iCLEF 2003 at Maryland: Translation Selection and Document Selection

Bonnie Dorr,* Daqing He, Jun Luo, Douglas W. Oard, Richard Schwartz,**
Jianqiang Wang, David Zajic

Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA
{dorr,daqingd,jun,oard,wangjq,dmzajic}@umiacs.umd.edu, schwartz@bbn.com

Abstract. Maryland performed two sets of experiments for the 2003 Cross-Language Evaluation Forum's interactive track, one focused on interactive selection of appropriate translations for query terms, the second focused on interactive selection of relevant documents. Translation selection was supported using possible synonyms discovered through back translation and two techniques for generating KeyWord In Context (KWIC) examples of usage. The results indicate that searchers typically achieved a similar search effectiveness using fewer query iterations when interactive translation selection was available. For document selection, a complete extract of the first 40 words of each news story was compared to a compressed extract generated using an automated parse-and-trim approach that approximates one way in which people can produce headlines. The results indicate that compressed "headlines" result in faster assessment, but with a 20% relative reduction in the $F_{\alpha=0.8}$ search effectiveness measure.

1 Introduction

The goal of Cross-Language Information Retrieval (CLIR) is to help searchers find relevant documents even when their query terms are chosen from a language different from the language in which the document are written. For the Cross-Language Evaluation Forum's (CLEF) interactive track (iCLEF), we have been exploring the most challenging such case—when the searcher's knowledge of the document language is so limited that they would be unable to formulate queries or recognize relevant documents in that language. Our challenge is thus to provide searchers with translation tools that are tuned to the tasks that they must perform during an interactive search process: choosing query terms and recognizing relevant documents. For iCLEF 2003, we ran two sets of experiments, one focusing on each aspect of this challenge.

Although query formulation for CLIR might appear on the surface to be similar to query formulation in monolingual applications, the iterative nature

* Authors names in alphabetical order, affiliations as shown except as noted
** BBN Technologies, 9861 Broken Land Parkway, Suite 156, Columbia, MD 21046

of the interactive search process results in a difference that is of fundamental importance. When searching within the same language, searchers seek to choose terms that authors actually used in documents. That is clearly not possible when using query translation for CLIR; instead, searchers must choose query terms that the *system* will translate into terms that were used by authors. In CLIR, searchers interact with a translation system, and not (directly) with documents. We know much about how users interact with documents in an iterative search process, but we know comparatively little about how they interact with translation systems for this purpose. That was the focus of our first set of iCLEF experiments this year.

Document selection might initially seem to be a more straightforward task; here, we want to support the searcher's decisions about relevance with the best possible translations. But, again, the iterative nature of the search process complicates the question; for interactive searching, assessment speed can be as important as assessment accuracy. The reason for this is simple; longer assessment times would mean either fewer documents read or few iterations performed, both of which can adversely affect the effectiveness of a search process. We therefore need translations that can be assessed rapidly and accurately. That was the focus of our second set of iCLEF experiments this year.

The remainder of this paper is organized as follows. We first introduce the design of our query translation experiments, present our results, and draw some conclusions. We then do the same for our document selection experiments. Finally, we conclude with some ideas for future work that have been inspired by the experiments reported in this paper.

2 Query Translation Experiments

Ultimately, we are interested in learning whether a searcher's ability to interactively employ a CLIR system based on query translation can be improved by providing greater transparency for the query translation process. Our experiments last year for iCLEF 2002 demonstrated the utility of user-assisted query translation in an interactive CLIR system [4]. With user-assisted query translation function, a CLIR system provides an additional interaction opportunity where the searcher can optionally select (or remove) translations of individual query terms before and/or after viewing the retrieved documents. Our experiment results demonstrate that the average scores of $F_{\alpha=0.8}$ obtained from searches on a system with user-assisted query translation appeared better than that of a system without it. However, due to the small sample size (four searchers), we did not obtain statistical significance. In addition, we believe that the quality of the search also lies in the whole process of the search, rather than the search outcome alone. This motivated us to design an iCLEF 2003 experiment to explore the following questions:

1. What strategies do searchers apply when formulating their initial query, and when reformulating that query? Would the availability of a translation selection function lead searchers to adopt different strategies? Can we observe any

relationship between subject knowledge, search experience, or other similar factors on the choice of strategies in each condition?

2. Is there a statistically significant difference in search effectiveness (as measured by $F_{\alpha=0.8}$) between searches performed using a system with user-assisted query translation and a system that lacks that capability? Formally, we sought to reject the null hypotheses that there is no difference between the $F_{\alpha=0.8}$ for the manual and automatic conditions that are defined below.

2.1 System Design

We used the Maryland Interactive Retrieval Advanced Cross-Language Engine (MIRACLE) for the query translation experiments reported in this paper. MIRACLE is an improved version of the CLIR system that we used for iCLEF 2002. MIRACLE has recently evolved rapidly during the DARPA Surprise Language Exercise, but at the time of our iCLEF experiments the basic architecture of the system and the layout of the user interface were quite similar to that of our iCLEF 2002 system (see Figure 1). The system supports two conditions, an "automatic" condition, with a design similar to that of present Web search engines, and a "manual" condition, in which the user can participate in the construction of a translated query.

MIRACLE uses the InQuery text retrieval system (version 3.1p1) from the University of Massachusetts to implement Pirkola's structured query technique (which has been shown to be relatively robust in the presence of unresolved translation ambiguity) [9]. For the automatic condition, Pirkola's method is applied over all known translations; for the manual condition, only selected translations are used. A backoff translation strategy is used when the term to be translated is not known; first the term is stemmed, if translation still fails then a stemmed version of the term list is also used. This serves to maximize the coverage of the bilingual term list [8].

Since we are interested in the case in which searchers have no useful knowledge of the document language, we must provide the user with some evidence about the meaning of each translation in the manual condition. Optimally, we would like to provide query-language definitions of each document language-term. Although dictionaries that contain such definitions do exist for some language pairs, they are relatively rare in print, and extremely rare in an accessible electronic form. Therefore, we present searchers with as many of the following three sources of evidence as can be constructed for each term:

The document-language term. For languages in the same writing system, this can sometimes be informative, since some translations may be cognates or loan words that the searcher can recognize. For example, one Spanish translation of "ball" is "fiesta;" clearly that is the "party" sense of ball (as in "to attend a ball").

Possible synonyms. Translations are cross-language synonyms, so round-trip translation can reveal useful synonyms. For example, one possible translation

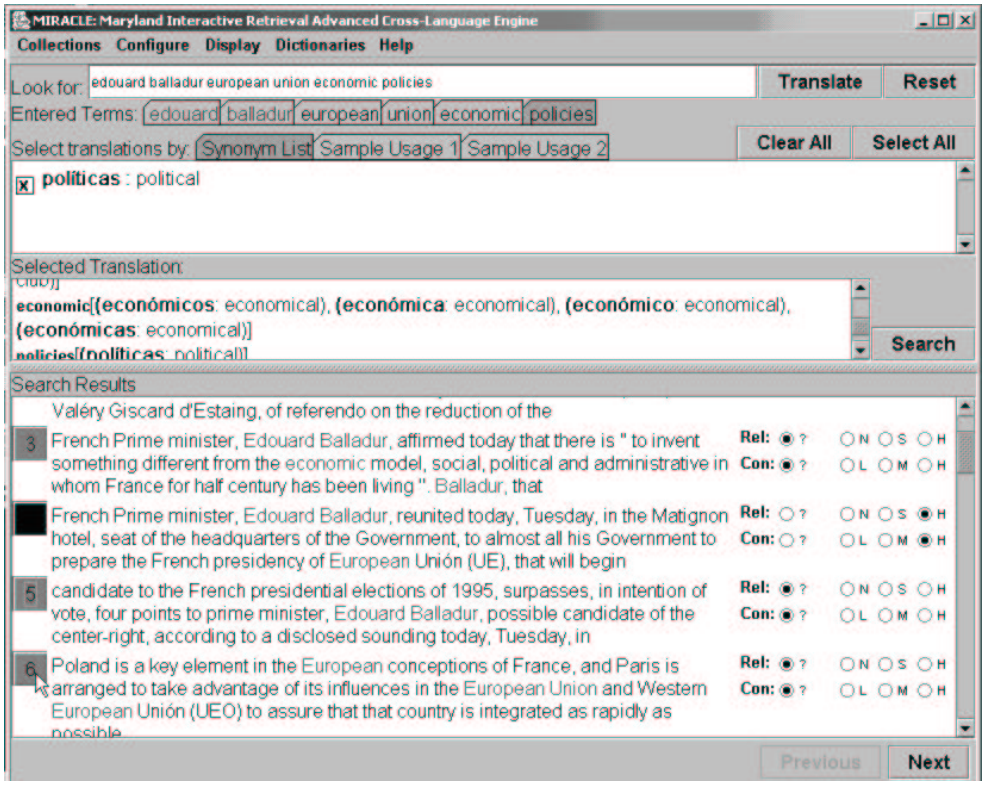

Fig. 1. The MIRACLE user interface for iCLEF 2003, manual condition.

of "bank" into Spanish is "ribera." The following translations can be found for "ribera:" bank, shore, seashore, riverside, and waterfront. This leaves little question that "ribera" does not refer to a financial sense of "bank."

Examples of usage. For iCLEF 2002, we found examples of usage (which we call "KeyWord in Context, or KWIC) in word-aligned parallel text. While examples found in this way are generally correct and often informative, the size and scope of available parallel text collections may not be sufficient to find examples for some terms in this way. We therefore added a second way of constructing KWIC examples of usage that we call KWIC-IR to our iCLEF 2003 system.

KWIC-IR The searchers in iCLEF 2002 liked KWIC based on parallel text because it provided more context information than the possible synonyms that result from round-trip translation. We therefore sought to extend KWIC to additional cases using an approach based on round-trip translation and a monolingual English text collection. It is straightforward to obtain "back translations" once we have a bilingual term list, and it is much easier to obtain a large and representative English text collection than it would be to assemble a compara-

ble amount of parallel text for every language pair that might be of interest. KWIC-IR works as follows:

- Given a query term e and one of its translations s, we can obtain a set of back translations;
- For each back translation bte_i, search the monolingual English text collection C to obtain a set of sentences containing bte_i.
- Merge all the sentences of all the back translations to build a sentence pool P;
- After removing stopwords, select representative terms from the sentence pool P by using $tf * idf$ scheme. Here tf is defined as the frequency of the term appearing in P; and idf is the inversion of document frequency, which is defined as how many sentences containing the term out there in the whole text collection C. The selection algorithm is tuned to select terms that co-occur sharply in the context of these back translations, (which include, of course, the query term e), thus preferring terms that co-occur with e that are strongly related to many translations of s;
- Search the monolingual English text collection to obtain a set of sentences containing the query term e;
- Use the set of context words to rank the sentences containing e, and pick up the top one as the KWIC-IR example for s.

Our initial testing indicated that this approach could generate reasonable examples of usage much of the time; our iCLEF 2003 experiments provided our first opportunity to try it in the context of an extrinsic (task-centered) evaluation.

2.2 Experiment Design

We followed the standard protocol for iCLEF 2003 experiments. Searchers were sequentially given eight topics (stated in English), four using the manual condition, and four using the automatic condition. Presentation order for topics and system was varied systematically across searchers as specified in the track guidelines. After an initial training session, they were given 10 minutes for each search to identify relevant documents using the radio buttons provided for that purpose in our user interface. The searchers were asked to emphasize precision over recall (by telling them that it was more important that the document that they selected be truly relevant than that they find every possible relevant document). We asked each searcher to fill out brief questionnaires before the first search (for demographic data), after each search, and after using each system. Each searcher used the same system at a different time, so we were able to observe each individually and make extensive observational notes. We also conducted a semi-structured interview (in which we tailored our questions based on our observations) after all searches were completed.

We conducted a small pilot study with a single searcher (umd00) to exercise our new system and refine our data collection procedures. Eight searchers

(umd01-umd08) then performed the experiment using the eight-subject design specified in the track guidelines.[1] We submitted all eight runs (umd01-umd08) for use in forming relevance pools.

Resources We chose English as the query language and Spanish as the document language. The Spanish document collection contained 215,738 news stores from EFE News Agency. We used the Spanish-to-English translations provided by the iCLEF organizers for construction of document surrogates and for viewing the full document translations. The translations were created by using Systran Professional 3.0.

We obtained our Spanish-English bilingual term list in our lab. It contains 24,278 words and was constructed from multiple sources [3]. We used the InQuery built-in Spanish stemmer to stem both the collection and the Spanish translations of the English queries. Our KWIC techniques (called "sample usages" in the MIRACLE system for easy understanding by the searchers) require parallel Spanish/English texts and a big monolingual English collection. We obtained the first one from the Foreign Broadcast Information Service (FBIS) TIDES data disk, release 2, and the second one from the TDT-4 collection English news part, which was collected and released by Linguistic Data Consortium (http://www.ldc.upenn.edu).

Measures We computed the following measures in order to gain insight into search behavior and search results:

- $F_{\alpha=0.8}$, as defined in the track guidelines (with "somewhat relevant" documents treated as not relevant). We refer to this condition as "strict" relevance judgments. This value was computed at the end of each search session.
- $F_{\alpha=0.8}$, but with "somewhat relevant" documents treated as relevant. We refer to this condition as "loose" relevance judgments. This value was also computed for each session.
- The total number of query iterations for each search.

The F measure is an outcome measure; it cannot tell us what happened along the way. We therefore also used Camtasia Studio (www.techsmith.com) to record the searcher's activities during each session.

2.3 Results

Searcher characteristics There were four female and four male searchers. Otherwise, the searcher population was relatively homogeneous. Specifically, our searchers were:

Educated. Six of the eight searchers were either enrolled in a Masters program or had already earned at least a masters degree. The remaining two were one undergraduate student that was near graduation and one person with a Bachelors degree.

[1] http://terral.lsi.uned.es/iCLEF/2003/guidelines.html

Mature. The average age over all eight searchers was 33, with the youngest of 21 and the oldest of 45.

Experienced searchers. Five of the eight searchers held degrees in library science. The searchers reported an average of about 7 years of on-line searching experience, with a minimum of 3 years and maximum of 10 years. All searchers reported extensive experience with Web search services, and all reported at least some experience searching computerized library catalogs (ranging from "some" to "a great deal"). Seven of the eight reported that they search at least once or twice a day.

Inexperienced with machine translation. Seven of the eight searchers reported never having, or having only some, experience with any machine translation software or free Web translation services. The remaining one reported having more than "some experience" with machine translation software or services.

Not previous study participants. None of the eight subjects had previously participated in a TREC or or iCLEF study.

Native English speakers. All eight searchers were native speakers of English.

Not skilled in Spanish. Five of the eight searchers reported no reading skills in Spanish at all. Another three reported poor reading skills in Spanish.

Results for relevance judgments Our official results are based on the $F_{\alpha=0.8}$ values averaged over all searchers that ran each condition. We use what we call *strict relevance*, i.e., treating "somewhat relevant" as "irrelevant" in the calculation of precision and recall. We found that the manual and automatic conditions achieved very nearly the same value for this measure when averaged over all eight topics (0.2272 and 0.2014, respectively). For comparison, we recomputed the same results with *loose relevance*, i.e., treating "somewhat relevant" as "relevant" in the calculation. Again, the difference in average values for $F_{\alpha=0.8}$ across all eight topics between the two conditions is small (0.3031 for manual, 0.2850 for automatic). According to Wilcoxon sign-rank test, neither of the differences is statistically significant.

Looking at the results by topic (see Figure 2), only Topic 5 exhibits a clear difference between the $F_{\alpha=0.8}$ values for the manual and automatic conditions, and only with strict relevance judgments. Topic 5 is about the economic policies of the French politician Edouard Balladur. It was neither the topic that the searchers felt was most difficult, nor the topic that they felt was easiest. We do not presently have an explanation for this effect. The reason that there is no difference on Topic 4 is that all the searchers did not find any official relevant document in both conditions, although they marked several. One possible explanation is that there are only 3 official relevant documents in the whole collection of 215,738 documents.

Similar values of $F_{\alpha=0.8}$ can mask offsetting differences in precision and recall, so we also examined precision and recall separately. As Figure 3 illustrates, precision seems to be more sensitive to the manual/automatic distinction than recall, with the precision for Topic 5 strongly favoring the manual condition and

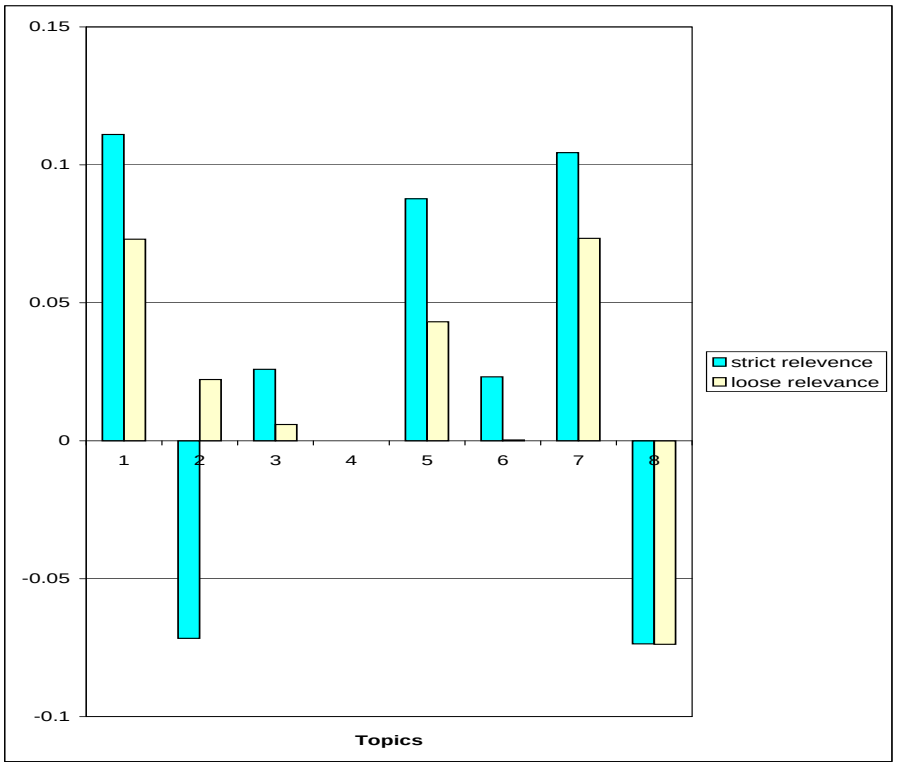

Fig. 2. Absolute improvement in $F_{\alpha=0.8}$ from the manual condition. Bars above the x axis favor the manual condition, below favor the automatic condition. Loose judgments treat "somewhat relevant" as relevant.

the precision for Topic 8 (EU fishing quotas) strongly favoring the automatic condition. On the other hand, recall is somewhat sensitive to the manual/automatic distinction for Topics 1 (The Ames espionage case) and 7 (German Armed Forces out-of-area), in both cases favoring the manual condition. Since we would normally expect recall and precision to change in opposite directions, Topics 5 and 8 (which clearly lack this effect) and, to some degree, Topic 7, deserve further exploration.

One interesting comparison we can draw is between this year's results and those from last year. Although there is language difference (Spanish vs German), topic difference (one of this year's topic only has 3 relevant documents (Topic 4), another has 181 relevant documents (Topic 8)), and some small system differences, it is interesting to note that last year's searchers achieved $F_{\alpha=0.8} = 0.4995$ for the manual condition, and 0.3371 for the automatic condition. The manual condition thus achieved a 48% improvement in $F_{\alpha=0.8}$ over the automatic condition in last year's experiment, but we only saw a 13% improvement this year. Last year, searchers were allowed 20 minutes for each topic; this year they were

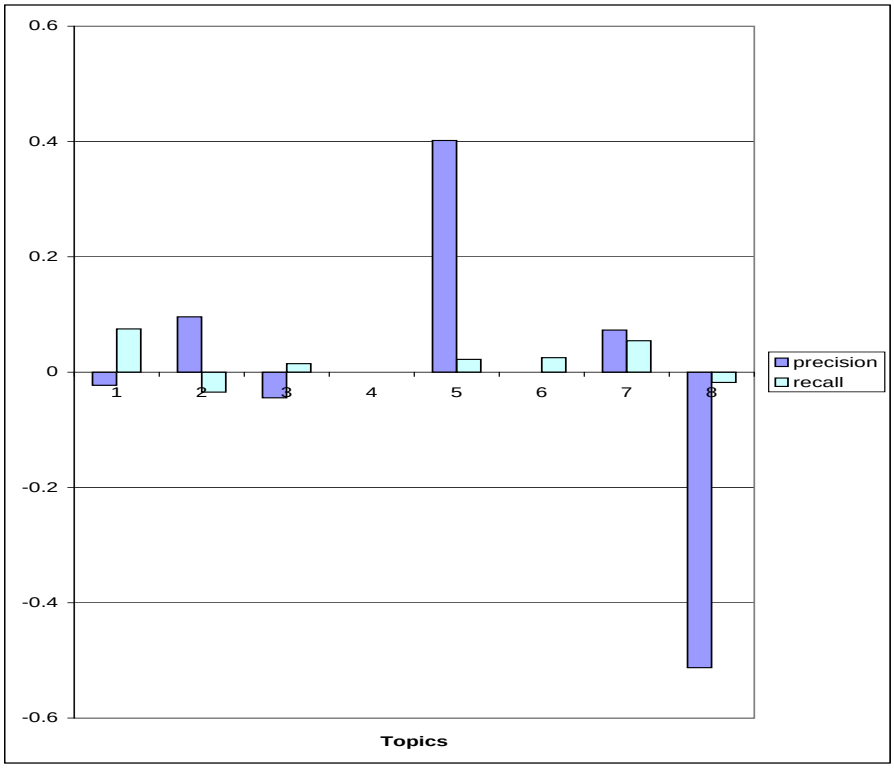

Fig. 3. Absolute improvement in precision and recall from the manual condition, strict judgement. Bars above the x axis favor the manual condition, below favor the automatic condition.

allowed only 10. Perhaps user-assisted query translation is of greater value later in the search process, after some of the more obvious ways of formulating the query have already been tried.

Query iteration analysis We obtained the number of iterations for each search session through log file analysis. On average, searches in the manual condition exhibited fewer query iterations than searches in the automatic condition (3.72 vs 5.22). Looking at individual topics, topic 4 (computer animation) has the largest average number of iterations (8.25) for the automatic condition, followed by topics 1 (5.75) and 7 (5.50), whereas topic 5 has the largest average number of iterations (6) in the manual condition, followed by topics 3 (4.5) and 8 (4.25).

Subjective Assessment We analyzed questionnaire data and interview responses in an effort to understand how participants employed the systems and to better understand their impressions about the systems. The searchers re-

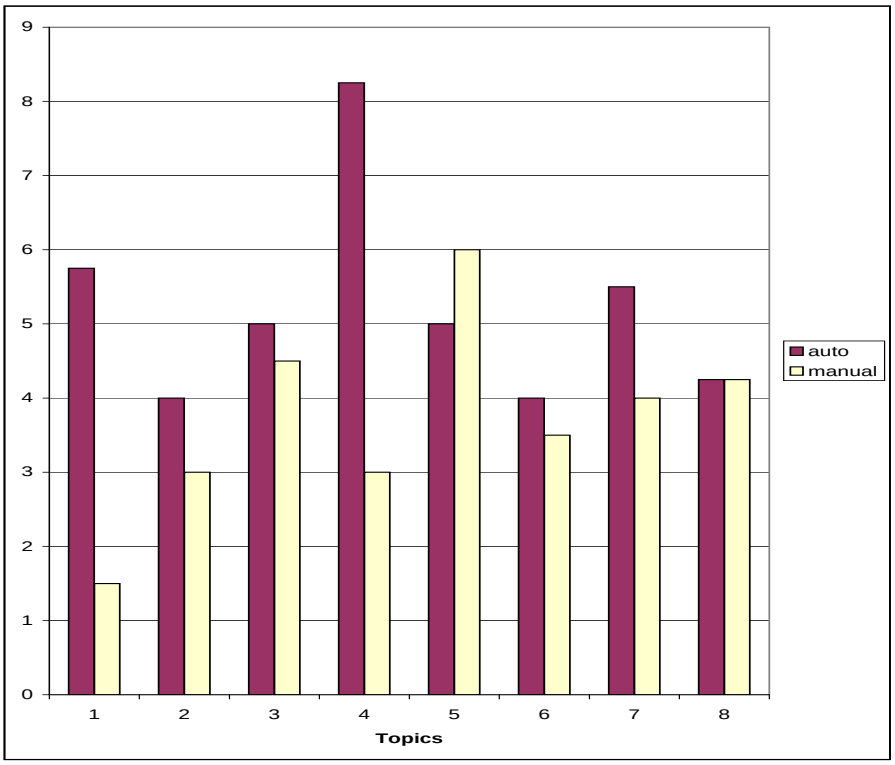

Fig. 4. Average number of query iterations, by topic.

ported that the manual and automatic conditions were equally easy to search with, but because the manual condition was under their control, they had a more satisfying overall experience under this condition.

When talking about the reasons for difficulty finding relevant documents, among the searchers using the manual condition, "unfamiliar topics" and "quality of the translation" were the most often mentioned (4 times each). The same general pattern was observed in the responses of searchers using the automatic condition: three mentions of "translation quality," two mentions of "unfamiliar topics," and two mentions of "no relevant documents returned."

For questions unique to the manual condition, all of the searchers reported that they could (somewhat or very) confidently select/remove translations, and all of them felt that it was (somewhat or very) useful to have the ability to modify the system's translation selections. However, there was no clear preference among the available indications of the meaning of each translation. Three searchers reported that the synonym list was the most useful cue, while two searchers found the examples of usage to be the most useful (three were not sure).

Most participants reported that they were not familiar with the topics, with topic 3 having the greatest reported familiarity and topic 8 having the least.

Search strategies Although we have noticed clear differences between searches performed under the manual condition and than under the automatic condition, we have not noticed any clear difference in constructing the initial queries between searches on the two conditions. The terms of the initial queries were mostly from the provided topic statement, however, some extra terms were used for various reasons. For example, "spy" was used by a searcher in the initial query instead of "espionage", and "intrusion detection" was used by another searcher because the searcher happened to be the domain expert on that topic.

However, there were clear differences in searchers' subsequent behaviors on the two systems. Searchers did not have many extra choices in the automatic condition than in a monolingual search system. Therefore, their tactics were very similar to monolingual tactics, e.g., using terms from relevant documents, adding or removing terms, using synonyms or hyponyms, etc. Interestingly, there were cases in which the searchers picked up some untranslated but obviously important terms from the documents to modify the query. For example, a searcher expanded the query with "king Leon", which is probably a Spanish/English variant of the phrase "Lion King", and found several relevant documents because of it.

Searchers' behaviors varied widely in the context of the manual condition. In all 12 sessions associated with the manual condition, the searchers checked and removed unwanted translations with the help of our three cues. In some cases where the searchers did not like the returned results, they returned to check and/or change the translations, but this happened less frequently than query changes. There were also cases in which searchers returned to change the queries after looking at the translations. In particular, one subject searched only once for topic 4 (computer animation) during the entire 10-minute session, but the searcher issued four different queries, of which the first three were changed based on only looking at translations. Another interesting observation is that translations provide an extra resource for generating query terms. For example, after getting frustrated with the query term "European Union" and its variants, a searcher decided to select one Spanish translation from each word of the term (i.e. "europeo" for european and "sindicato" for union) and put them directly into the query. This subject was able to mark two relevant documents based on this query.

While user-assisted translation selection has proved useful, there are a number of limitations. In particular, translation selection is only one step in the overall CLIR process. It provides an extra interaction anchor that searchers found to be helpful. However, its effect is not as great as query reformulation. We have found many more cases of query reformulation than translation re-selection. In addition, our use of cues to assist searchers limited the usefulness of user-assisted translation selection, to some degree, because our cues do not always provide the *best* explanations. Our goal was to give three different cues which provide *good*

explanations—with the hope that the combination of these three would cover a wide range of situations. In practice, however, the searchers were sometimes confused by contradictory explanations and, as a result, they frequently decided to stick to one cue and ignore the other two.

3 Document Selection Experiments

In addition to the above investigation of the searcher's ability to choose queries using different techniques, we also ran an experiment to determine the user's ability to recognize relevant documents using different techniques. In particular, we compared the searcher's results using two different approaches to presenting document surrogates: (1) a complete extract of the first 40 (translated) words of each news story (2) a compressed extract generated using an automated parse-and-trim approach that approximates one way in which people can produce headlines. This section describes the parse-and-trim approach to headline generation and presents our results.

3.1 Parse-and-Trim Headline Generation

Headline generation is a form of text summarization in which the summarization is required to be informative and extremely short, and to mimic the condensed language features of newspaper headlines. Informative summaries answer the questions "what happened in this document?" or "what is claimed in this document?," rather than the question "what is this document about?"

Our headline generation system, *Hedge Trimmer*, constructs headlines automatically by selecting high-content words from the *lead* sentence of the document.[2] It does this by iteratively removing grammatical constituents from a parse of the lead sentence until a length threshold has been met. The parses are created by the BBN SIFT parser. As described in [6] the BBN SIFT parser builds augmented parse trees according to a process similar to that described in [2]. The BBN SIFT parser has been used successfully for the task of information extraction in the SIFT system [7].

The approach taken by Hedge Trimmer is most similar to that of [5], where a single sentence is shortened using statistical compression. However, Hedge Trimmer uses linguistically motivated heuristics for shortening the sentence. There is no statistical model, so prior training on a large corpus of stories and headlines is not required.

The input to Hedge Trimmer is a story. The first sentence of the story is passed through the BBN SIFT parser. The parse-tree result serves as input to a linguistically motivated module that selects story words to form headlines based on insights gained from observations of human-constructed headlines.

[2] We currently take the first sentence to be the lead sentence of the document; but further investigation is currently underway for selecting the most appropriate lead sentence.

At present, Hedge Trimmer is applied to the problem of cross-language headline generation by translating the first sentence of a story into English and running the Hedge Trimmer process on the resulting translation.

3.2 Use of Hedge Trimmer in iCLEF

We used Hedge Trimmer as the basis of forming document surrogates in the Interactive track for the 2003 Cross-Language Evaluation Forum. In this experiment two methods were used to produce English surrogates for Spanish documents. Surrogate A ("F40") consisted of the first 40 words of a machine translation of the document. Surrogate B ("HT") was a headline constructed by Hedge Trimmer from the machine translation of the first sentence. Eight subjects were shown surrogates for the results of IR searches on eight topics. The translations and search results were provided by iCLEF to all participants.

Each search result consisted of 50 documents. For each topic, the subjects were shown a description of the topic and surrogates for the 50 documents. The subjects were asked to judge whether the document was highly relevant, somewhat relevant or not relevant to the topic and whether they were highly confident, somewhat confident or not confident in their relevance judgment. The order of topics, and whether the subject saw F40 or HT for a particular topic was varied according to the Latin Square provided by iCLEF as part of the standard experiment design.

Our goal was to show that the two surrogates had close recall and precision, but that HT took the subjects less time to perform the task. Subjects were able to complete 1189 judgments in a total of 290:34 minutes with F40, while they completed 1388 judgments in 272:37 minutes with HT. That is, using F40 subjects made 4.09 judgments per minute, while with HT they made 5.09 judgments per minute. However the results of the experiment showed that over 32 searches F40 had an average precision of 0.5939, average recall of 0.3769 and average $F_{\alpha=0.8}$ of 0.4737, while HT had average precision of 0.4883, average recall of 0.2805 and average $F_{\alpha=0.8}$ 0.3798.

Inter-annotator agreement did not differ much between the two systems. We used Cohen's κ [1] to measure the pairwise inter-annotator agreement. κ is 0 when the agreement between annotators is what would be expected by chance, and is 1 when there is perfect agreement. Due to the experiment design, it was not possible to calculate system-specific inter-annotator agreement for each pair of annotators because some pairs of annotators never used the same surrogate for judging the same documents. The average overall κ score was for those cases in which subjects did see the same surrogate for the same document was 0.2455, while the average pairwise κ score for F40 was 0.2601 and the average pairwise κ score for HT was 0.2704.

After the subjects completed judging the documents for a topic, they were asked the following questions:

1. Were you familiar with this topic before the search?
2. Was it easy to guess what the document was about based on the surrogate?

3. Was it easy to make relevance judgments for this topic?
4. Do you have confidence in your judgments for this topic?

The subjects answered each question by selecting a number from 1 to 5, where 1 meant "not at all", 3 meant "somewhat" and 5 meant "extremely." The responses are shown in Table 1.

Table 1. Average Question Responses by System

	F40	HT
Question 1	2.09	1.97
Question 2	3.65	2.91
Question 3	3.75	3.28
Question 4	3.78	3.13

We do not take this result necessarily to mean that informative headlines are worse surrogates than the first forty words. It is likely that the headlines used in HT were not good enough headlines to make a conclusion about informative summaries in general. Also, the average length of the headlines used in HT was much shorter than forty words, giving F40 the advantage of including more topic information.

4 Conclusion and Future Work

We focused on testing the effectiveness of user-assisted translation selection in interactive CLIR application, and observing different search strategies/tactics that the searchers could used in their interaction with a CLIR system with user-assisted translation selection feature. Our analysis suggests the usefulness of the approach, and the diversity of tactics the searchers adapted to take advantage of the extra interaction opportunity provided by user-assisted translation selection. However, the effectiveness of the approach is dependent on the characteristics of the topic, the time pressure, and the quality of the cues. Further development of user-assisted translation selection will be on 1) finihsing analyzing searchers search behaviors, and design better evaluation measures; 2) designing an easy-obtained and robust cue that can provide best explanation all the time.

In addition, we focused on the tasks of determining document relevance. Although the HT system shows some promise for this task, the results indicate that the system has not yet reached a point where better results are consistently obtained. Continued development on headline generation will focus on improving the quality of the outputs that are generated. In particular, the following improvements are planned for headline generation: (1) Use of an n-gram language model for selecting the best translation surrogate produced by the system; (2) Better selection of the window of words from the article from which the headline should be chosen; (3) Use of topic detection to identify words that should

not be deleted. Moreover, one of our next steps is to use HT in the context of summarization of broadcast news. Experiments will compare how well headlines support human performance on an extrinsic task with respecto topic lists, sentence extraction, first-N-words, and other summarization approaches.

Acknowledgments

The authors would like to thank Julio Gonzalo and Fernando López-Ostenero for their tireless efforts to coordinate iCLEF, Nizar Habash for providing us the Spanish-English bilingual termlist, and Michael Nossal for developing the earlier version of the Miracle system. This work has been supported in part by DARPA cooperative agreements N660010028910 and BBNT Contract 9500004957.

References

1. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measures, **20** (1960) 37–46
2. Collins, M.: Three generative lexicalised models for statistical parsing. Proceedings of the 35th ACL (1997)
3. Habash, N. Y.:. Generation-heavy Hybrid Machine Translation. PhD thesis, Department of Computer Science, University of Maryland at College Park (2003)
4. He, D., Wang, J., Oard, D. W., Nossal, M.: Comparing user-assisted and Automatic Query Translation. Proceedings of CLEF'02 (2002)
5. Knight, K., Marcu, D.: Statistics-based summarization step one; sentence compression. Proceedings of AAAI-2001 (2001)
6. Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., Weischedel, R.: Algorithms that Learn to Extract Information; BBN: Description of the SIFT System as Used for MUC-7. Proceedings of the MUC-7 (1998)
7. Miller, S., Ramshaw, L., Fox, H., Weischedel, R.: A novel use of statistical parsing to extract information from text. Proceedings of the 1st Meeting of the North Amererican Chapter of the ACL, Seattle, WA (2000) 226–233
8. Oard, D. W., Levow, G., Cabezas, C.: CLEF Experiments at Maryland: Statistical Stemming and backoff translation. Peters, C. editor, Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal (2000) 176-187
9. Pirkola, A.: The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia (1998)