

The Evaluation of Systems for Cross-Language Information Retrieval

Martin Braschler¹, Donna Harman², Michael Hess³, Michael Kluck⁴, Carol Peters⁵ and Peter Schäuble¹

¹Eurospider Information Technology AG, Zurich, Switzerland. braschler|schauble@eurospider.com

²National Institute of Standards and Technology, Gaithersburg, Md, USA. donna.harman@nist.gov

³Dept. of Information Technology, University of Zurich, Switzerland. hess@ifi.unizh.ch

⁴InformationsZentrum Sozialwissenschaften (IZ), Bonn, Germany. mkl@bonn.iz-soz.de

⁵Istituto di Elaborazione della Informazione - CNR, Pisa, Italy. carol@iei.pi.cnr.it

Abstract

We describe the creation of an infrastructure for the testing of cross-language text retrieval systems within the context of the Text REtrieval Conferences (TREC) organised by the US National Institute of Standards and Technology (NIST). The approach adopted and the issues that had to be taken into consideration when building a multilingual test suite and developing appropriate evaluation procedures to test cross-language systems are described. From 2000 on, a cross-language evaluation activity for European languages known as CLEF (Cross-Language Evaluation Forum) will be coordinated in Europe, while TREC will focus on Asian languages. The implications of the move to Europe and the intentions for the future are discussed.

1. Introduction

The availability of powerful cross-language information retrieval (CLIR) systems that enable users to find and retrieve relevant information in whatever language it has been stored is a key factor for global access and sharing of knowledge. Users of internationally distributed information networks need to be able to find, retrieve and understand relevant information, in whatever language and form it may have been stored. Situations where the user is faced with the task of querying a multilingual document collection are becoming increasingly common.

Many users have some foreign language knowledge but their proficiency may not be good enough to formulate queries to appropriately express their information needs. Such users will benefit enormously if they can enter their query in their native language because they are able to examine relevant documents in other languages, even if they have not been translated. Users with no target language knowledge can send relevant retrieved material to a translation service. The key issue is to be able to find relevant information in the first place, and to know that it is relevant. For this reason, much attention has been given over the last few years to the study and development of tools for cross-language information retrieval, i.e. tools that allow users of document collections in multiple languages to formulate queries in their preferred language and retrieve relevant information, in whatever language it is stored.

However, the development of such tools is not an easy task. Various approaches are currently being studied involving different strategies, such as machine translation, multilingual lexicons, thesauri and ontologies, parallel or comparable corpora (Oard, 1997). All these approaches imply the development of complex systems that integrate methodologies and tools developed for Natural Language Processing (NLP) with techniques from Information Retrieval (IR). These systems require an intensive process of testing and tuning before they can be implemented successfully in end user applications.

2. CLIR at TREC

It has been demonstrated extensively by the Text REtrieval Conference series that the availability of evaluation procedures can contribute significantly to the improvement of system. For this reason, in 1997, it was decided to include cross-language system evaluation as one of the tracks at TREC. The aim was to provide developers with an infrastructure enabling them to test and tune their systems and compare the results achieved using different cross-language strategies.

The main goals of the CLIR track in TREC have been to:

1. create the infrastructure for testing cross-language information retrieval technology through the creation of a large-scale multilingual test collection and a common evaluation setting;
2. investigate effective evaluation procedures in a multilingual context;
3. provide a forum for the exchange of research ideas.

The first cross-language retrieval task in TREC-6 used documents in English, French and German and topics (query statements) in English, French, German, Spanish and Dutch (Schäuble & Sheridan, 1998). Participating groups submitted runs for any pair-wise combination of languages. Comparison was based on monolingual baselines in the document language, which had to be submitted for each cross-language experiment. However, the large number of possible language combinations meant that there was little possibility to make comparisons between different types of systems or approaches.

The principal focus of the activity was thus altered in TREC-7 and TREC-8, where the main task was for groups to use topics written in one language in order to retrieve documents from a single pool of documents in four languages – English, French, German and Italian (Braschler et al. 1999; 2000). The results were to be submitted in a merged, ranked list. As all systems were querying the same multilingual collection, a better comparison between different techniques and strategies was possible. Although this was a hard task, it was felt to reflect more accurately the kind of performance the user

community is now demanding of cross-language systems.

The document collections consisted mainly of newswires, mostly for the same time period (1988-90) for the four languages¹. The comparability of the collections was guaranteed by style (all news stories) and period, which ensured that many current events of more than purely national interest would be covered by more than one if not all of the collections.

By increasing the difficulty of the task, many research groups were stimulated to develop or extend systems originally built to run for pairs of languages in order to handle several languages. In any case, bilingual runs were still accepted from those groups who did not feel like coping with the full task. A subtask was also provided for domain specific cross-language retrieval (English-German) on a structured database of social science documents.

2.1 Why System Evaluation is Important

The CLIR tracks at TREC have witnessed a wide range of approaches to the problem. Of course, all the classical techniques have been extensively adopted. These include the use of machine translation for query translation, the use of bilingual dictionaries and associated disambiguation techniques, and various corpus-based approaches that build data structures for CLIR from comparable or parallel text corpora.

However, there have also been notable experiments that diverged widely from some standard assumptions with regard to CLIR – such as the common claim that a translation of all the documents in a collection is impractical. In TREC-6, the University of Maryland translated the entire collection using a machine-translation system - although this did require very substantial resources (Oard & Hackett, 1998). Another group even did CLIR without doing any translation at all: Cornell produced an English -> French run for TREC-6 by treating English words as potentially misspelt French words (Buckley et al., 1998). An original experiment was contributed by New Mexico State University in TREC-8: they tested how well monolingual users can perform cross-language retrieval if they are assisted in manually disambiguating query terms (Ogden et al., 2000).

This variety of approaches demonstrates the ability of participants to use new ideas. It also leads to other groups picking up ideas and refining them, therefore contributing substantially to advancements in the field. In TREC-8 we observed how groups increasingly try to build on concepts proven to be successful in earlier TRECs, for example for merging strategies. By expanding the number of subtasks available to participants in the future, we expect to keep attracting new, innovative approaches.

2.2 Topic Creation and Relevance Assessment

The evaluation methodology adopted for the CLIR activity has been an adaptation of the strategy previously studied for the ad-hoc task, the main monolingual system

evaluation track in TREC performance (Voorhees & Harman, 1999). However, a number of task-specific issues had to be investigated when defining the criteria for topic development, relevance assessment and results pooling in a multilingual environment. These will be discussed in the following.

For TREC-6, the CLIR track topics were developed centrally at NIST. 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 and TREC-8, this work was distributed over sites where the different languages are spoken natively².

The ad-hoc TREC formula, consisting of a three word title, a brief description and a longer narrative, was followed. Participants could submit runs using any or all of these three elements, using their preferred topic language. An example taken from last year's topic set for English is:

Title: Statistics about abortion

Description: Find statistics on legal and/or illegal abortions throughout the world.

Narrative: Relevant documents report statistics on the number of legal or illegal abortions in the world or in different countries, for instance with reference to the total number of pregnancies. Documents containing only opinions about abortion or giving the number of abortions by a single person or in a single clinic will be rejected.

Each site prepared a number of topics in one of the four languages of the document collection. Topics were created to reflect real world information needs and, for each set of documents, to cover national, European and international issues (in approximately equal parts). This meant that queries were not necessarily matched by relevant documents in all the collections. Certain, very specific queries focussing on topics of purely national interest may only retrieve documents from a single collection; other topics may find far greater coverage in some collections rather than others. There was a deliberate imbalance in topic vs. collection representativeness. Participating systems could not rely on any assumptions with respect to retrieval rate against collections.

The final topic set was chosen from the input provided by each group and then produced in all the topic languages. All translations were by fluent target language speakers and (almost) always were directly from the source to the target language. Generally, it is not permitted to translate a topic from an already translated one. This is because there is a risk that meaning can shift from one version to the next.

The translation techniques adopted have been studied to ensure an acceptable balance between precision with respect to the source and naturalness with respect to the target language. However, at times, for culturally sensitive material, direct translations are not possible. In these cases, it is necessary for the translator to provide a

¹ The English texts were provided by the Associated Press; German, French and Italian documents came from SDA, the "Schweizerische Depeschagentur" (Swiss News Agency). For German, texts from the Swiss newspaper "Neue Zürcher Zeitung" (NZZ) for 1994 were also added.

² English: NIST, USA; French: Univ. Zurich, CH; German: IZ-Bonn & Univ. Koblenz, DE; Italian: IEI-CNR, IT.

paraphrase. For example, a topic originally formulated in French on the Swiss public debt included the statement that "la plus grande partie de la dette publique est couverte par les placements". This was rendered in English as: "However the major part of the public debt is covered by the equivalent of U.S. Treasury bonds". Similar expedients, using the terms for the national equivalent, were adopted in the Italian and German versions.

While preserving the topic meaning, terms must be used in the target topic that can realistically be expected in the document collection for that language. Thus a high level of performance is required of topic translators to avoid an imbalance in topic authenticity. The aim is a complete set of source language equivalent topics for each language in the document collection, in order to create as close to real world conditions as possible.

The participating groups then select one of the topic languages as the language used by their system to query the multilingual document pool. They can submit runs using different topic languages if they wish.

The relevance assessments are also produced in the same distributed setting. Methods have been studied by TREC to ensure a high degree of consistency in the relevance judgements. All assessors follow the same criteria when judging the documents. An accurate assessment of relevance for retrieved documents for a given topic implies a good understanding of the topic. This is much harder to achieve in the distributed scenario of the CLIR track where understanding is influenced by language and cultural factors.

Although the topic creators initially work on the basis of their knowledge of possible events for the years covered by the document collections, the final choice and refinement of the topics is made with respect to the contents of the document collections. The way a particular argument is presented in a collection tends to influence its formulation. Thus a topic which does not appear to raise problems of interpretation in the language used for its preparation may be far more difficult to assess against the documents in another language. Some topics, although perfectly clear to the creator, may be found by the assessors to be too vague or difficult to interpret, while others require very specific knowledge that may have been underestimated at the moment of creation. When, for example, it is a question of understanding whether a particular tropical forest is in South America or whether a named Chinese town is actually in the Yunnan region, this is not too much of a problem, but at times a correct interpretation of a topic requires specific knowledge in a particular domain in order to be able to assess all the documents correctly. Depending on the domain, this is more difficult to guarantee.

A continual process of electronic discussion and verification between the assessors at each site is thus necessary during the relevance assessment stage in order to ensure, as far as possible, that the decisions taken as to relevance are consistent over sites, over languages and over collections. Nevertheless, even though we are convinced of the importance of such consistency checking, it has been demonstrated that variations in relevance judgements generally do not invalidate overall

relative evaluation. This will be discussed in the section describing the analysis of the results.

2.3 Domain-specific Evaluation

A domain-specific subtask using the GIRT³ test collection, which covers a vertical domain (social sciences), has been offered since TREC-7. The rationale of this subtask is to test retrieval on another type of document collection, serving a different kind of information need. The information which is provided by these social science documents is far more targeted than news stories and contains a lot of terminology. The users of this kind of domain-specific collections tend to be recall-oriented: they are typically interested in the completeness of results. This means that they are generally not satisfied with finding just some relevant documents from a collection that may contain much more. Developers of domain-specific cross-language retrieval systems need to be able to tune their systems to meet this requirement.

The GIRT database contains nearly 80,000 documents, extracted from the FORIS database (descriptions of ongoing social science research projects in German speaking countries - Austria, Germany, Switzerland) and SOLIS (social science literature in German). These databases are made available by IZ-Bonn (Centre for Social Sciences) via traditional host services and the Internet. The data contains bibliographical information and additional indexing terms, classification texts, and abstracts. Besides the German texts, there are English translations of the titles (for nearly all documents) and about 20% of the documents have additional English abstracts. English indexing terms can be used for retrieval purposes by taking the English equivalents for the German terms from the bilingual social science thesaurus of IZ, also provided with the data.

Specific topics have to be built for this subtask because the data is very different from that found in newspapers or newswires. The documents treat more long-term societal problems in an in-depth manner; current problems are dealt with some time-lag. Thus the topics must be built with respect to the characteristics of the collection and the relevance judgements of the results sets have to be related to the domain. These judgements are made by social science experts working at IZ.

A domain-specific language requires appropriate indexing and retrieval systems. Recent results clearly show the difficulty of differentiating between specific sociological terms and common language terms: "words [used in sociology] are common words that are [also] in general use, such as community or immigrant"⁴. In many cases there exists a clear difference between the scientific

³ German Indexing and Retrieval Test Database. The first use of the GIRT test database for the evaluation of retrieval systems is described in detail in German (Frisch & Kluck, 1997) and more briefly in English (Kluck, 1998).

⁴ Haas 1997, p. 78; cf. p. 74: "T tests between discipline pairs showed that physics, electrical engineering, and biology had significantly more domain terms in sequences than history, psychology, and sociology (...) the domains with more term sequences are those which may be considered the hard sciences, while those with more isolated domain terms tend to be the social sciences and humanities."

meaning and the common meaning. Furthermore, there are often considerable differences between scientific terms when used in different domains, owing to different connotations, theories, political implications, ethical convictions, and so on. This means that it can be more difficult to use automatically generated terms and queries for retrieval in a vertical collection. For example, as noted by Ballesteros & Croft (1998), when using a dictionary-based cross-language query system "queries containing domain-specific terminology which is not found in general dictionaries were shown to suffer an additional loss in performance". The challenge is to adjust general concepts for retrieval systems to the domain-specific area.

2.4 The Role of Relevance Judgements

The analysis of results for cross-language system evaluation at TREC is based on proven measures for IR evaluation, such as recall/precision graphs and average precision figures. These measures rely on dependable relevance assessments. This dependence means that the quality and reliability of relevance assessment is (justifiably) constantly questioned. For people to be able to rely on the results they obtain, they need to have confidence in the quality of the underlying relevance judgements.

The central difficulty with relevance assessment is caused by the fact that relevance is usually subjective. Whether someone considers a document as relevant with respect to a particular information need is dependent on factors such as personal background knowledge and personal preference and context, not specified in the information request. This means that two people might have very different opinions on the relevance of a particular document. However, even the same person can judge a document relevant that he/she previously considered irrelevant when these factors change.

The importance of this issue has led to a considerable amount of research. An extensive study was made by Voorhees (1998). Voorhees investigated the properties of relevance assessments produced for the TREC-4 and TREC-6 conferences. She found that the relative effectiveness of different retrieval strategies remains stable despite marked differences in the relevance judgements used to measure retrieval. This finding is also consistent with earlier studies. In other words, it means that while the actual values of the effectiveness measure (e.g. average precision) are affected by the differences in relevance judgements, the retrieval performance remains almost always constant. This is true for cross-system comparisons, and even more so when comparing algorithmic variants of the same retrieval system.

Voorhees has tested a variety of set-ups to support this conclusion, including judgements made by topic authors vs. judgements made by non-authors, judgements made by a single judge vs. group judgements and judgements made in very different environments. Even though the disagreement between judges can be (surprisingly) high⁵, she concludes that the relevance assessments remain valid for all these cases.

⁵ Voorhees states, for example, that across all topics she investigated, 30% of the documents that the topic author marked as relevant were rejected as non-relevant by both the additional independent assessors that she used in her study.

While these findings referred to the monolingual ad-hoc task of TREC, many of them are very important for CLIR relevance assessments. The distributed relevance assessment procedure means that we have to deal with issues of multiple assessors, potentially working in very different environments. It also means that a topic author cannot usually assess the documents in languages other than the one used for the formulation of the topic.

Additional problems in cross-language relevance assessment arise because of the added effort of performing topic creation and relevance assessment in multiple languages. The added effort in topic creation means that it is more difficult to increase the number of topics. On the other hand, the stability of relevance assessments increases with the number of topics, since effects of changes in individual topics even out when averaged over a sufficient number of topics. Similarly, the pool of assessed documents has to be of adequate size. Again this is difficult because, for every language, a pool of documents of sufficient size has to be assessed. Consequently, questions about the quality of CLIR assessments have surfaced, starting with TREC-7.

An experiment with respect to the influence of the CLIR pool was made by the TwentyOne group (Kraaij et al., 2000). When doing an analysis based on their figures, it appears that the central claim by Voorhees, namely that relative performance is stable, holds for the CLIR TREC-7 relevance assessments they examined (Braschler et al., 2000). This means that we feel confident with respect to the usability of the assessments for the participants in the respective TREC-CLIR tasks. Because of the limited size of the pools used for past years, however, more caution is needed when making comparisons between systems that did not participate. With the expansion of our CLIR evaluation activities in the future, we are convinced that the reliability of the assessments will be increased for such non-participating systems. With the addition of new participants, and the planned increase in the number of topics, we hope to maintain a high level of quality, and this will be carefully monitored.

3. Move to Europe

From 2000 on, this cross-language evaluation activity for European languages will be coordinated in Europe rather than in the U.S, while TREC will focus on Asian languages. The launching of an independent activity – known as CLEF (Cross-Language Evaluation Forum)⁶ and pronounced "clé" – will allow us to focus on a wider range of issues.

The systems participating in the CLEF series will have to solve three major problems: finding translations, pruning translations, and weighting translation alternatives. Many additional issues must also be dealt with: maintaining variant collating orders, normalising accents, separating languages within a collection, language specific stemming and morphology, merging results (cf. Grefenstette 1998).

⁶ CLEF is organised in collaboration between the US National Institute of Standards and Technology (NIST) and the DELOS Network of Excellence for Digital Libraries, funded by the European Commission.

In CLEF 2000, there will be three main evaluation tasks, testing multilingual, bilingual and monolingual (non-English) information retrieval systems, plus the subtask for domain-specific cross-language evaluation.

Similarly to TREC-7 and 8, the main task of CLEF 2000 requires searching a multilingual document collection for relevant documents. The multilingual collection now consists of newspaper documents in four languages (again English, French, German and Italian) for the same time period⁷. For each language at least one national newspaper is represented. This should offer a more linguistically and culturally representative corpus than the previously used multilingual collection of (mainly) newswire stories in which the French, Italian and most of the German texts all came from a single source: SDA, the Swiss news agency.

In order to encourage European participation, we have increased the number of official topic languages. CLEF 2000 will also have Spanish, Dutch, and possibly Swedish and Finnish, as topic languages in addition to the languages of the multilingual collection. This, however, raises a number of issues. Previous criteria for the preparation of the topic set established that an equal number of topics had to be produced with respect to each collection in the multilingual set, and that topic translation should be directly from source to target. As more topic languages are added, it will become impossible to respect either of these conditions. In any case, the requirement remains that, for each language, the topic set should be as linguistically representative as possible, i.e. using the terms that would naturally be expected to represent the set of query concepts in the given language.

A bilingual system evaluation task will also be offered. In CLEF 2000, this will consist of querying the Los Angeles Times collection using any topic language (other than English). In later years, the target collection may well be in another language, French or German for example.

In order to be successful with multilingual retrieval, a good understanding of the questions involved in monolingual information retrieval is necessary. 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. One of the aims of the CLEF activity is to encourage the development of tools to manipulate and process languages other than English. For this reason, in CLEF 2000, a third task will regard monolingual (non-English) information retrieval for systems developed to run on French, German and Italian.

The addition of both the bilingual and the monolingual tasks will help us to build up the pool of assessed documents quickly, so that effective and reliable test collections in a number of European languages can be made available to the research community for system testing and tuning purposes.

However, the goal of CLEF is not only to offer a system evaluation infrastructure but also to provide researchers with a suitable forum to discuss ongoing work and present new ideas and approaches. A yearly workshop will thus be held in which the results of the activity can be reported and compared.

4. Intentions for the Future

Our objectives for the future are ambitious; it remains to be seen how many of them are realisable. There are three directions in which we would like to go: (i) the addition of more languages – at least to cover the major European ones; (ii) the addition of new tasks; (iii) the setting up of an infrastructure to handle cross-language retrieval for other types of media in addition to text.

The first of these points - the inclusion of more languages – is not an easy target to meet. The addition of extra languages is very costly in terms of resources and time. It is clear that it is not possible to include too many languages in the collections for the multilingual task as it will become too daunting for most potential participants. A large number of languages also raises questions with respect to ensuring an acceptable level of comparability between results. This question will have to be investigated considerably. We have already mentioned the difficulty of creating authentic topics in all topic languages. In CLEF 2001, we hope to include Greek, thus including the problem of handling not only a large number of languages but also of dealing with multiple character sets.

The addition of new tasks could be very interesting. As has been stated, CLEF aims at reflecting real world information needs; an evident need is for tools for multilingual access and retrieval on the Web. Since 1999, TREC has introduced a Web track; it would be wonderful if CLEF could do likewise.

Text is just one of the media in which information is made available in digital form. Another future goal could be to establish a set of metrics for evaluating systems for cross-language speech retrieval.

Last but not least, we would like to study evaluation methodologies with respect to user needs. Very little is known as yet with respect to the expectations and real needs of the users of systems for multilingual information access. Even less is known as to how far the current evaluation infrastructure is really providing the best metrics to stimulate systems to meet these – as yet largely unknown – needs. This would be an important and valuable area for future research.

All this implies the creation of a complex infrastructure involving much effort and considerable resources. Immense rigour in task setting, topic creation, relevance assessment, and results analysis is needed in order to provide a quality service to the CLIR community with the aim of stimulating development of cross-language systems that are capable of handling all kinds of situations and satisfying a wide range of needs.

5. Acknowledgements

We gratefully acknowledge the support of all the data providers and copyright holders, and in particular:

Newswires: Associated Press, USA; SDA - Schweizerische Depeschagentur, Switzerland.

⁷ Los Angeles Times 1994; Le Monde 1994; Frankfurter Rundschau 1994, Der Spiegel 1994, 1995; La Stampa 1994.

Newspapers: English texts provided by the Los Angeles Times. French data from Le Monde and ELDA: European Language Resources Distribution Agency. German collections by Frankfurter Rundschau, Druck und Verlagshaus Frankfurt am Main; Der Spiegel, Spiegel Verlag, Hamburg; Neue Zürcher Zeitung, Zurich; GIRT - InformationsZentrum Sozialwissenschaften, Bonn. Italian data from La Stampa 1994 - Hypersystems Srl, Torino.

Without their help such a large scale evaluation would be impossible.

6. References

- Ballesteros, L. and Croft W.B. 1998. Statistical Methods for Cross-Language Information Retrieval. In G. Grefenstette (ed.), *Cross-Language Information Retrieval* (pp. 21-40). Boston.
- 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 (TREC-7)*. NIST, Gaithersburg, MD.
- Braschler, M., Peters, C., and Schäuble, P. 2000. Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*. NIST, Gaithersburg, MD.
- Buckley, C., Mitra, M., Walz, J. and Cardie, C. 1998. Using Clustering and SuperConcepts Within SMART: TREC6. In *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*. NIST, Gaithersburg, MD.
- Frisch, E. and Kluck, M. 1997. *Pretest zum Projekt German Indexing and Retrieval Testdatabase (GIRT) unter Anwendung der Retrievalsysteme Messenger und freeWAIssf*. Bonn: InformationsZentrum Sozialwissenschaften (IZ-Arbeitsberichte, Nr. 10).
- Grefenstette, G. 1998. The Problem of Cross-Language Information Retrieval. In G. Grefenstette (ed.), *Cross-Language Information Retrieval* (1-9), Boston.
- Haas, D.K. 1997. Disciplinary Variation in Automatic Sublanguage Term Identification. *J ASIS*, 48: 67-79.
- Kluck, M. 1998. German Indexing and Retrieval Test Data Base (GIRT): Some Results of the Pre-test. In M.D. Dunlop (ed.), *The 20th BCS IRSG Colloquium: Discovering New Worlds of IR (IRSG-98)*, Grenoble, France.
<http://www.ewic.org.uk/ewic/workshop/view.cfm/IRSG-98>
- 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 (TREC-8)*. NIST, Gaithersburg, MD.
- Oard, D.W. 1997. Alternative Approaches for Cross-Language Text Retrieval. In *Cross-Language Text and Speech Retrieval*, AAAI Technical Report SS-97-05.
<http://www.clis.umd.edu/dlrg/filter/sss/papers/>
- Oard, D. W. and Hackett, P. 1998. Document Translation for Cross-Language Text Retrieval at the University of Maryland. In *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*. NIST, Gaithersburg, MD.
- Ogden, B., Cowie, J., Ludovik, E., Molina-Salgado, H., Nirenburg, S., Sharples, N., Sheremtyeva, S. 2000. CRL's TREC-8 Systems. Cross-Lingual IR, and Q&A. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*. NIST, Gaithersburg, MD.
- Schäuble, P. and Sheridan, P. 1998. Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*. NIST, Gaithersburg, MD.
- 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*.
- Voorhees, E.M. and Harman, D. 1998. Overview of the Seventh Text REtrieval Conference. In *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*. NIST, Gaithersburg, MD.