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. Participants 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 is 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 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 (e.g, 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 interactive document selection by cross-language searchers. In one, Oard and Resnik adopted a classification paradigm to evaluate browsing effectiveness in cross-language applications,

*Human-Computer Interaction Laboratory, 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

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **2006**	2. REPORT TYPE	3. DATES COVERED **00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE **TREC-9 Experiments at Maryland: Interactive CLIR**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,Institute for Advanced Computer Studies,Department of Computer Science,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**7**	

finding that simple gloss (i.e., word-by-word) translations allowed users to outperform a Naive Bayes classifier [3]. In 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 [4]. 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 nonrelevant 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 participants 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 participants 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 participants 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 participant (i.e., topics CH55-CH59 to participant 1, topics CH60-CH64 to participant 2, etc.). The participants' 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. Participants were instructed to click on the appropriate topic number from an initial selection page, resulting in display of the full topic description. After all participants completed reading the topic description, the participants 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. For each document in the list, a gloss translation of the title and three radio buttons (‘Relevant’, ‘Not relevant,’ and ‘No response’) were presented. The participant was 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 more than any participant in our pilot study could evaluate in 5 minutes. The participant could look at the translation of any document by clicking on the translated title. The first time this was done, a second window was created in which the translation was displayed. The document selection window remained visible in order to facilitate recording the relevance judgment and selecting the next document.

- Relevance judgment. If the participant was able to decide on the relevance of a document based on either the translated title or the translated document, he/she could select the ‘Relevant’ or ‘Not Relevant’ button for that document in the document selection window. The third option, ’No Response,’ was initially automatically selected, and could be left selected if no judgment could be made.
- Recording relevance judgments. After five minutes, participants were instructed to manually select a button 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 term–by–term 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 candidate translations in decreasing order of English usage (see [1] for additional details on this process). 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.

We displayed the same 2-best balanced translations to the user. 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.[1] We used this as a basis for comparing three reranking approaches in our official run: Maximum, Partial, and Balanced.

[1] Our baseline run is unofficial, having been scored locally using the published relevance judgments.

In Maximum reranking,[2] documents marked by the participants as relevant 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 participant'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 participant might disagree with the TREC assessor's judgment of relevance, even if they fully understood the document.
- The participant might not be able to accurately assess the relevance of the document to the topic based on the gloss translation.
- The participant 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,[3] 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 participant marked it as relevant. Because of the way that uninterpolated average precision is computed, achieving a similar effect from demoting nonrelevant documents requires that 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 variant on Partial reranking that we called "Partial2".[4] As with Partial reranking, in Partial2 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 1, 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 2 shows, when relevant documents are moved down the list, there can be a severe adverse impact on retrieval effectiveness, as these results on these two topics demonstrate. This suggests that we should adopt a more conservative strategy towards demotion.

[2] Official run 'TB.'

[3] Official run 'mixed.'

[4] Official run 'percent.'

	Baseline	Maximum	Partial	Partial2
All topics	0.2477	0.1710	0.1801	0.2183
Without CH60-CH64	0.1947	0.1803	0.1917	0.1916

Table 1: Official results and contrastive results with one participant removed.

Topic	Relevant Docs	Baseline	Maximum	Partial	Partial2
CH60	4	1	0.0031	0.0031	0.6429
CH62	1	0.5	0.0011	0.0011	0.025

Table 2: Degradation in average precision when the Baseline does well.

As it turns out, both of these topics were assigned to the same participant. On closer inspection, it is clear that our results seem to be adversely affected by a single participant. Each participant inspected the results of five queries and, due to time constraints, each query was inspected by only a single participant. As table 1 shows, when the topics presented to that participant (CH60-CH64) are excluded, virtually all of the differences are removed. We observed that the participant in question had judged two to three times as many retrieval results as other participants, 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, and in future studies it would clearly be desirable to assign the same topic to more than one participant [2].

We conducted some additional *post-hoc* analysis to find the optimum way of using the relevance judgments that we obtained. We grouped the relevance judgments into four categories:

TR Judged by the user as relevant based on the title

TN Judged by the user as not relevant based on the title

DR Judged by the user as relevant based on the document text

DN Judged by the user as not relevant based on the document text

For each category (and for the full set of user judgments), we computed the mean average precision for what we call Balanced reranking, using the following formula:

$$R' = \lfloor R(1-\Delta) \rfloor \tag{1}$$

$$R' = \lceil \frac{R}{1-\Delta} \rceil \tag{2}$$

where R' is the new rank, R is the original rank and Δ is a number between 0 and 1 that specified the increment size. Equation (1) is used for upward movement of documents judged to be relevant and equation (2) is used for downward movement of non-relevant documents.

We call this Balanced reranking because moving a document down by an increment of size Δ and then back up by an increment of size Δ would return it to its original position (except as influenced by roundoff errors). We tried every value for Δ between 0.0 and 1.0 in increments of 0.05.

Figure 1 shows the results of this *post hoc* analysis. None of the judgment subsets or values for Δ produced more than a 1% relative improvement in uninterpolated mean average precision. For higher values of Δ, it does appear that judgments based on examination of the full text of a glossed document were more reliable than judgments based on examination of the glossed title alone. When only titles were observed, the results suggest that decisions that a document was relevant may have been more reliable than decisions that a document was not relevant. Both results should be interpreted with caution, however. It is not possible to conclude that glossed documents are more informative than glossed titles, for example, because other factors (e.g., more careful participants) might explain the observed relationship equally well. Similarly, the relative effect of relevant and not-relevant judgments is sensitive to both the reliability of the judgments and the design of the Balanced reranking technique.

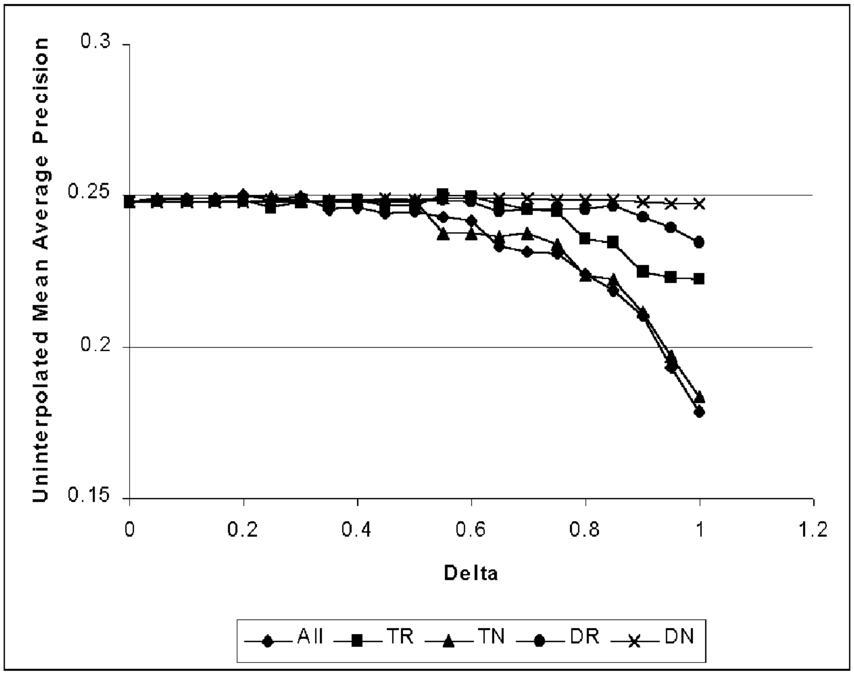

Figure 1: Balanced reranking with various subsets of the user judgments.

One important qualitative observation that we made is that our participants seemed to find the assessment process itself to be fairly difficult. NIST assessors are generally highly trained analysts, but our participants were (by design) novice users. If we were to provide more training before the study and more time to perform assessments, we might be able to minimize the effect of this factor.

4 Conclusion

We have tried to study interactive CLIR in the context of the present TREC CLIR track. The study reported here, with only a single participant for each block of five queries, a limit of five minutes to examine a full set of translations, and only a single interface design, is clearly of such limited scope that it would be difficult to draw any firm conclusions. Viewed as a pilot study for a more comprehensive interactive cross-language TREC evaluation, it offers some useful insights into the challenges of conducting such evaluations that we expect will inform our future work. Interactive reranking may ultimately have potential as a way of assessing user-system synergy, but clearly several issues of user training and study design remain to be worked out.

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] Gina-Anne Levow and Douglas W. Oard. Translingual topic tracking with PRISE. In *Working Notes of the Third Topic Detection and Tracking Workshop*, February 2000.

[2] Douglas W. Oard, Gina-Anne Levow, and Clara I. Cabezas. CLEF experiments at Maryland: Statistical stemming and backoff translation. In Carol Peters, editor, *Proceedings of the First Cross-Language Evaluation Forum*. 2001. To appear. http://www.glue.umd.edu/∼oard/research.html.

[3] 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.

[4] 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.