured" weighting scheme (Hiemstra et al., 1999). Eurospider employed corpus-based techniques, using similarity thesauri and pseudo-relevance feedback on aligned documents (Braschler et al., 1999). The Berkeley (Gey et al., 1999) and Maryland groups used off-the-shelf machine translation systems.

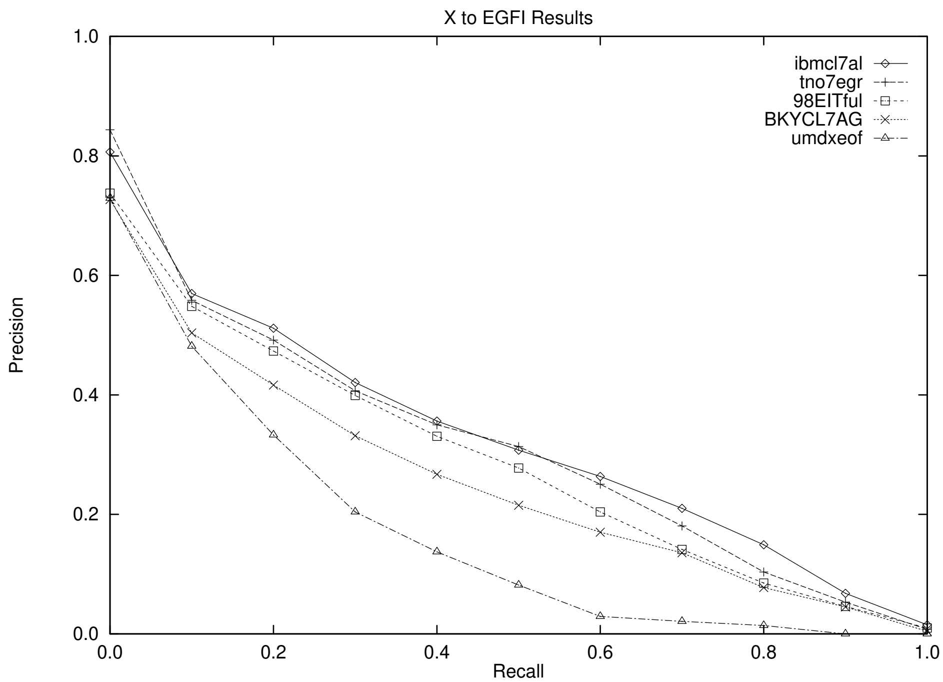

Figure 2: Results of the main evaluation X→EGFI.

Figure 3 shows results from the secondary evaluation. Again, the graph shows the best automatic run submitted by each participant. The run of the Los Alamos group is an exception, as it is classified as a manual run. Here too, the top three runs are quite close. IBM was again using their statistical translation models, Twenty-One was using dictionary-based translation with fuzzy query expansion terms and Berkeley was again using their MT approach. Following are four groups only participating in the secondary evaluation: Université de Montréal, CEA, TextWise and Los Alamos.

A particularly interesting aspect of this year's track was how participants approached the merging problem. Again, many interesting methods were used: Among the solutions proposed were: Twenty-One compared averages of similarity values of individual runs, Eurospider used document alignments to map runs to comparable score ranges through linear regression and IBM used modeling of system-wide probabilities of relevance. But it was also possible to avoid the merging problem: the Berkeley group expanded the topics to all languages and then ran them against an index containing documents from all languages, therefore directly retrieving a multilingual result list.

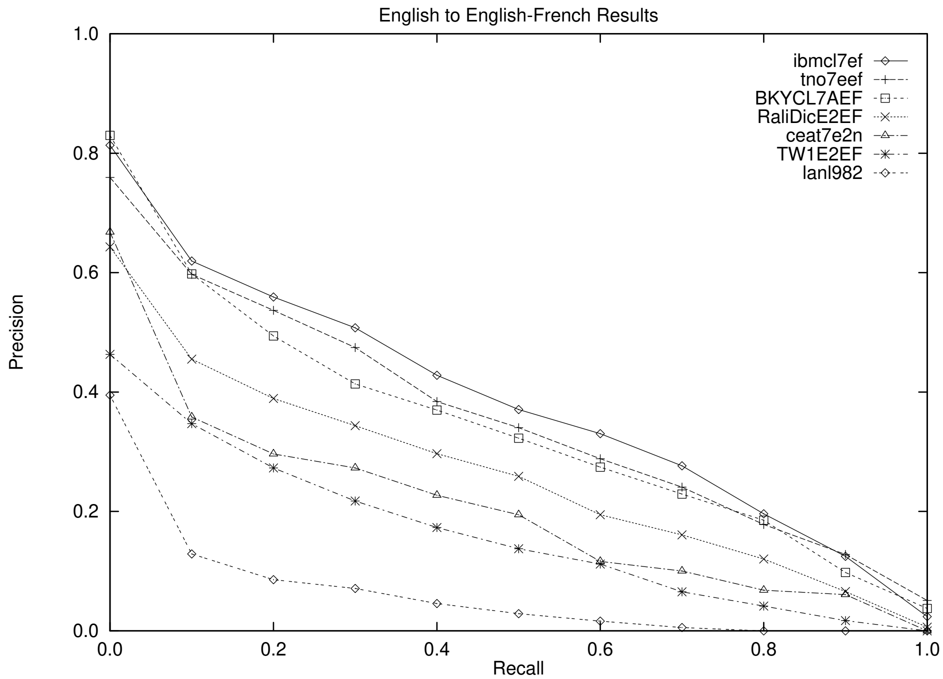

Figure 3: Results of the secondary evaluation E→EF.

5 Evaluation issues

As mentioned, one of the distinguishing features of the CLIR track is that topic development is distributed. Topic development is clearly subjective, and tends to depend on the creator's own particular background. However, for CLIR it is presumed that both the language and cultural background also impact on the choice and phrasing of topics. A close examination of this year's topics would probably permit an astute observer to group them fairly accurately, according to source language and creation site. This should not be considered negative in the participants' viewpoint nor should it affect the validity of the exercise. However, it causes some problems both with reference to translation of the topics and their assessment.

Since it is unrealistic to find topic creators that have total competence in all four languages, each topic is developed in one language and then translated at the other sites. Topic translation thus raises the typical problems involved in any translation: a total understanding of the source in order to achieve a perfect rendering of the target. The conflict is as to how far the target version can deviate from the source in terms of style, vocabulary, and authenticity. It is necessary to find an acceptable balance between precision with respect to the source and naturalness with respect to the target language. While preserving the topic meaning, terms must be used in the target topic that are actually found in the documents of that language. Thus, a high level of performance is required of the topic translators to avoid an imbalance in topic authenticity.

The relevance assessments were also produced in the same distributed setting. Of course, an accurate assessment of relevance for retrieved documents for a given topic implies a good understanding of the topic. In the distributed scenario of the CLIR track, understanding is also influenced by the multilingual/multicultural characteristics of the task. Although the topic creators initially worked on the basis of their knowledge of possible events for the years covered by the document collections, the final decision and refinement with regard to the topics was made based on the contents of the document collection. The way a particular argument is presented in a collection therefore will influence its formulation. However, this presentation is not necessarily reproduced in the documents in other languages. Thus a topic which did not appear to raise problems of interpretation in the language used for its preparation, may be much more difficult to assess against documents in another language. Some of the topics were found by the assessors to be too vague or difficult to interpret, while others required very specific knowledge. The decision for each local topic developer to include one or two topics of high local significance also caused some difficulties. Some more political arguments, that were well known and much discussed in the local document collection but still had wider implications were difficult to understand and recognize in other collections.

Obviously, it tends to be easier to assess the results for a collection in the original language used for a topic. This fact needs further investigation to assess its real effect (if any) on the overall results.

6 Outlook

The CLIR track will return next year. It was agreed to keep the main task; retrieving documents in many different languages. There also will be a secondary evaluation, retrieving documents from a pool of English documents and one additional chosen language. The GIRT subtask will be offered again next year, and will also allow to send in monolingual runs, something that is not planned for the other evaluations.

The evaluation issues mentioned above mean that there will be emphasis on clear rules for translation, and topics will be circulated to check for problems of interpretation. "Difficult" topics will be possibly accompanied by interpretation aids or training of assessors.

Acknowledgements

We thank the Neue Zürcher Zeitung (NZZ), the Schweizerische Depeschenagentur (SDA) and the Asssociated Press (AP) for making their data available to the TREC community. We would also like to express our gratitude to everyone involved in topic creation and relevance assessment at NIST, the IZ Sozialwissenschaften, CNR and the EPFL.

References

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

Braschler, M. and Schäuble, P. (1998). Multilingual Information Retrieval Based on Document Alignment Techniques. In *Second European Conference on Research and Advanced Technology for Digital Libraries, Heraklion, Greece*.

Braschler, M., Mateev, B., Mittendorf, E., Schäuble, P., and Wechsler, M. (1999). SPIDER Retrieval System at TREC7. In *Proceedings of the Seventh Text Retrieval Conference (TREC7), 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*.

Diekema, A., Oroumchian, F., Sheridan, P., and Liddy, E. D. (1999). TREC-7 Evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French. In *Proceedings of the Seventh Text Retrieval Conference (TREC7), 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. C., Jiang, H., and Chen, A. (1999). Manual queries and Machine Translation in Cross-Language Retrieval at TREC-7. In *Proceedings of the Seventh Text Retrieval Conference (TREC7), National Institute of Standards and Technology (NIST), Gaithersburg, MD*.

Hiemstra, D. and Kraaij, W. (1999). TREC-7 working notes: Twenty-One in ad-hoc and CLIR. In *Proceedings of the Seventh Text Retrieval Conference (TREC7), 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 Retrieval. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland*.

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.

McCarley, J. S. (1999). Multilingual Information Retrieval at IBM. In *Proceedings of the Seventh Text Retrieval Conference (TREC7), National Institute of Standards and Technology (NIST), Gaithersburg, MD*.

Schäuble, P. and Sheridan, P. (1998). Cross-Language Information Retrieval (CLIR) Track Overview. 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.