Cross-Language Information Retrieval (CLIR) Track Overview

Martin Braschler[1], Carol Peters[2], Peter Schäuble[1]

[1] Eurospider Information Tech. AG, Schaffhauserstr. 18, CH-8006 Zürich, Switzerland
[2] Istituto Elaborazione Informazione (CNR), Via Alfieri 1, 56010 Ghezzano, Pisa, Italy

1 Introduction

A cross-language retrieval track was offered for the third time at TREC-8. The main task was the same as that of the previous year: 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. Compared to the usual definition of cross-language information retrieval, where systems work with a single language pair, retrieving documents in a language L1 using queries in language L2, this is a slightly more comprehensive task, and we feel one that more closely meets the demands of real world applications.

The document languages used were the same as for TREC-7: English, German, French and Italian. The queries were available in all of these languages. Monolingual non-English retrieval was offered to new participants who preferred to begin with an easier task. However, all the groups which did not tackle the full task opted for limited cross-language rather than monolingual runs. These experiments were evaluated by NIST and are published as unofficial ("alternate") runs. We also offered a subtask, working with documents from the field of social sciences. This collection (known as "GIRT") has some very interesting features, such as controlled vocabulary terms, title translations, and an associated multilingual thesaurus.

The track was coordinated at Eurospider Information Technology AG in Zurich. Due to its multilingual nature, the topic creation and relevance assessment tasks were distributed over four sites in different countries: NIST (English), IZ Bonn (German), IEI-CNR (Italian) and University of Zurich (French). The University of Hildesheim invested considerable effort into rendering the topics homogeneous and consistent over languages.

The participating groups experimented with a wide variety of strategies, ranging machine translation, corpus-, and dictionary-based approaches. Some results are given in Section 4. There were, however, also some striking similarities between many of the runs, such as the choice of English as topic language the majority, and the use of Systran by a lot of groups. Some implications of these findings are discussed in Section 5.

The main goal of the TREC CLIR activities has been the creation of a multilingual test collection that is re-usable for a wide range of evaluation experiments. This means that the quality of the relevance assessments is very important. The Twenty-One group conducted an interesting analysis with respect to the completeness of the assessments and the impact of this on the pool. We address some of their findings in Section 5.

The paper concludes with an indication of our plans for the future of the cross-language track, which will bring substantial changes to the format and coordination of the activities.

2 Overview of CLIR

There are three main ways in which cross-language information retrieval approaches attempt to "cross the language barrier" – through query translation, or document translation, or both. (Oard, 1997). CLIR research started out with experiments using controlled vocabularies and associated dictionaries and thesauri, but nowadays free text approaches are most common. These approaches also dominate experiments in past and present CLIR tracks. Free text methods can be further classified according to the resources 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. It also played a large role in the TREC-8 experiments of a number of groups. However, CLIR is a difficult problem to solve on the basis of 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 are also popular. Groups experimenting with such approaches during this or former CLIR tracks include Eurospider, IBM and the University of Montreal.

Lastly, a significant number of cross-language retrieval approaches make use of existing linguistic resources, mainly machine-readable bilingual dictionaries. Various ideas have been proposed to address some of the problems associated with dictionary-based translations, such as ambiguities and vocabulary coverage. One of the groups that have investigated the use of such dictionaries is the Twenty-One consortium.

3 CLIR-Track Task Description

Similarly to last year, CLIR track participants were asked to retrieve documents from a multilingual pool containing documents in four different languages. They were free to choose the topic language, and then had to find relevant documents in the pool regardless of the languages in which the texts were formulated. Most groups approached this task by performing separate bilingual retrieval runs, and then combining the results. The merging of their retrieval results was therefore an additional problem for these groups.

Documents for TREC-8 were in English, German, French and Italian. There were 28 topics, each one provided in all four languages. In order to attract newcomers, monolingual non-English runs were accepted; however, participants preferred to do bilingual cross-language runs when they could not do the full task.

The TREC-8 task description also included a vertical domain subtask, working with a second data collection, containing documents from a structured database in the field of social science (the "GIRT" collection). This collection comes with English titles for most documents, and a matching bilingual thesaurus. The University of Berkeley conducted some very extensive experiments with this collection.

The document collection for the main task contained mainly news-wire articles. The English texts were taken from three years (1988 to 1990) of Associated Press news stories. For German, French and Italian, news stories were taken from SDA, the "Schweizerische Depeschenagentur" (Swiss News Agency), covering the same time period. While these texts were produced by the same agency, this does not mean that they contain actual translations. However, there is a sizeable topic overlap between the texts in the three languages, enabling experiments with alignment on these collections (for example experiments by Eurospider and IBM). For German, texts from the Swiss newspaper "Neue Zürcher Zeitung" (NZZ) for 1994 were also added. Table 1 gives more details on the document collections.

Document collections			
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: figures for the document collections.

For TREC-6, the CLIR track topics were developed centrally at NIST (Schäuble and Sheridan, 1998). However, problems during the topic creation and relevance assessment process and reactions from participants showed that this was not an optimal solution. A good translation has to take regional and cultural differences into account, and this is very hard to achieve if there is just one topic creation site. Consequently, in TREC-7, a distributed topic creation and relevance assessment setup was introduced (Braschler et al., 1999). This made it much easier to use native speakers in the translation stage which helped to improve overall quality. However, spreading this process over several sites means increased coordination overheads. The danger of producing inconsistent translations was addressed by active communication between the sites through e-mail and meetings. We retained this distributed setup for TREC-8. In addition, we received valuable help from University of Hildesheim in ensuring the consistency and quality of the topics.

The topic creation and results assessment sites for TREC-8 were:

- English: NIST, Gaithersburg, MD, USA (Ellen Voorhees)
- French: University of Zurich, Switzerland (Michael Hess)
- German: IZ Sozialwissenschaften, Germany (Jürgen Krause, Michael Kluck)
- Italian: IEI-CNR, Pisa, Italy (Carol Peters)

At each site, an initial 10 topics were formulated. At a topic selection meeting, the seven topics from each site that were felt to be best suited for the multilingual retrieval setting were then selected. Each site then translated the 21 topics formulated by the others into the local language. This ultimately led to a pool of 28 topics, each available in all four languages. It was decided that roughly one third of the topics should address national/regional, European and international issues, respectively. To ensure that topics were not too broad or too narrow and were easily interpretable against all document collections, monolingual test searches were conducted.

Participants were free to experiment with different topic fields (using either the title, description or narrative – or all three), and with both automatic and manual runs, similar to the definitions of the TREC adhoc task.

4 Results

A total of twelve groups from six different countries submitted results for the TREC-8 CLIR track (see Table 2). Eight participants tackled the full task (up from last year's five), submitting 27 runs (up from 17). The remainder of the participants either submitted runs using a subset of languages, or concentrated on the GIRT subtask only. English was the dominant topic language, even more so than last year. This development was not anticipated in such a pronounced form. Still, each language was used by at least one group as the topic language.

Participant	**Country**
Claritech	USA
Eurospider Information Technology AG	Switzerland
IBM	USA
IRIT/SIG	France
Johns Hopkins University APL	USA
MNIS-Textwise Labs	USA
New Mexico State University	USA
Sharp Laboratories of Europe Ltd	UK
Twenty-One	Netherlands
University of California, Berkeley	USA
University of Maryland	USA
University of Montreal	Canada

Table 2: Distribution of participants.

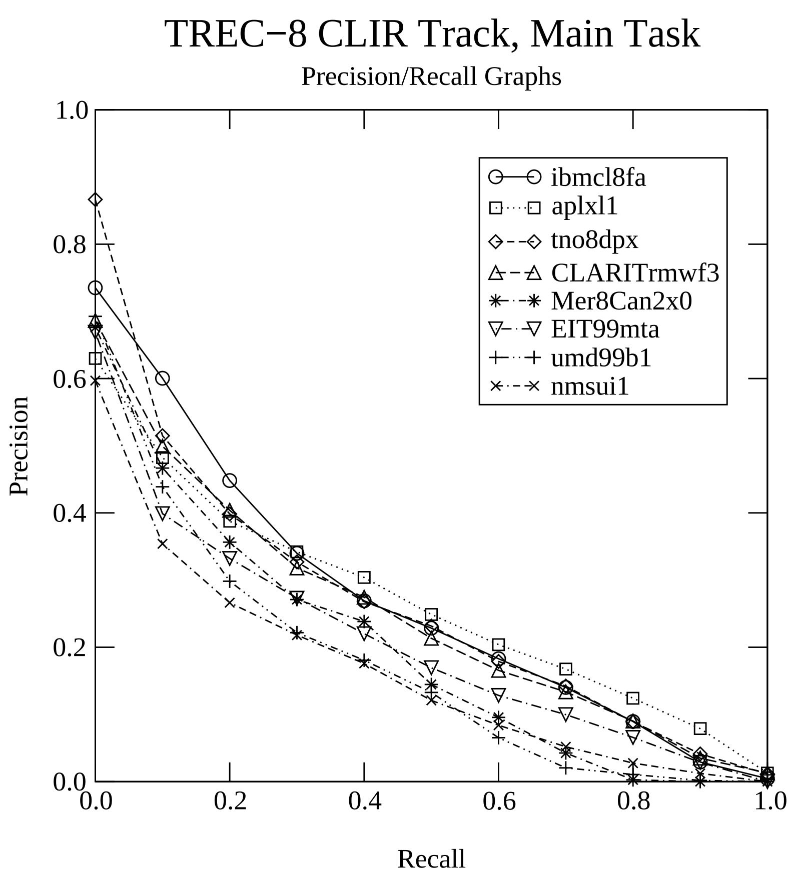

Figure 1: Runs for the main task

The relevance assessments used for the evaluation of these runs were performed by the same four sites listed above.

While the average precision numbers improved in TREC-7 with respect to TREC-6, they fell slightly in TREC-8; this is perhaps due to having a smaller average number of relevant documents per topic.

Figure 1 shows a comparison of runs for the main task. The graph shows the best automatic runs against the full document pool for each of the eight groups. Because of the diversity of the experiments conducted, the figures are best compared on the basis of the specific features of the individual runs. These can be found in the track papers. For example, New Mexico State runs use manually translated queries, which are the result of a monolingual user interactively picking good terms. This is clearly an experiment that is very different from the runs of some other groups that are essentially doing "ad-hoc" style cross-language retrieval, using no manual intervention whatever.

Approaches employed in TREC-8 by individual groups include:

- experiments on pseudo relevance feedback by Claritech (Qu et al., 2000)
- similarity thesaurus based translation by Eurospider (Braschler et al., 2000)
- statistical machine translation by IBM (Franz et al., 2000)
- combinations of n-grams and words by JHU (Mayfield et al., 2000)
- use of conceptual interlingua by Textwise (Ruiz et al., 2000)
- query translation using bilingual dictionaries by Twenty-One (Kraaij et al., 2000)
- evaluation of the Pirkola measure by University of Maryland (Oard et al., 2000)
- transaction models derived from parallel text by University of Montreal (Nie, 2000)
- use of an online machine translation system by Mercure/IRIT (Boughanem et al., 2000)

This diversity of approaches is one of the characteristics that makes the CLIR track extremely interesting and shows that there is still a lot of room for further studies and development.

Merging remained an important issue for most participants. University of Maryland tried to circumvent the problem by using an unified index in some of their runs, but the other groups working on the main task all had to rely on merging of some sort to combine their individual, bilingual cross-language runs. Some of the approaches this year include: merging based on probabilities - calculated using log(Rank) by various groups including IBM, merging using linear regression on document alignments by Eurospider, linear combinations of scores by JHU, and of course, straight, score-based merging.

Two groups submitted runs for the GIRT subtask. Berkeley even participated exclusively in the subtask only, and did some very comprehensive experiments using both the English titles of the documents and the English/German thesaurus supplied with the collection (Gey and Jiang, 2000). These runs show some of the interesting properties of GIRT, and we hope that this subtask will have more participants in the future.

It is also possible to do ad-hoc style runs on GIRT, ignoring controlled vocabulary, English titles and the thesaurus. This approach was taken by Eurospider.

5 Observations and Trends

It is interesting to note certain similarities between the submissions of a number of participants this year. Two main points stand out with respect to the main task: first, 21 out of

27 submitted runs used English as the topic language, and second, that at least half of all groups used the Systran machine translation system in some form for parts of their experiments.

Although it is not surprising that English is a popular choice as topic language, we did not expect this language to be so dominant. While English was also the most popular choice for TREC-7, the percentage of runs that used non-English topics was substantially higher (7 out of 17). We had hoped that with the CLIR track in its third year, more groups would start to experiment with non-English query languages. That this has not been not the case could be due to several factors. The fact that three quarters of the participants are located in English speaking countries certainly plays an important role. If we can encourage more European groups to participate in this activity, the ratio should become more balanced.

However, we believe it is also a result of a lack of resources available to some of the groups. The coordinators have always been aware that the main task of handling four languages may appear daunting to newcomers. In the past, we attempted to lessen the "shock" by allowing either cross-language runs on subsets of languages, or monolingual non-English runs. The intention was to allow groups that did not have access to resources for all languages, or were lacking experience in handling some of the languages, to start slowly and then expand their participation in the future.

While it is encouraging to see that most groups did try to tackle the main task, the fact that the majority of them chose English as their topic language may indicate that they are still constrained in the kind of resources available to them. They may have found dictionaries for English and the other languages, but not for e.g. German to Italian. The resource problem therefore seems to remain as a stumbling block. In the future, we hope to invest some efforts into building a repository for such resources that will allow participants to share whatever free components they have available. Together with the continued offer to start with easy tasks, this should also contribute to encouraging new groups to participate in cross-language system evaluation activities.

Similarly, we feel that part of the reason for the choice of Systran by so many groups also lies in a lack of resources: using Systran allowed the groups to do at least something with certain language pairs that they would otherwise not have been able to include in their experiments. That Systran offers mainly combinations of English with other languages probably also contributed to the domination of English as topic language.

Another area that merits attention this year is that of the relevance assessments. The Twenty-One group made an interesting analysis of the TREC-7 pool of relevance judgments. The quality of the pool and the judgments was also a topic of discussion on the mailing list leading up to the TREC-8 conference. The literature reports a considerable number of interesting experiments aimed at testing the quality and the properties of relevance assessments. The work by Voohees (Voorhees, 1998) is particularly notable. Working with the relevance assessments of the TREC-4 and TREC-6 ad-hoc task, Voorhees found that the relative effectiveness of different retrieval strategies remains stable despite marked differences in the relevance judgments used to measure retrieval. This means that while the actual values of the effectiveness measure (i.e. average precision) are affected by differences in relevance judgements, the relative retrieval performance remains almost always constant. While the analysis by the Twenty-One group was concerned with a slightly different question, namely if the size of the pool is sufficient, we felt it would be interesting to spot-check the hypothesis that the ordering remain mostly stable even when the values of the relevance judgments are altered. In fact, we found that, on the basis of the numbers given by Twenty-One in their paper, the ranking of the systems would probably have remained nearly identical, even if individual runs were not judged. Since the runs that were analyzed by Twenty-One are a mix of multilingual and bilingual experiments, and since it was not possible to re-run all the experiments in time for this paper, unfortunately, we cannot give exact figures. However, the only two runs that seem to have any real potential for changing ranks are the RaliDicAPf2e and ceat7f2 runs. As can seen from the numbers given in the Twenty-One paper, these are the two runs that provide the most unique relevant documents. They are also very close to some other runs in their absolute values. These two factors combine to increase the probability of a change in ranking. Note also that for the three groups that had multiple runs judged (Berkeley, Eurospider and Twenty-One), the ordering of the runs does not change in any case. This is consistent with the findings of Voorhees for the TREC-style relevance judgments analyzed in her paper, where she states that comparing algorithmic variants of the same system is very reliable.

Constantly questioning the relevance assessments and analyzing their quality remains very important when the goal is to create a reliable test suite for cross-language system evaluation. Most research on the topic is encouraging, and the considerations outlined above that indicate a stable ranking seems to imply that such findings are also valid in the case of the cross-language pool. We have to remain vigilant with respect to the quality of that pool since, as the Twenty-One group points out, it is still rather small. We are however confident that participants receive valuable results from their evaluation through the CLIR track. It is certainly true that non-participants might have more difficulties in interpreting their results based on the small size of the CLIR pool, as Twenty-One points out. We hope, however, that this will encourage these people to participate in the future, thus increasing the size of the pool. This is the best way to improve the pool.

6 Move to Europe and CLEF

From 2000 on, it has been decided to coordinate cross-language system evaluation for what are traditionally considered as European languages in Europe rather than in the U.S, although still in collaboration with NIST and TREC. The European side is sponsored by the DELOS Network of Excellence for Digital Libraries and funded by the European Commission.

There are several reasons that have lead to this decision. Perhaps the main one is that, as already mentioned, much of the work was already being done in Europe. However, moving the coordination to Europe not only makes logistic sense but also leaves NIST freer to concentrate on cross-language evaluation on other language groups. In fact, in 2000, TREC will be offering a cross-language track using English and Mandarin documents and English topics. Depending on data availability, the track may also involve Tamil and Malay documents.

More importantly, this move and the launching of an independent activity – known as CLEF (for Cross-Language Evaluation Forum) - allows us to focus on a wider range of issues. As has been stated, the main task offered in TREC-7 and 8 - the multilingual retrieval task - was a hard task and possibly discouraged some potential participants who did not have the resources (or the confidence) to tackle cross-language retrieval with all four languages. Thus, we have decided to provide a greater variety of tasks in CLEF 2000. The aim is both to encourage the participation of groups who are only now beginning to tackle the issues involved in cross-language retrieval, and also to extend the possibility of participation to groups developing systems for other European languages.

There will thus be three main evaluation tasks in CLEF 2000: multilingual information retrieval, bilingual information retrieval, and monolingual (non-English) retrieval, plus again the GIRT sub-task for cross-language retrieval in a special domain. Interested groups can participate in any one or in all four tracks.

Similarly to TREC-8, the main task of CLEF 2000 requires searching a multilingual document collection for relevant documents, and listing the results in a merged, ranked list. Although the official languages are again English, French, German and Italian, it is also possible to submit runs in which the document collection is queried in other languages. In this case, participants will be responsible for the translation of the query into their selected language. The results for such runs will be given separately. A pair-wise cross-language task is provided in which the query language can be French, German or Italian and the target document collection is English. Many IR groups are now beginning to work on retrieval over pairs of languages and this will give them a chance to participate officially in the CLEF activity. Unofficial bilingual runs in which the query to the English document collection can be in any other European language can also be submitted and will be evaluated.

Multilingual information retrieval implies a good understanding of the issues involved in monolingual retrieval. It is often asserted that procedures for monolingual information retrieval are (almost) completely language independent. This is not however true; different languages present different problems. Methods that may be highly efficient for certain language typologies may not be so effective for others. Issues that have to be catered for include word order, morphology, diacritic characters, language variants. So far, most IR system evaluation has focussed on English. CLEF will provide the opportunity for monolingual system testing and tuning and build up test suites in other European languages (beginning with French, German and Italian in CLEF 2000).

The CLEF multilingual document pool for 2000 consists of comparable corpus consisting national newspapers for all four languages from the same time period; a change from the news agency stories of previous years. Topics will be developed much as before; however, the use of Italian French and German national papers rather than Swiss sources will perhaps extend the multicultural aspect. It is hoped to be able to offer additional languages in future years. The number of topics will be increased with the aim of building up the size of the pool as quickly as possible.

The results of CLEF 2000 will be presented at a two-day workshop to be held in September in Lisbon, Portugal, immediately after the fourth European Conference on Digital Libraries (ECDL 2000). The first day will be open to all interested participants and focussed on research related issues in Multilingual Information Access. The second day will report and discuss the results of the CLEF activity and will be restricted to active CLEF participants.

More information on CLEF can be found at http://www.iei.pi.cnr.it/DELOS/CLEF/.

Acknowledgements

Our thanks go to the data providers, Neue Zürcher Zeitung (NZZ), Schweizerische Depeschenagentur (SDA) and the Associated Press (AP). We would also like to express our gratitude to the topic creation and relevance assessment sites for their great work.

References

Boughanem, M., Julien, C., Mothe, J., and Soule-Dupuy C. (2000). Mercure at trec8: Adhoc, Web, CLIR and Filtering tasks. In *Proceedings of the Eighth Text Retrieval Conference (TREC8).*

Braschler, M., Kan, M.-Y., Schäuble, P., and Klavans, J. (2000). The Eurospider Retrieval System and the TREC-8 Cross-Language Track. In *Proceedings of the Eighth Text Retrieval Conference (TREC8).*

Braschler, M., Krause, J., Peters, C., and Schäuble, P. (1999). Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of the Seventh Text Retrieval Conference (TREC7).*

Franz, M., McCarley, J. S., and Ward, R. T. (2000). Ad hoc, Cross-language and Spoken Document Information Retrieval at IBM. In *Proceedings of the Eighth Text Retrieval Conference (TREC8).*

Gey, F. C. and Jiang, H. (2000). English-German Cross-Language Retrieval for the GIRT Collection - Exploiting a Multilingual Thesaurus. In *Proceedings of the Eighth Text Retrieval Conference (TREC8).*

Kraaij, W., Pohlmann, R., and Hiemstra, D. (2000). Twenty-One at TREC-8: using Language Technology for Information Retrieval. In *Proceedings of the Eighth Text Retrieval Conference (TREC8).*

Mayfield, J., McNamee, P., and Piatko, C. (2000). The JHU/APL HAIRCUT System at TREC-8. In *Proceedings of the Eighth Text Retrieval Conference (TREC8).*

Nie, J.-Y. (2000). CLIR using a Probabilistic Translation Model based on Web Documents.

Oard, D. W. (1997). Cross-Language Text Retrieval Research in the USA. Presented at *3rd ERCIM DELOS Workshop, Zurich, Switzerland.*
Available from http://www.clis.umd.edu/dlrg/filter/papers/delos.ps

Oard, D. W., Wang, J., Lin, D., and Soboroff, I. (2000). TREC-8 Experiments at Maryland: CLIR, QA and Routing. In *Proceedings of the Eighth Text Retrieval Conference (TREC8).*

Qu, Y., Jin, H., Eilerman, A. N., Stoica, E., and Evans D. A. (2000). CLARIT TREC-8 CLIR Experiments. In *Proceedings of the Eighth Text Retrieval Conference (TREC8).*

Ruiz, M., Diekema, A., and Sheridan, P. (2000). CINDOR Conceptual Interlingua Document Retrieval: TREC-8 Evaluation. In *Proceedings of the Eighth Text Retrieval Conference (TREC8).*

Schäuble, P. and Sheridan, P. (1998). Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of the Sixth Text Retrieval Conference (TREC6).*

Voorhees, E. M. (1998). Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*