TREC-9 Experiments at Maryland: Interactive CLIR

Douglas W. Oard,* Gina-Anne Levow,†and Clara I. Cabezas,‡
University of Maryland, College Park, MD, 20742

Abstract

The University of Maryland team participated in the TREC-9 Cross-Language Information Retrieval Track, with the goal of exploring evaluation paradigms for interactive cross-language retrieval. Subjects were asked to examine gloss translations of highly ranked documents and make relevance judgments, and those judgments were used to produce a new ranked list in which documents assessed as relevant were promoted and those assessed as nonrelevant were demoted. No improvement over fully automatic ranking was found which suggests that additional work on user interface design and evaluation metrics are required.

1 Introduction

The principal goal of our research on cross-language information retrieval (CLIR) is, of course, to build systems that are useful for a some ultimate purpose. In the Text Retrieval Conferences (TREC), ad hoc retrieval tasks such as the CLIR track are designed to model the process of an individual user searching for one or more previously unseen documents on some topic. Although some applications such as document alerting require fully autonomous operation, we are particularly interested in interactive applications in which user and machine seek to synergistically exploit the strengths of each to search more effectively together than either could in isolation. Our principal goal in TREC-9 was to begin our exploration of this synergy in the context of CLIR.

Interactive retrieval can be roughly divided into three stages: query formulation, search, and browsing. In the context of CLIR, search has received the vast majority of the attention (e.g., at the TREC, NTCIR, and CLEF evaluations). There has also been some attention given to query formulation issues such as user-assisted query translation, both in research systems and in deployed applications (c.f., http://messene.nmsu.edu/ursa/arctos). We are, however, aware of only two reported user studies that have explored issues related to cross-language browsing effectiveness. In one, Oard and Resnik adopted a classification paradigm to evaluate browsing effectiveness in cross-language applications, finding that a simple gloss translation approach allowed users to outperform a Naive Bayes classifier [1]. In

*College of Information Studies and Institute for Advanced Computer Studies, oard@glue.umd.edu
†Institute for Advanced Computer Studies, gina@umiacs.umd.edu
‡Department of Linguistics, clarac@umiacs.umd.edu

the other study, Ogden et al., evaluated a language-independent thumbnail representation in the TREC-7 interactive track, finding that the use of thumbnail representations alone resulted in even better recall at 20 documents than was achieved using English document titles [2]. The logical next step is to combine the best of both experiments, working directly on retrieval results as Ogden et al. did, while focusing on the marginal improvement over fully automatic processing as Oard and Resnik have done.

One obvious approach to this challenge would be to organize a TREC track at the intersection of the present CLIR and interactive tracks. We used the TREC-9 CLIR track for exploratory work in that direction, providing users with simple gloss translations of the retrieved documents and allowing them to improve the ranked list by moving documents that they believe to be relevant higher in the list and documents that they believe to be irrelevant lower. Since either change would improve mean uninterpolated average precision if the user's judgment were correct, we adopted the change in mean uninterpolated average precision between the automatically generated ranked list and manually corrected ranked list as a metric for assessing the effect of the user's contribution.

2 Experiment Design

We initially conducted two small pilot studies to refine our user interface and experiment procedures using graduate students from our laboratory. Five graduate students from outside our laboratory with no self-reported Chinese language skills were then recruited as subjects for the experiments reported below. All were proficient or native speakers of English, and all reported experience with document retrieval that was limited to the use of search engines such as AltaVista or Google. We offered to buy pizza for our subjects upon completion of the experiment session, but all of them declined our offer!

The experiment was conducted during a single session. A Web-based user interface was designed specifically to support these experiments. The subjects were first provided with an opportunity to become familiar with the system and the experiment protocol using two topics from the TREC-5/6 Chinese collection. For the experiment itself, we divided the 25 TREC-9 CLIR topics into sets of five, and assigned one set to each subject (i.e., topics CH55-CH59 to user 1, topics CH60-CH64 to user 2, etc.). The subjects' task was to sequentially perform a search using a query that was automatically derived from the topic description and then judge the relevance of as many documents as time allowed based on their understanding of the full topic description. This process involved four steps for each topic:

- Topic selection. Subjects were instructed to click on the appropriate topic number from an initial selection page, resulting in display of the full topic description. After all subjects completed reading the topic description, the subjects were instructed to select the 'Search' button.
- Document selection. Selecting the search button resulted in addition of a ranked list of document titles to the same window. A new screen with a list of the top fifty glossed document titles would appear. For each document in the list, the title and three radio buttons ('Relevant', 'Not relevant,' and 'No response') were presented. Subjects

were given five minutes from the time he/she hit 'search' to evaluate the relevance of as many documents as possible. We limited the displayed portion of the ranked list to fifty documents because that was far greater than any subjects in our pilot study could evaluate in 5 minutes. The subject could look at the translation of any document by clicking on the gloss of the document title. The first time this was done, a second window was created in which the glossed translation of the document would be displayed. The document selection window remained visible in order to facilitate recording the relevance judgment and selecting the next document.

- Relevance judgment. In the document display page, gloss translations were provided using a technique described below. If the subject was able to decide on the relevance of a document based on the gloss translation, he/she could select from 'Relevant' or 'Not Relevant' buttons. The third option, 'No Response,' was initially automatically selected, and would be left selected if no judgment could be made.

- Recording relevance judgments. After five minutes, subjects were instructed to manually select a function to submit their relevance judgments. The judgments were then recorded, and the topic selection screen was displayed to begin a new search.

We used a document translation strategy for CLIR, which is a natural choice when browsable translations must be immediately available. Our tools are designed to work with the GB character set, so we used the commercial NJStar Communicator package to convert from Big5 to GB. Fully automatic segmentation was then performed using the ch_seg package from New Mexico State University. We performed a termbyterm translation from Chinese into English using a balanced translation strategy to produce exactly two terms for each Chinese term in the original documents. For Chinese terms with no known translation, the untranslated Chinese term was converted to pinyin (without tone) and generated twice. For Chinese terms with one known translation, that translation was generated twice. Terms with two or more known translations resulted in generation of each of the "best" two translations once. The Brown Corpus served as a side collection to sort the alternates. In prior experiments we have found that such a balanced translation strategy significantly outperforms a more naive (unbalanced) technique in which all known translations are included because it avoids over-weighting terms that have many translations. The resulting English collection was then indexed using Inquery (version 3.1p1), with the default kstem stemmer and the default English stopword list.

2.1 Cross-Language Browsing

We displayed the same 2-best balanced translations to the user that were used for automatic search. To improve readability we grouped alternate translations using parentheses and showed the most common translation first using a bold font. Query terms were highlighted in red in an effort to help guide the user's eye to relevant passages. Our baseline for retrieval effectiveness was the mean uninterpolated average precision achieved by the automatically generated ranked list. We used this as a basis for comparing three reranking approaches: Maximum, Partial, and Balanced.

In Maximum reranking, documents marked by the user as relevant based on their examination of gloss translations were moved to the top of the list (position 1) and documents marked as irrelevant were moved to the bottom (position 1000), with the relative order between documents marked in the same way preserved. The remaining documents (labeled 'No Response') appeared in their original (automatically computed) order between those two sets. If the user's relevance judgments were perfect (perfection here being defined by TREC assessors, of course), Maximum reranking would produce the greatest possible improvement in mean uninterpolated average precision. At least three possible sources of error are possible however:

- The user might disagree with the TREC assessor's judgment of relevance, even if they fully understood the document.
- The user might not be able to accurately assess the relevance of the document to the topic based on the gloss translation.
- The user might select the wrong button by mistake.

Since moving a document all the way to the wrong end of the list could mask the beneficial effect on our metric of several correct assessments, we also tried two more conservative strategies. In Partial reranking, documents marked as relevant were moved halfway to the top of the list. For example, a document in position 11 would be moved to position 6 if the user marked it as relevant. Because of the way that uninterpolated average precision is computed, achieving a similar effect form demoting nonrelevant documents requires a the documents be moved further—we thus continued to move documents marked as 'not relevant' to the bottom of the ranked list (position 1000).

It is not clear how far down the list a document marked as nonrelevant should be moved, so we also tried a third approach that we called Balanced reranking. As with Partial reranking, in Balanced reranking we moved documents judged as relevant up by 50% of the distance to the top. When moving documents down, however, we limited their demotion to 10 times as far from the top of the list as they were in the automatically computed list. For example, a document in position 2 would move to position 11.

3 Results and Analysis

As shown in Table 2, we found that the best effectiveness of these four conditions was achieved by the Baseline (completely automatic) condition, although the differences were not statistically significant at $p < 0.05$ by a paired two-tailed t-test. We performed a query-by-query analysis to better understand this result and observed two important effects. First, as table 1 shows, when relevant documents are moved down the list, there can be a severe adverse impact on retrieval effectiveness. This suggests that we should adopt a more conservative strategy towards demotion, but we have not yet had the time to explore that idea.

As it turns out, both of these topics were assigned to the same subject. On closer inspection, it is clear that our results seem to be adversely affected by a single subject.

Topic	Relevant Docs	Baseline	Partial	Maximum	Balanced
CH60	4	1	0.0031	0.0031	0.6429
CH62	1	0.5	0.0011	0.0011	0.025

Table 1: Degradation in average precision when the Baseline does well.

Topic	Relevant Docs	Baseline	Partial	Maximum	Balanced
CH60	4	1	0.0031	0.0031	0.6429
CH62	1	0.5	0.0011	0.0011	0.025

Table 2: Official results and contrastive results with topics CH60-CH64 removed.

Each subject inspected the results of five queries and, due to time constraints, each query was inspected by only a single subject. As table 2 shows, when the topics presented to that subject (CH60-CH64) are excluded, virtually all of the differences are removed. We observed that the subject in question had judged two to three times as many retrieval results as other subjects, and had marked the vast majority as not relevant, even when the title alone seemed to us to provide explicit evidence that the document was indeed on topic. These judgments are thus highly suspect.

While the susceptibility to individual performance might be mitigated by averaging the effect of several subjects, no subject obtained reliable improvements with any of our three reranking techniques. One factor that we believe contributes significantly to these problems is that the assessment process itself is quite difficult for novice users to perform. NIST assessors are generally highly trained analysts, but our subjects were (be design) novice users. We believe that the interactive reranking approach has potential as a way of assessing user-system synergy, but clearly several issues of user training and study design remain to be overcome.

4 Conclusion

[being written now]

Acknowledgments

The authors are grateful to Bill Ogden for several helpful discussions about evaluation of interactive CLIR. This work was supported in part by DARPA contract N6600197C8540 and DARPA cooperative agreement N660010028910.

References

[1] Douglas W. Oard and Philip Resnik. Support for interactive document selection in cross-language information retrieval. *Information Processing and Management*, 35(3):363–379, July 1999.

[2] William Ogden, James Cowie, Mark Davis, Eugene Ludovik, Hugo Molina-Salgado, and Hyopil Shin. Getting information from documents you cannot read: An interactive cross-language text retrieval and summarization system. In *Joint ACM DL/SIGIR Workshop on Multilingual Information Discovery and Access*, August 1999. http://www.clis.umd.edu/conferences/midas.html.