TREC-7 Experiments at the University of Maryland

Douglas W. Oard
Digital Library Research Group
College of Library and Information Services
University of Maryland, College Park, MD 20742
oard@glue.umd.edu

Abstract

The University of Maryland participated in three TREC-7 tasks: ad hoc retrieval, cross-language retrieval, and spoken document retrieval. The principal focus of the work was evaluation of merging techniques for cross-language text retrieval from mixed language collections. The results show that biasing the merging strategy in favor of documents in the query language can be helpful. Ad hoc and spoken document retrieval results are also presented.

1 Introduction

The principal goal of the University of Maryland's participation in the Seventh Text REtrieval Conference (TREC-7) was to evaluate the performance of alternative merging strategies for Cross- Language Information Retrieval (CLIR) from mixed language collections. The Logos machine translation system[1] was used in a fully automatic mode for query translation, and PRISE from the National Institutes of Standards and Technology was used for all runs. We participated in the Ad Hoc task as well in order to gain experience with PRISE, and we also used PRISE for Spoken Document Retrieval (SDR) track runs. No manual processing was done, and all of our runs were submitted in the automatic category.

2 Cross-Language Information Retrieval

As typically formulated, interactive information retrieval involves at least three stages: query formulation, searching the document collection using the query to identify a set of possibly relevant documents, and selection of desirable documents by the user [1]. CLIR potentially adds complexity to each stage. The focus of our work in the CLIR track at TREC has been on fully automatic techniques that are appropriate for the middle stage, finding possibly relevant documents when the query and document may not be in the same language. At TREC-6 we compared query translation and document translation approaches, finding little difference in overall retrieval effectiveness [2]. Query translation is the more efficient of the two approaches, and that advantage is magnified when documents in several languages are present in the collection as is the case in the TREC-7 CLIR track. We have thus chosen query translation as the basis for our experiments this year.

In TREC-6 we learned that language-specific processing such as stemming can have a substantial effect on retrieval effectiveness, a lesson that others have learned before [3]. In those experiments we used Inquery version 3.1, which was capable of stemming English but not German. With long queries, we observed that indexing English translations of German documents (with stemming) gave better results than indexing the documents in German (without stemming or compound splitting). We initially believed that this gave evidence favoring document translation. After seeing the same effect on English (AP) documents, however, we now believe that the differences resulted from a failure to perform stemming or compound splitting in German.

[1]Logos Corporation, 111 Howard Boulevard, Suite 214, Mount Arlington, NJ 07856 USA

2.1 Experiment Design

The TREC-7 CLIR track requires that documents in German, French, Italian, and English be processed. Since we had reliable *a priori* knowledge of the language contained in each portion of the collection, we used that knowledge to select appropriate language-specific processing. Documents in the AP collection were treated as English, documents in the “French SDA” collection were treated as French, documents in the “Italian SDA” collection were treated as Italian, and documents in both the “German SDA” and the “NZZ” collections were treated as German.

The Logos machine translation system can translate from English to French, German, Italian and Spanish. Our queries were thus based on the English topics. We began by translating the queries from English into each other language, using the Logos system in a fully automatic mode with no application-specific additions to the lexicon or semantic rules. We then formed title queries from the words in the title field, and long queries from every topic word except SGML markup, the contents of the query number field, and the terms “Description:” and “Narrative:” that appear in every query.

PRISE includes the Porter stemmer for English, a German stemmer implemented by Martin Braschler, and a French stemmer implemented by Jacques Savoy. We did not have an Italian stemmer, and no compound splitting was performed in any language. The stopword list from Inquery version 3.1 was used in English, and degenerate stopword lists were used in the other languages (“le” in French, “die” and “dir” in German, and “du” in Italian — PRISE choked if the stopword list was empty). No stop-structure removal was performed. Separate PRISE indexes were built for each language, with the German index covering both the “German SDA” and the “NZZ” collections. Index construction required between two and four hours on a dedicated Sparc 20, depending on the number of documents in each language, and retrieval results for all 25 queries were typically computed in a few minutes (varying slightly with query length and whether stopwords were used). In our official runs we inadvertently omitted the 1989 and 1990 AP documents from the English index, and this adversely affected our results. That has been corrected in the results reported here.

Vector space text retrieval systems such as PRISE typically produce retrieval status values that lack comparability across collections, so rank-based merging generally outperforms strategies based on retrieval status values. Voorhees demonstrated that giving more weight to collections that are historically more productive can yield better results that a uniform rank-based merging strategy [5]. In TREC-6 we observed that machine translation of German queries into English achieved 56% of the average precision that was observed when English queries were used for monolingual retrieval, and we expected that a strategy which selected more documents from the English collection than from the other three collections would perform well. We thus implemented a uniform weighted merge in which the top N documents were selected (without replacement) from English every time the top document was selected (without replacement) from each of the other languages.

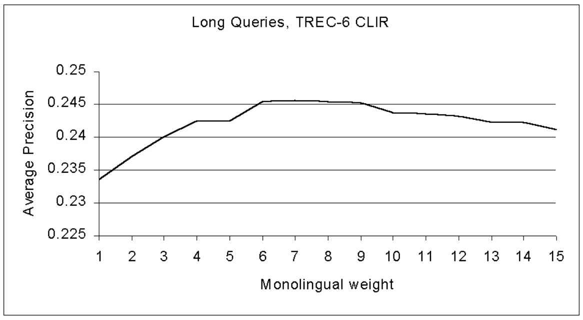

Figure 1: The effect of varying N on TREC-6 CLIR long queries.

2.2 Parameter Selection

In order to get some idea of a reasonable range for N, we tried our strategy on the TREC-6 CLIR collection. The TREC-6 document collection is a substantial subset of the TREC-7 document collection, lacking only the Italian SDA documents. The limited pool of participating systems may, however, have limited the completeness of the TREC-6 relevance judgments in some languages, and there were some differences in the way queries were formulated in the two evaluations. In our official runs the omitted 1990 and 1991 portions of the AP collection reduced the performance of the English collection. Not surprisingly, $N = 1$ outperformed higher values of N under those conditions, so our two official TREC-7 CLIR submissions were produced with an even merging strategy ($N = 1$) on title (run umdxeot) and long (run umdxeof) queries. When we reran our experiments on the complete TREC-6 collection we found that weighted merging outperformed an even merging strategy by about 5% on long queries (at $N = 6$), but that no more than a 0.3% advantage could be achieved on title queries (at $N = 1.4$). Figure 1 illustrates the long-query results.

2.3 Results

When the TREC-7 CLIR relevance judgments became available we observed a similar advantage for strongly weighted merging, achieving an 9% improvement on long queries at the $N = 6$ parameter learned on the TREC-6 data and an 11% improvement on long queries at the *post hoc* optimum parameter value ($N = 9$). Weighted merging again produced only a modest improvement (2% at $N = 5$) on title queries. Figure 2 illustrates these results.

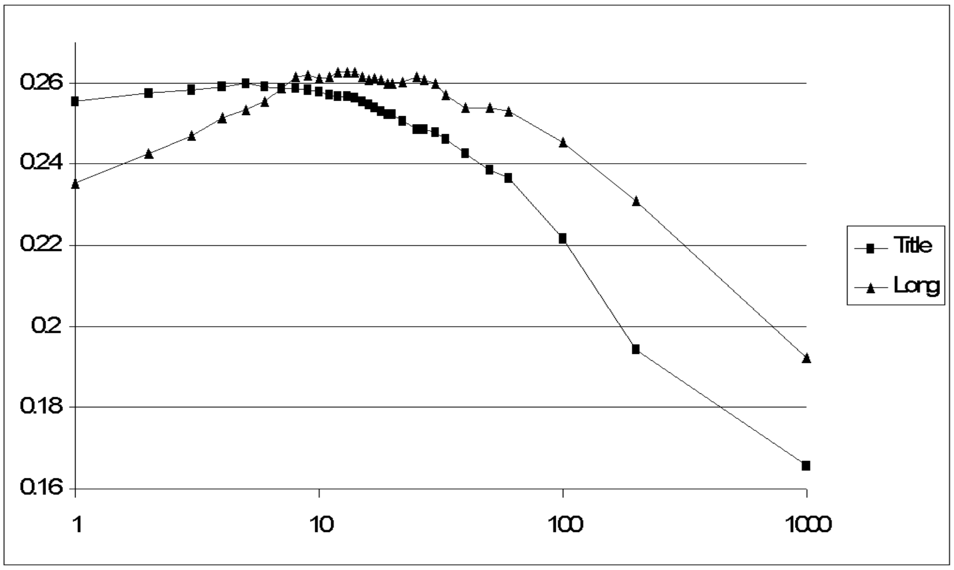

Figure 2: The effect of varying N on TREC-7 CLIR title and long queries.

We were surprised by how large the large values of N were that produced the best results for long queries and by the consistent difference between the effectiveness of weighted merging for title and long queries, so we decided to examine the monolingual performance of our system for each language pair using the TREC-7 data. Table 1 shows the uninterpolated average precision obtained when English queries were used to retrieve documents in a single language. In this case, only relevance judgments for documents in that language were considered. Some topics lack known relevant documents in some languages, so the number of queries over which the averages are calculated are shown for each language. There is some variation evident between the two query lengths in German, but no systematic differences are evident.

Table 2 shows some collection statistics. On average, nearly twice as many relevant documents are known for English as for any other language, and there are even fewer known relevant documents in the Italian collection. The average density of relevant documents is somewhat more consistent, however.

Our results suggest two factors that might be useful when selecting collection weights if a uniform merge strategy is used. The most obvious is the expected performance of each system - a monolingual system would be expected to outperform a cross-language one, for example. The second possibly useful factor is collection size, which should predict the number of relevant documents well if the collections and queries

Doc Lang	Title Queries	% of English	Long	% of English	Num of Queries
English	0.4357		0.5290		26
French	0.2827	65%	0.3420	65%	28
German	0.2265	52%	0.2311	44%	27
Italian	0.2453	56%	0.2874	54%	25

Table 1: Non-interpolated average precision with English queries for documents each language.

Doc Lang	Documents	Average Relevant	Average Density
English	242,917	60	2.5
French	141,656	35	2.5
German	251,850	33	1.3
Italian	62,359	18	2.9

Table 2: Density of known relevant documents per 10,000 documents, averaged over 28 topics.

are chosen in a way that produces similar densities of relevant documents across the collections. It is not yet clear whether the number of relevant documents is actually more important than their density, but our results suggest that a focused investigation of that issue could prove useful in this context.

A note of caution should be sounded regarding our use of the average precision measure. Our monolingual English run achieved higher precision at 5, 10, 15, 20, and 30 documents than the best merged run on both title and long queries.[2] The advantage of the merging strategy is only evident at 100, 200, 500 and 1000 documents. The average precision measure is useful because it balances precision and recall, but other measures may me more appropriate for specific applications.

3 Ad Hoc and Spoken Document Retrieval Tasks

We used our participation in the ad hoc retrieval task to become familiar with PRISE. The official run was submitted using the default term weighting strategy in PRISE, which does not do as well as the "okapi1" weights that we used for our CLIR and SDR experiments.

We are working on user interface design for information retrieval systems that provide access to large collections of recorded speech [4], and the SDR track offers an opportunity to gain additional experience with content-based retrieval using speech recognition output. Our speech recognition system was not ready in time for these runs, so we submitted results only for the baseline recognizer output. We used a modified version of PRISE for these experiments in which some changes had been made to the numerical details of retrieval status value computation, but a comparison with the original system revealed no significant differences in the ranked output. The Porter stemmer, okapi1 weights, and the Inquery stopword list were the only deviations from the default settings in the indexer. Indexing took approximately 15 minutes for each of the three runs, and batch processing of the queries was completed in under a minute per collection. The queries used were identical for each of the three runs.

4 Conclusion

We have demonstrated one useful strategy for merging retrieval results from collections in different languages. As the richness of the TREC CLIR corpus grows, we plan to exploit it to investigate more sophisticated

[2] In this case, precision values for the monolingual English runs were computed using all relevance judgments rather than those for English alone in order to produce comparable results.

strategies. We are also interested in integrating automatic language identification in order to investigate a whether applications in which the document languages cannot be reliably determined from *a priori* information will pose substantially greater challenges. The TREC CLIR corpus also provides an excellent resource for evaluating other approaches to CLIR, and we hope to use it to explore both cognate matching and corpus-based techniques.

Acknowledgments

The author is grateful to Paul Hackett and Skip Warnick for their assistance with the experiments, to NIST for their extremely helpful support as we learned to use PRISE, and to the Logos Corporation for the use of their machine translation system. This work has been supported in part by DARPA contract N6600197C8540.

References

[1] Douglas W. Oard. Serving users in many languages: Cross-language information retrieval for digital libraries. *D-Lib Magazine*, December 1997. http://www.dlib.org.

[2] Douglas W. Oard and Paul G. Hackett. Document translation for cross-language text retrieval at the University of Maryland. In *The Sixth Text REtrieval Conference (TREC-6)*. National Institutes of Standards and Technology, November 1997. http://trec.nist.gov/.

[3] Páraic Sheridan and Jean Paul Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 1996. http://www-ir.inf.ethz.ch/Public-Web/sheridan/papers/SIGIR96.ps.

[4] Laura Slaughter, Douglas W. Oard, Vernon L. Warnick, Julie L. Harding, and Galen J. Wilkerson. A graphical interface for speech-based retrieval. In Ian Witten, Rob Akscyn, and Frank M. Shipman, III, editors, *The Third ACM Conference on Digital Libraries*, pages 305–306, June 1998.

[5] Ellen M. Voorhees, Narendra K. Gupta, and Ben Johnson-Laird. Learning collection fusion strategies. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 172–179. ACM Press, July 1995.