The CLEF Campaign

Martin BRASCHLER
Eurospider IT AG, Schaffhauserstr. 18, 8006 Zürich, Switzerland
braschler@eurospider.com

Carol PETERS
IEI-CNR, Area della Ricerca di San Cataldo, 561000 Pisa, Italy
carol@iei.pi.cnr.it

Abstract

The Cross-Language Evaluation Forum (CLEF) provides an infrastructure aimed at supporting the development, testing and evaluation of systems for cross-language information retrieval, and for monolingual information retrieval of European languages other than English. Originally started as a track at the TREC-6 conference, CLEF became an independent initiative in 2000 when the coordination moved to Europe. The diversity of languages commonly spoken in Europe has led to a multilingual, distributed setup that was chosen to best accommodate the unique linguistic properties of each language and the implications for topic development and relevance assessments. The tasks in CLEF 2000 involved a multilingual document collection in four core languages and several additional topic languages. Twenty groups participated in the campaign, submitting a wide range of experiments. The test collection was subsequently analyzed with respect to the completeness of the assessments in order to ensure the validity for future evaluation and benchmarking activities. CLEF will continue in 2001, with several additions to the individual tasks.

Keywords: CLEF, Cross-Language Information Retrieval, Multilingual Information Retrieval, Evaluation Forum.

1 Introduction

The first evaluation campaign of the Cross-Language Evaluation Forum (CLEF) took place in 2000. The current goals of CLEF are to

- Support the evaluation of systems for cross-language retrieval
- Encourage the development of strategies and tools for advanced non-English monolingual retrieval
- Attract more European participants to this kind of evaluation task

The main focus of CLEF are European languages. English is specifically excluded from the monolingual retrieval task, in order to avoid duplication of effort with the TREC conference series, but is included for all other types of experiments.

CLEF is the successor of the TREC Cross-Language (CLIR) track as offered in TREC-6 through TREC-8 [9] [1] [2]. A distinctive feature of CLEF is its distributed organizational setup, which was chosen to accommodate the unique linguistic properties of the languages that are covered. The implementation of this distributed setup began during the CLIR activity within TREC and had a determining influence on the decision to create an independent activity.

CLEF 2000 included four core languages in its document collections (English, French, German, Italian) and a fifth (Spanish) is being added in 2001. For each of these languages, groups providing the respective linguistic knowledge are involved in the organization of CLEF. This particular setup has implications on the topic development and relevance assessment, which are implemented as distributed processes.

The paper first gives a brief outline of the activities that led to the founding of CLEF. It then describes the setup of the first CLEF campaign that took place in 2000. This is followed by an analysis of the issues that arise from distributed topic development and relevance assessment. After a summary of the results of the initial campaign, the paper closes with an outline of the plans for CLEF 2001.

2 History

In 1997, a new cross-language information retrieval (CLIR) track was introduced at the TREC-6 conference [9]. TREC, the popular series of IR evaluation conferences organized by the National Institute of Standards and Technology (NIST) [5] [10], had included non-English (Spanish and Chinese) retrieval evaluation in the past, but these were

strictly monolingual experiments. The new TREC-6 CLIR track defined an evaluation task that required participants to use topics in one language in order to retrieve documents formulated in a second language (bilingual retrieval). Document languages were English, French and German. Topic languages were the same three languages plus, additionally, Spanish and Dutch. This was the first evaluation campaign for cross-language information retrieval and was met by considerable interest in the research community. Since then, other evaluation forums have also started to offer cross-language experiments with various languages, among them NTCIR [6] and Amaryllis [4].

A total of thirteen groups from six different countries participated in the first CLIR at TREC campaign in 1997. However, two main problems became quickly evident. First, the many different language pairs that were used by the participants made it difficult to compare across systems. Second, it proved to be difficult for NIST to find people for topic development and relevance assessment who were sufficiently proficient in all the necessary languages.

The first issue was addressed by introducing a truly multilingual "main" task at TREC-7. In this task, participating groups were required to tune their systems for search on all document languages simultaneously, i.e. on a multilingual collection. This also more accurately reflects real-world scenarios, where users often access documents in their native as well as in some additional, foreign languages. Participants approached this new task by either building unified, multilingual search indexes, or by merging the results of multiple bilingual retrieval results.

It was felt that the second issue could only be resolved by radically changing the organization of the track, and by moving from a centralized setup at NIST to a new, distributed setup involving additional organizing groups that provide the language knowledge for individual languages. This idea was realized by involving new groups based in Germany, Italy (the TREC-7 CLIR track added Italian as a fourth language), and Switzerland (for French). NIST remained actively involved by handling English, and providing the infrastructure.

After successful campaigns at TREC-7 and TREC-8, it became clear that in order for the activity to grow, and to expand to cover more languages, it would be beneficial to move it to Europe, and turn it into an independent activity. It was also hoped that this move would attract more European groups, which had been under-represented while the campaign was centered in the United States. The new activity was named Cross-Language Evaluation Forum (CLEF).

3 CLEF 2000 Setup

CLEF 2000 was sponsored and promoted by the DELOS Network of Excellence for Digital Libraries [8], which is funded by the European Commission. The first evaluation campaign started in early 2000, and ended with a workshop in Lisbon, Portugal, in September 2000, held in conjunction with ECDL 2000 (the European Conference on Digital Libraries). The CLEF steering committee consisted of six groups:

- IEI-CNR, Pisa, Italy (Coordinator, responsible for Italian)
- NIST, Gaithersburg, USA (English)
- IZ Sozialwissenschaften, Bonn, Germany (German)
- University of Zurich, Zurich, Switzerland (French)
- Eurospider IT AG, Zurich, Switzerland (Technical coordinator)
- UNED, Madrid, Spain (Spanish topic translations; added during the campaign)

The multilingual document collection in CLEF 2000 covered four core languages (English, French, German, Italian) and, in addition, four more languages (Dutch, Finnish, Spanish, Swedish) were included in the topic set. Four different tasks were offered:

- *Multilingual Information Retrieval* (X > EFGI). Search on a multilingual document collection containing documents in all four core languages. Free choice of topic language out of the eight made available (different topic languages can be used for different experiments)
- *Bilingual Information Retrieval* (X > E). Search on the English document collection. Free choice of topic language (other bilingual language combinations were discouraged to maintain comparability of results across systems)
- *Monolingual Information Retrieval* (F > F, G > G, I > I). Monolingual retrieval for German, French and Italian. No English monolingual retrieval was offered, in order to avoid duplication with TREC.
- *Domain-Specific Retrieval*: Choice of German or English topics to search a German document collections containing texts from the domain of social sciences. First introduced in the TREC-7 CLIR track, the goal of this task is to investigate the special issues in CLIR on texts with domain-specific terminology. The documents contain descriptors, and a corresponding bilingual German/English thesaurus is provided to the participants.

Apart from the domain-specific task, that used its own data set (the GIRT German document collection), the other subtasks used documents from a multilingual collection of newspaper texts in all four core languages. All these texts are from the years 1994 and 1995, and were taken from the LA Times (English), Le Monde (French), Frankfurter Rundschau and Der Spiegel (German) and La Stampa (Italian). The complete multilingual collection contained approximately 360,000 articles totaling around 1.1 GB. This document collection can be considered as a comparable corpus as, to a large extent, the texts in the different languages are similar in content, style and time. Each language is represented by a national newspaper for the same period which, with the exclusion of local news, is reporting events of international and general interest. Thus much of the content of each collection is similar, although rendered in different languages and from different national and cultural perspectives. The CLEF document collection does not contain direct translations.

Forty topics were developed and translated in all four core languages. External groups then translated these topics into four additional languages. As in TREC, topics in CLEF are statements of user needs. In contrast, queries are the actual strings submitted to the individual systems used by the participants. The domain-specific GIRT task used its own set of 25 topics. The goal of all tasks was to retrieve for each topic the best 1000 matching documents from the respective collections. Both automatic (no manual interference) and manual (all other) experiments were allowed.

4 Topic development

The evaluation methodology adopted for the CLEF activity has been an adaptation of the strategy previously used for the TREC ad-hoc task, the main monolingual system evaluation track in TREC [10]. However, mainly due to the distributed setup, a number of issues had to be investigated when defining the criteria for topic development, relevance assessment and results pooling.

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 hard to achieve if there is just one topic creation site. Consequently, starting with TREC-7, and continuing in CLEF, this work has been distributed over sites where the different languages are spoken natively.

The ad-hoc TREC formula, consisting of a very short title (typically two or three words), 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 the CLEF topic set for English is:

Title: The Electroweak Theory
Description: Find documents that report recent discoveries in the field of subnuclear physics that confirm the unified electroweak theory of Weinberg-Salam-Glashow.
Narrative: Relevant documents report on discoveries in the last ten years of subatomic particles, such as quarks or photons, which provide experimental confirmation of the standard theoretical model of nuclear interactions proposed by Weinberg-Salam-Glashow. Other work in the field of nuclear physics that does not have a direct connection with this theory is not pertinent.

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). Queries were therefore 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. This was a deliberate imbalance: 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 translated to all core languages. All translations were by fluent target language speakers and (almost) always were directly from the source to the target language, in order to avoid the meaning shift that can occur when translating from non-source versions. In a second step, the topics were translated to four additional topic languages by volunteer groups.

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 paraphrase. For example, in the TREC-7 CLIR track, a topic originally formulated in French on the subject of 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 problems occurred regularly in all languages.

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. The translation must also be a realistic reflection of the way actual users would formulate the topic. Thus a high level of performance is required of

the 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.

5 Relevance assessments

The relevance assessments are also produced in the same distributed setting. CLEF uses methods adapted from 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 CLEF 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 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 exchange of e-mail for discussion and verification between the assessors at each site thus takes place 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.

CLEF currently adopts the binary relevance judgment approach by which a document is judged on a relevant/non-relevant basis, depending on whether some portion of the document actually contains information responding to the conditions stipulated in the Narrative part of the topic. This strategy has its advantages from the perspective of the calculation of the results, but imposes certain difficulties on the assessors, who are often uneasy about making such strict decisions. It is specifically the "partially relevant" documents that cause problems and lead to "subjective" judgments. For this reason, we view the multigrade approach to relevance judgment, such as that adopted by NTCIR and others, with interest, and may move towards a similar strategy in the future.

6 CLEF 2000 Results

In total, 20 groups from 10 different countries participated in one or more of the subtasks that were offered for CLEF 2000 (see Table 1) [3]. This is a significant increase from the 12 groups that took part in the TREC-8 CLIR track. The participants represent a nice mix of "veterans" that have been participating in TREC for years, and “first-timers” Most of these newcomers were from Europe, thus satisfying our goal of increasing European participation.

While most of the North American groups from earlier CLIR at TREC tracks returned for CLEF, and some new US groups also joined in, unfortunately no Asian group participated in 2000. Hopefully, in the future CLEF will be successful in attracting Asian participants who are interested in extending their cross-language systems to European languages beyond English.

Table 1. List of participants

CWI (NL)	U Dortmund (DE)
Eurospider (CH)	U Glasgow (UK)
IAI (DE)	U Maryland (US)
IRIT (FR)	U Montreal/RALI (CA)
ITC-irst (IT)	U Salamanca (ES)
Johns Hopkins/APL (US)	U Sheffield (UK)
New Mexico SU (US)	U Tampere (FI)
Syracuse U (US)	UC Berkeley (US)
TNO/U Twente (NL)	West Group (US)
U Chicago (US)	Xerox XRCE (FR)

Sixteen of the twenty participating groups did some form of cross-language experiments (either multilingual, bilingual or both), while the remaining 4 concentrated exclusively on monolingual retrieval. Three groups worked on the GIRT domain-specific task. Nine groups participated in more than one task, but no group tried all four.

All topic languages were used, with English and German being most popular. The multilingual "main" task received the most submissions. Most groups concentrated on automatic experiments, with only a few participants contributing manual runs.

The main translation strategy was query translation, although two groups also tried document translations and combined approaches. A lot of groups used some form of dictionaries. This marks a shift from the TREC-8 CLIR track, where machine translation (MT) was popular. The pronounced emphasis on work in stemming and morphological analysis was also worthy of note. A number of groups performed detailed experiments on linguistic features

such as "compounding" in German and Dutch. The results from the monolingual tasks indicate that these groups usually benefited from their work in this area as they generally obtained the best scores.

Multilingual

Figure 1 shows the recall/precision curves of the best five results (automatic runs) from the multilingual "main" task.

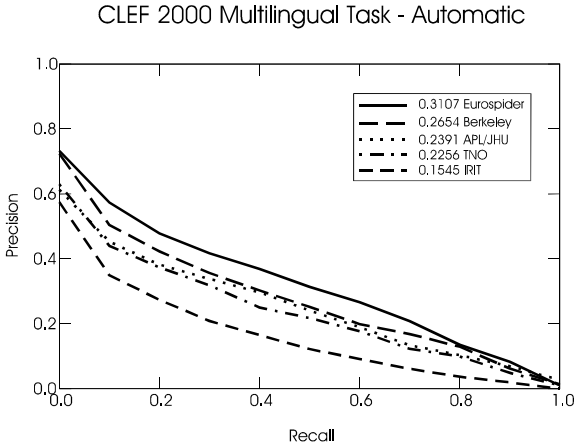

Figure 1. Best five results from the multilingual task.

It is interesting to note that all five top groups are previous TREC participants, with one of them going all the way back to TREC1 (Berkeley). These groups considerably outperformed newcomers. This may be an indication that the "veterans" benefited from the experience they gained in previous years, whereas the new groups still experienced some "growing pains". It will be interesting to see if the newcomers catch up next year. The two top performing entries (Eurospider and Berkeley) both used a combination of translations from multiple sources. The entry from Johns Hopkins University achieved good results even though avoiding the use of language-specific resources.

Bilingual

The bilingual subtasks showed a similar pattern (see Figure 2): groups that did well in the multilingual task generally also submitted bilingual runs that performed well. The University of Tampere and CWI also submitted experiments among the top five. These two groups made use of compound-splitting techniques for their topic languages (German and Dutch, respectively). This was likely beneficial to their dictionary-based approaches.

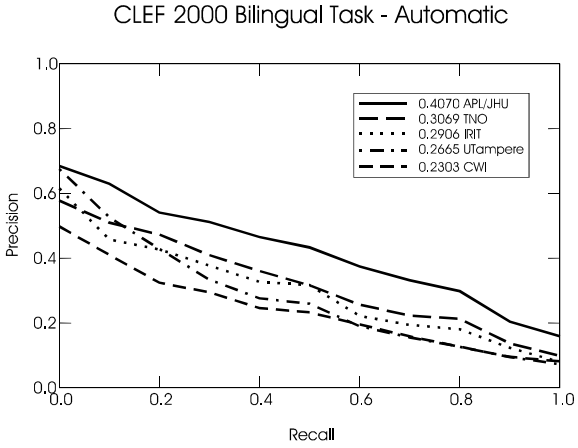

Figure 2. Best five results from the bilingual task.

Monolingual

Some of the best performing entries in the monolingual task came from groups that did not conduct cross-language experiments and instead concentrated on monolingual retrieval. Two such groups are West Group and ITC-irst, which produced the top-performing French and Italian entries, respectively. Both groups used elaborate morphological analysis in order to obtain base forms of query words and document terms. However, the performance of the top groups in French and Italian monolingual retrieval was in general very comparable.

In contrast, the differences between German monolingual entries were substantially greater. One likely explanation for the wider range in results is the decompounding issue: the four best performing groups all addressed this peculiarity of the German language, either by splitting the compounds (Eurospider, TNO, West Group) or through the use of n-grams (Johns Hopkins). Correct handling of compounds therefore appears to be crucial for good performance in German retrieval. These findings were also confirmed in the CLEF report by West Group [7].

Domain-Specific (GIRT)

Three groups submitted a total of seven runs for the domain-specific task. Xerox focused on monolingual experiments, whereas University of California at Berkeley investigated only cross-language retrieval on this collection. University of Dortmund submitted results from both monolingual and cross-language experiments.

While the Dortmund group used machine translation, a range of different translation approaches was used by Berkeley: thesaurus lookup, "entry vocabulary module (EVM)" and machine translation. They used a combination of all three approaches as well, giving them superior performance to any of the single approaches.

7 Pool Quality

The value of the results reported from experiments such as those conducted in CLEF, and the value of the resulting test collection, depend heavily on the quality of the relevance assessments.

While a range of issues were considered to ensure the consistency and accuracy of the relevance assessments, as outlined in Section 5, concern often focuses on the completeness of the judgments. More specifically, the number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance, as would theoretically be required to calculate recall-based evaluation measures. Instead, approximate recall figures are calculated by using pooling techniques.

The assumption is that if a sufficient number of diverse systems contribute results to a pool, it is likely that a large percentage of all relevant documents will be included. Thus, instead of judging all documents, the "pool" of highly ranked documents from all the systems is judged. All unjudged documents are assumed to be irrelevant.

A main concern with this strategy is that if the number of relevant documents not detected is above a certain (low) threshold, the resulting test collection will be of limited future use for non-participants.

One way to address this issue is by determining the stability of the evaluation results when groups are treated as if they had not directly participated [12] [11]. In order to do this, all relevant documents that were detected by just one group are removed (marked irrelevant), and then the evaluation measures are recalculated. When there are few such "unique relevant documents", and if little variation in evaluation measures and rankings is observed, then the judgments are likely to be sufficiently complete and stable.

Analysis of the multilingual CLEF 2000 pool showed a mean change of only 0.80% in average precision (standard deviation of ±1.15%, maximum 5.99%). This compares favorably with the figures published for TREC ad-hoc collections [10], and is a marked improvement over previous TREC CLIR tracks. The ranking was also extremely stable, reinforcing the indications that the CLEF collection is well-suited for future testing and benchmarking activities by CLIR system developers. For a more detailed discussion see [3].

8 Plans for 2001

The second CLEF campaign will continue to offer the same basic tasks, with the following differences:

- The multilingual task will be extended with a fifth important core language, Spanish (previously only offered as a topic language).
- Bilingual retrieval will still be offered for X > E language combinations, but additionally, a Dutch collection will be added, and bilingual retrieval results for X > D will be accepted. As Dutch is a less widely used language compared to the core set, this will provide an interesting task for groups that want to adapt their system to a new, “unknown” language. In order to encourage participation, simple Dutch-English word lists, a Dutch stemmer and Dutch stopwords will be made available to participants.
- The same Dutch collection will also be used for a Dutch monolingual retrieval task. This, and a new Spanish monolingual task, bring the number of languages for monolingual retrieval experiments to five.
- In addition to the new document collections for Spanish and Dutch, the French, German and Italian collections will also be enlarged. In total, the CLEF collections contain approximately 1 million documents for 2001.
- The domain-specific GIRT task will continue essentially unchanged. Russian topics will also be provided this year, in order to encourage groups to make use of a Russian/German translation list that is available for this collection.

As in 2000, additional topic translations from interested organizations will be accepted and included in the evaluation.

Specifically, a new cooperation with NII (National Institute of Informatics) in Japan will allow to produce Japanese translations of the CLEF 2001 topics. NII will participate in the topic development/-selection as well, with the aim of having topics that can be interchanged between the NTCIR and CLEF evaluations. Further coordination and discussion between the CLEF and NTCIR initiatives will take place, and will be very beneficial for future campaigns.

Additionally, the topics will be distributed in Chinese. The translations are provided by the National Taiwan University. The availability of both Japanese and Chinese topics will hopefully open up a whole range of new language combinations, and will make the activity more attractive for Asian research groups.

A working group is also examining the feasibility of organizing an experimental interactive cross-language system evaluation task.

CLEF 2001 will start in early spring with the release of the document collections and topics. The campaign will end in September with a two-day workshop meeting in Darmstadt, Germany (in conjunction with the Fifth European Conference on Digital Libraries ECDL 2001).

The CLEF steering committee will be joined by University of Hildesheim, which takes over the han-

dling of French, and University of Twente, responsible for Dutch.

More information about CLEF can be found on the CLEF website: www.clef-campaign.org

9 Acknowledgements

The topic sets were all prepared by independent groups, i.e. by groups not participating in the system evaluation tasks. The main topic sets (E, F, G, I) and the Spanish topics were prepared by the project partners. Here, we should like to express our gratitude to the following organizations who voluntarily engaged translators to provide additional topic sets, working on the basis of the set of source topics: the DRUID project for the Dutch topics; the Department of Information Studies (University of Tampere, Finland) and the UTA Language Centre for the Finnish topics; SICS Human Computer Interaction and Language Engineering Laboratory for the Swedish topics.

We also gratefully acknowledge the support of all the data providers and copyright holders, and in particular: The Los Angeles Times, for the English data collection; Le Monde S.A. and ELDA: European Language Resources Distribution Agency, for the French data; Frankfurter Rundschau, Druck und Verlagshaus Frankfurt am Main; Der Spiegel, Spiegel Verlag, Hamburg, for the German newspaper collections; InformationsZentrum Sozialwissenschaften, Bonn, for the GIRT database; Hypersystems Srl, Torino and La Stampa, for the Italian data; Schweizerische Depeschenagentur (SDA) and Associated Press (AP) for the newswire data of the training collection.

Without their help, this evaluation activity would be impossible.

We should like to thank the ECDL 2000 Conference organizers for all their assistance in the organization of the CLEF Workshop, and in particular Caroline Hagège and Nuno Mamede, (Local Coordinators) and Eulália Carvalho, and José Luis Borbinha (ECDL Chair).

References

[1] M. Braschler, J. Krause, C. Peters, P. Schäuble. Cross-Language Information Retrieval (CLIR) Track Overview. *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, 1998.

[2] M. Braschler, C. Peters, P. Schäuble. Cross-Language Information Retrieval (CLIR) Track Overview. *Proceedings of the Eighth Text Retrieval Conference (TREC8)*, 1999.

[3] M. Braschler. CLEF 2000 – Overview of Results. *Proceedings of CLEF 2001*. Lecture Notes in Computer Science, Springer. To appear.

[4] S. Chaudiron, L. Schmitt. Amaryllis: An evaluation-based program for text retrieval in French. *In Proceedings of the 2nd International Conference on Language Resources &Evaluation (LREC 2000)*, 2000.

[5] D. Harman. The TREC Conferences. In *Proceedings of HIM '9*5. Reprint in Sparck-Jones and Willett (eds.): *Readings in Information Retrieva*l. Morgan Kaufmann Publishers, 1995.

[6] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, S. Hidaka, J. Adachi. The NTCIR Workshop: The First Evaluation Workshop on Japanese Text Retrieval and Cross-Language Information Retrieval. In *Proceedings of the 4th International Workshop on Information Retrieval with Asian Language*s, 1999.

[7] I. Moulinier, A. McCulloh, E. Lund. West Group at CLEF2000: Non-English Monolingual Retrieval. *Proceedings of CLEF 2001*. To appear.

[8] C. Peters, C. Thanos. DELOS: A Network of Excellence for Digital Libraries: Promoting and Sustaining Digital Library Research and Applications in Europe. *Cultivate Interactive, No. 1, Web Magazine of the EC Digicult Programme*, 2000.

[9] P. Schäuble, P. Sheridan. Cross-Language Information Retrieval (CLIR) Track Overview. *Proceedings of the Sixth Text Retrieval Conference (TREC6)*, 1997.

[10] E. M. Voorhees, D. Harman. Overview of the Eighth Text REtrieval Conference (TREC-8). *Proceedings of the Eighth Text Retrieval Conference (TREC8)*, 1999.

[11] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697-716, 2000.

[12] J. Zobel. How reliable are the results of large-scale information retrieval experiments? *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.