Summarization Design for Interactive Cross-Language Question Answering

Daqing He[1**], Jianqiang Wang[2], Jun Luo[2], Douglas W. Oard[1,2]

[1] Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA
daqingd@umiacs.umd.edu
[2] College of Information Studies
University of Maryland, College Park, MD 20742 USA
{wangjq,jun,oard}@glue.umd.edu

Abstract. This paper describes an experimental investigation of interactive techniques for cross-language information access. The task was to answer factual questions from a large collection of documents written in a language in which the user has little proficiency. An interactive cross-language retrieval system that included optional user-assisted query translation, display of translated summaries for individual document ranked in order of decreasing degree of match to the user's query, and optional full-text examination of individual documents was provided. Two alternative types of extractive summaries were tried using a systematically varied presentation order, one drawn from a single segment of the translated document and the other drawn from three (usually) shorter segments of the translated document. On average, users were able to correctly answer just 62% of the sixteen assigned questions in an average of 176 seconds per question. Little difference was found between the two summary types for this task in an experiment using eight human subjects. Time on task and the number of query iterations were found to exhibit a positive correlation with question difficulty.

1 Introduction

Question Answering (QA) is a type of information access task. It differs from the more traditional task of finding topically relevant documents in that the information need is modeled as a requirement for a specific factual answer (expressed as a short snippet of text), rather than relevant documents. Cross-Language Question Answering (CLQA) is a special case of the QA task in which the questions and the documents that contain the answers are expressed in different languages. Most QA research has focused on the design and evaluation of fully automatic QA systems. For the Cross-Language Evaluation Forum (CLEF) 2004 interactive

** Current address: School of Information Sciences, University of Pittsburgh, Pittsburgh, PA 15260 USA

track (iCLEF), we explored an interactive variant of CLQA in which the user and the system worked together to rapidly find answers to factual questions.

The usual approach to QA is to first identify a set of candidate documents using information retrieval techniques (e.g., term matching after question rewriting), and then to apply more sophisticated natural language processing (e.g., question type classification, named entity tagging, and logical inference) to identify the location and text span of the most likely answers in those candidate documents. We are not aware of any case in which fully automatic QA technology is yet deployed in an operational setting, but people routinely use information retrieval system of more traditional designs to find answers to factual questions. Therefore, we chose to assess the degree to which a Cross-Language Information Retrieval (CLIR) system could support the interactive CLQA task.

The QA task is a variant of the more traditional passage retrieval task. In the case of QA, however, an exact answer must be found. Our intuition suggested that searchers would be able to correctly recognize the exact answer if shown a longer passage that provided adequate context, so we elected to try two variants on passage retrieval. We rely on the searcher to reformulate the question appropriately for use in a term-based passage retrieval system, thus avoiding the complexity typically associated with the question rewriting component of present QA systems. While this decision places some burden on the searcher, it results in a simpler and less opaque system design, thus (hopefully) leveraging the searcher's ability to iterate towards an appropriate query formulation when their first attempt proves to be unsuitable.

We therefore chose to focus on two research questions:

- What are the effects of different types of summaries on the effectiveness of people finding answers in CLQA tasks?
- What types of search behavior do users of interactive CLQA systems exhibit, and in what ways does that behavior differ from that observed when CLIR systems are used to find entire documents that are relevant to a topic?

Passage selection is a form of extractive summarization. In iCLEF 2003, we explored the utility of alternative summarization techniques as a basis for making relevance judgment in interactive CLIR [1]. We are, however, not aware of any research on the application of summarization techniques for CLQA; our iCLEF 2004 experiments help to fill the gap.

Our interest in search behavior includes query formulation, query reformulation, translation disambiguation, relevance judgment, and search termination. Little is known about these topics for monolingual QA, and CLQA introduces additional complexity. In particular, we are interested in the effect of translation quality on the users ability to accurately recognize correct answers. We know from prior studies that present machine translation systems can often adequately support relevance judgment, even when it would not be adequate to convey a complete understanding of meaning [7].

We begin by describing the interactive CLIR system used in the experiment, including the two types of summaries that we tried, in Section 2. The design of

the experiment is then explained in Section 3, and the analysis of results is in Section 4. The paper concludes in Section 5.

2 The MIRACLE System

We used the Maryland Interactive Retrieval Advanced Cross-Language Engine (MIRACLE) for the interactive CLQA experiments reported in this paper. MIRACLE is the result of an extensive revision of the interactive CLIR system that we used for iCLEF 2003. We made modifications to both the basic architecture of the system and the layout of the user interface (see Figure 1). The system includes an optional user-assisted query translation capability

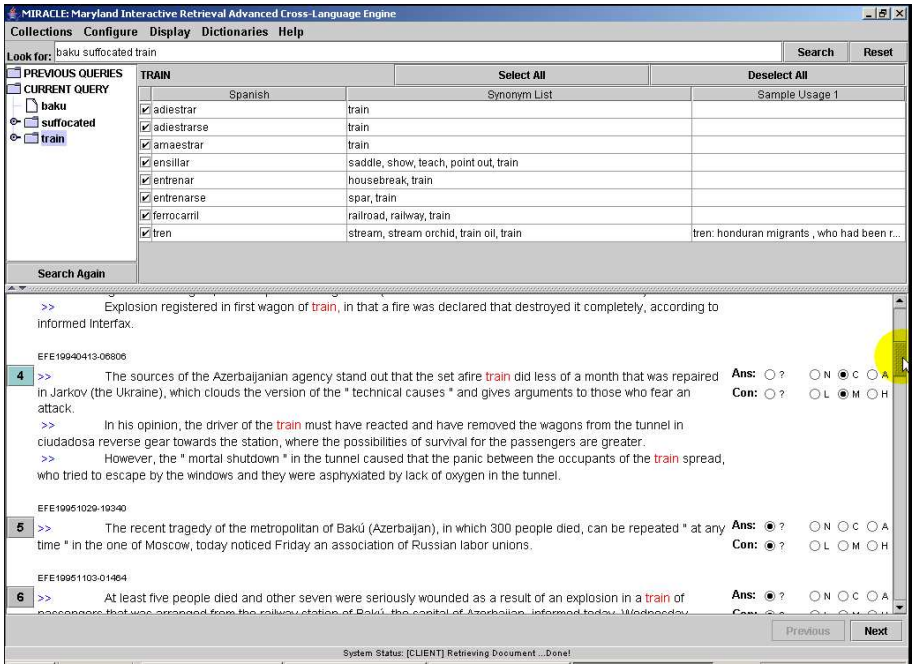

Fig. 1. The MIRACLE user interface for iCLEF 2004, showing KWIC summaries.

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) [6]. All known translations are initially selected, and the user is offered the opportunity to deselect inappropriate translations. Three cues are provided to facilitate this task: (1) The translation itself (which may be recognizable as a loan word), (2) possible synonyms that may help to illustrate the meaning of a translation (obtained through back-translation using the same

term list), and (3) examples of usage (extracted from either parallel text or the combination of the bilingual term list with a large English collection). 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 [5]. A fuller description of the MIRACLE system can be found in [4].

Users recorded answers by hand on the same form as the post-search questionnaire (which also asked for information about prior familiarity with the topic of the question and for an assessment of the subjective difficulty of the question). We modified the logging functions in MIRACLE to accommodate the requirement to designate supporting documents. The user could designate a document as supporting the answer based on either the summaries or on the full document by clicking the numbered button at the left side of the summary. Choices included "N" (not containing an answer), *"C"* (cannot tell), or *"A"* (containing an answer). He/She also could also optionally mark their confidence in that judgment as *"L"* (low confidence), *"M"* (medium confidence), or *H"* (high confidence). Users were not instructed to designate only one single supporting document when possible, and they were not told about the option of designating exactly two documents, each of which provided only partial support. When users designated more than one supporting document, we therefore chose one arbitrarily to submit for official scoring. Because this may not be the same document that the user would have chosen had we instructed them properly, our results may show a somewhat higher rate of unsupported (but otherwise correct) answers than would have been the case with proper user instruction.

2.1 Two Types of Summaries

To help users to identify potentially relevant documents in a ranked list, MIRACLE normally provides a Keyword-In-Context (KWIC) summary of the document. Each KWIC summary consists of up to three sentences that each contain at least one query term. In order to reflect the topical coverage of the document as accurately as possible in a limited space, we sample these three sentences from the beginning, the middle, and the end of the document respectively. This type of summary aims to provide a concise overview of the topical content of the document in order to support the task of relevance judgment. KWIC can therefore be viewed as a type of indicative summary. Figure 2 shows an example of a KWIC summary.

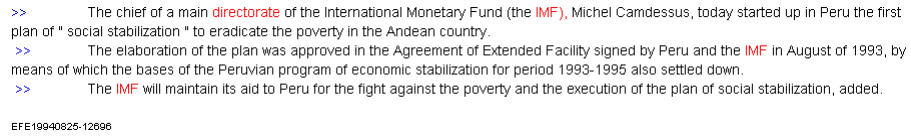

Fig. 2. KWIC summary.

To support our iCLEF 2004 experiment, we added longer single-passage summaries to MIRACLE in an effort to provide the user with more context than the single-sentence KWIC summaries can provide. Our goal in this case was to help the user find answers directly using the single-passage summary; these summaries were therefore intended to be informative rather than indicative. We adapted a passage retrieval module that we had developed for the High Accuracy Retrieval from Documents (HARD) track of the 2003 Text Retrieval Conference (TREC) [3]. The module first uses the density of unique query terms to identify the possible locations of relevant passages, then extends those passages to the nearby paragraph boundaries. When no clear annotation of paragraph boundaries can be found, the module extends the passage to a preset window size, and then further extends the passage to the next sentence boundary in each direction. If two passages are found that are adjacent or overlapping, they are then merged. Passages constructed in this way typically contain several sentences. When a document contains several passages, they are ranked based on a linear combination of the density of unique query terms in the passage and the score assigned by InQuery to the document hat contains the passage. In the passage retrieval condition, we rank passages rather than documents; multiple passages from the same document can appear in the ranked list. Figure 3 shows a one-sentence passage summary (many passage summaries are longer than this).

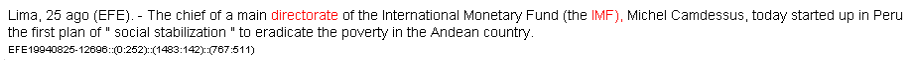

Fig. 3. Passage summary.

Results from the TREC 2003 HARD track indicated that our passage retrieval module typically identified the locations of relevant passages about as accurately as we were able to identify relevant documents, but that the passages we generated were typically far shorter (averaging 207 characters) than the ground truth passages specified by the HARD assessors (which averaged 5,945 characters). This probably is not a problem in the iCLEF 2004 setting, since we would expect that identifying short answers to factual questions would not require very long passages

We provided two variants of MIRACLE system to help the user to perform CLQA task. With all other components of the MIRACLE system remaining the same, one variant used the KWIC summaries as the surrogates of returned documents, which we call *KWIC condition*, and the other variant used the passage summaries, which we refer as *Passage condition*.

We can think of possible advantages for each condition in an interactive CLQA task. For example, KWIC summaries might help the user quickly identify documents that could contain the answer, and their inherent diversity may make them more robust in the presence of machine translation errors. Passage summaries, by contrast, may be more coherent and they might more often tell

the user the answer directly. We are not aware of any systematic study on this question for interactive CLQA; our work in iCLEF 2004 was intended to fill that gap. In particular, we were interested in the following questions:

1. Is there a measurable difference in task performance between using informative and indicative summaries for a CLQA task?
2. Is there a subjective preference for informative summaries over indicative summaries, or vice versa?
3. Is there a difference in users' search behavior (e.g., the frequency of consulting the full document) when the users are given informative summaries rather than indicative summaries?

3 Experiment Design

We followed the standard protocol for iCLEF 2004 experiments. Searchers were sequentially given 16 questions (stated in English), eight using the KWIC condition, and the other eight using the Passage condition. Eight searchers (umd01-umd08) performed the experiment using the eight-subject design specified in the track guidelines.[3] Presentation order for questions and systems was varied systematically across searchers using the required Latin Square design. After an initial training session, each searcher was given a maximum of 5 minutes for each search to find the answer, print it on a piece of that we provided, and (using the radio buttons) identify which documents supported that answer. The searchers were asked make sure that they actually found the correct answers.

We asked each searcher to fill out brief questionnaires before the first question (for demographic data), after each question, and after using each system. Each searcher completed the experiment at a different time, so we were able to observe each individually and make extensive observational notes. In addition, we used Camtasia Studio (www.techsmith.com) to record each searcher's screen activities and we asked searchers to think aloud. We also conducted a semi-structured interview (in which we tailored our questions based on our observations) after all questions were completed.

3.1 Resources

We chose English as the query language and Spanish as the document language. The Spanish document collection contained 454,045 news stores from EFE News Agency that were written in 1994 and 1995. We used the standard Spanish-to-English translations provided by the iCLEF organizers (which had been run using Systran Professional 3.0 at the University of Maryland) as a basis for construction of document summaries and for display of the full document translations.

We obtained a Spanish-English bilingual term list containing 24,278 terms that was constructed from multiple sources from earlier experiments that were

[3] http://nlp.uned.es/iCLEF/

run in our lab [2]. We used InQuery's built-in Spanish stemmer to stem both the collection and the Spanish translations of the English queries. The examples of usage shown in MIRACLE to support user-assisted query translation require a parallel Spanish/English text collection and a large monolingual English collection. We obtained the parallel text from the Foreign Broadcast Information Service (FBIS) TIDES data disk (release 2) and the large collection of English text from the English part of the TDT-4 collection (which is available from the Linguistic Data Consortium, http://www.ldc.upenn.edu).

3.2 Measures

We computed two types of measures to gain insight into search behavior and search results:

- Objective measures of the performance, such as the accuracy of identified answers;
- Objective measures of the search effort, such as the average time in seconds to find answers, the total number of query iterations for each search, and the fraction of answers found using a summary alone without examining the full document.

4 Results

Our analysis is not yet complete (notably, we have not yet looked at the data we collected on examination of full documents), but in this section we present the results that were available at the time this paper was due.

4.1 Searchers

We had relatively homogeneous searchers, who were:

Educated. All eight searchers either had already earned a Bachelors degree or were undergraduate students.

Mature. The average age was 26, with the youngest being 19 and the oldest 35.

Experienced searchers. The searchers reported an average of about 10 years of on-line searching experience, with a minimum of 7 years and maximum of 15 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"). All eight reported that they search at least once or twice a day.

Inexperienced with machine translation. All eight searchers reported never having, or having only some, experience with any machine translation software or Web translation services.

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

Native English speakers. All eight searchers were native speakers of English.
Not skilled in Spanish. Seven of the eight searchers reported no reading skills in Spanish at all. The remaining one reported poor reading skills in Spanish.

4.2 Quantitative Analysis

Searchers achieved over 60% accuracy in both conditions (0.625 for the Passage and 0.609 for the KWIC). The difference was not statistically significant (at p¡0.05) using a Wilcoxon signed-rank test (see Figure 4).

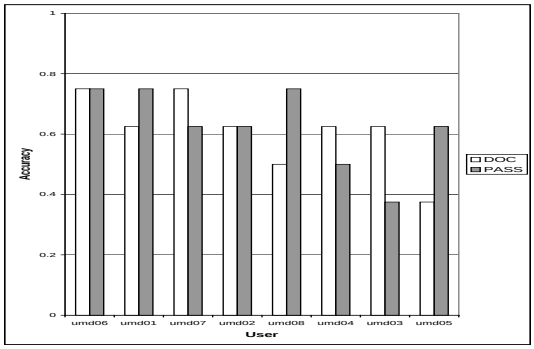

Fig. 4. Accuracy of answers identified under passage condition (i.e. PASS) and KWIC condition (i.e., DOC).

We observed that the questions roughly fall into three categories of difficulty according to the proportion of the correct answers to all answers: easy (Questions 8, 11, 13, 4, 16, 6), moderate (Questions 14, 7, 2, 10, 15, 12), and difficult (Questions 1, 3, 9, 5) (see Figure 5). Table 1 shows the questions themselves in order of increasing difficulty.

One possible factor contributing to question difficulty is the type of information that a question asks for. As Table 1 shows, questions asking for names (person's name, team's name) are generally easier than questions asking for quantities (e.g., number of people, amount of money). When seeking a person's name, queries consisting of terms describing the person's role (e.g., president, director, or winner) and terms naming a related organization (e.g., International Monetary Fund, Burundi, or Nobel Prize) were generally effective; such terms are typically highly selective. On the other hand, for questions about figures, good query terms may be harder to find (as was the case for "when did Lenin die?"), or it may be difficult to determine which of several possible answers is correct (particularly for events that evolve with time such as "How many people were declared missing in the Philippines after the typhoon 'Angela'?").

Searchers spent less time finding the answer to a question under the passage condition than under the KWIC condition (167 seconds vs. 185 seconds). Six of

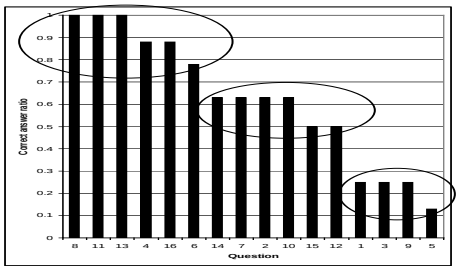

Fig. 5. Questions ranked in order of increasing difficulty (decreasing accuracy), clustered in three groups.

the eight searchers spent less time answering questions under the passage condition (see Figure 7), and Figure 6 shows a clear relationship between increasing question difficulty and increasing search time.

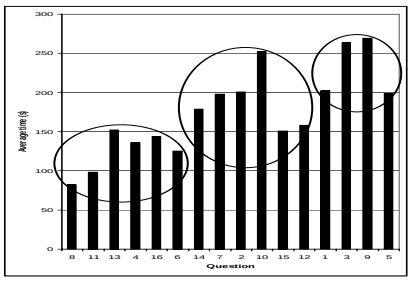

Fig. 6. Average search time, in order of question difficulty (grouped as in Figure 5

When people encounter a question whose answer is difficult to find, one of the strategies they often apply is to modify their query, a process that we call iterative query refinement. The average number of query iterations per question, shown in Figure 8, can be used as an alternative to search time as an indicator of effort. A general trend towards an increasing number of iterations with increasing question difficulty is evident, although there are several clear counterexamples. Five of the eight searchers performed fewer query iterations in the Passage condition (see Figure 9).

Question 8	Who is the managing director of the International Monetary Fund?
Question 11	Who is the president of Burundi?
Question 13	Of what team is Bobby Robson coach?
Question 4	Who committed the terrorist attack in the Tokyo underground?
Question 16	Who won the Nobel Prize for Literature in 1994?
Question 6	When did Latvia gain independence?
Question 14	When did the attack at the Saint-Michel underground station in Paris occur?
Question 7	How many people were declared missing in the Philippines after the typhoon "Angela"?
Question 2	How many human genes are there?
Question 10	How many people died of asphyxia in the Baku underground?
Question 15	How many people live in Bombay?
Question 12	What is Charles Millon's political party?
Question 1	What year was Thomas Mann awarded the Nobel Prize?
Question 3	Who is the German Minister for Economic Affairs?
Question 9	When did Lenin die?
Question 5	How much did the Channel Tunnel cost?

Table 1. The 16 questions, sorted in order of increasing difficulty (decreasing accuracy).

Figures 4, 7 and 9 reveal substantial differences among the eight users participating in the experiment. Accuracy varied between 0.5 and 0.75 for both systems, and average search time spanned an even larger range. For example, on average umd03 spent 108 seconds to find the answer to a question, while umd01 doubled that time for the same task. No correlation between time and accuracy was evident; spending more time doesn't necessarily lead to high accuracy. For example, umd07 achieved 32% better accuracy over umd05 while spending 44% less time. The average number of query iterations exhibited even larger variation, raging from 1.5 iterations per question to 5.2 iterations per question. Again, however, no correlation between the number of iterations and accuracy, nor between the number of iterations and the average time per question is evident. In other words, performing more query iterations does not necessarily lead to higher accuracy, nor does it necessarily take more time.

4.3 Search Behavior

We observed some clear differences between the search behavior exhibited in this CLQA task and the search behavior that we have previously observed when using a CLIR system to search for relevant documents. The most striking difference was that the searches were all precision oriented in the CLQA case. Searchers usually stopped their search after they became convinced that they had found the answer. This usually involved one document providing the answer, and then one or two additional documents providing confirming evidence. Searchers found confirmation in the text surrounding the answer string, either in the summary

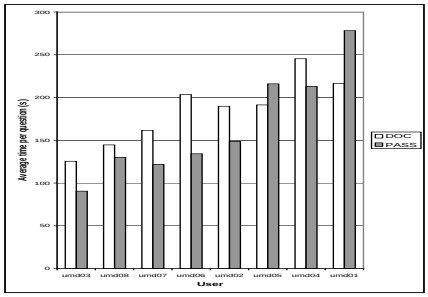

Fig. 7. Average search time per question under passage condition (PASS) and KWIC condition (DOC).

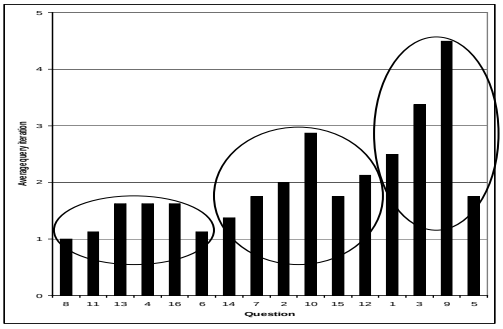

Fig. 8. Average query iterations, in order of question difficulty grouped as in Figure 5

or in the full document, or in the text of other documents. In some cases, these other documents were found in the same ranked list; in others the searchers reformulated the query to generate a more focused ranked list of documents. One tactic that was observed repeatedly was to include the answer as part of query. For example, one reformulated query for the initial query "charles millon political party" was "charles millon udf" ("udf" was the party abbreviation), which was the answer. This is very similar to the strategy used in the answer verification stage of many automatic QA systems; this coincidence suggests that observing search strategies in interactive CLQA may offer insights that could be useful in the design of fully automated systems.

One commonly search tactic observed in our previous experiments on finding topically relevant documents was that the searchers first identify the key concepts of a search topic and then formulate the query as a set of keywords that are synonyms or morphological variants expressing those key concepts, with the

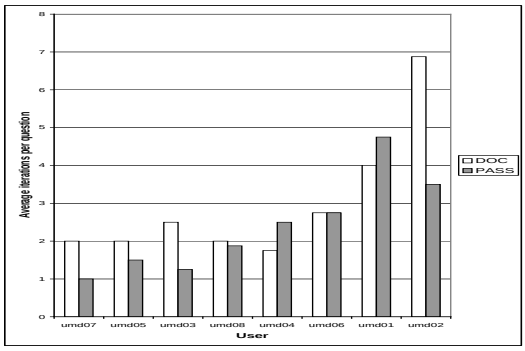

Fig. 9. Average query iterations per question under passage condition (PASS) and KWIC condition (DOC)

hope of bringing back as many relevant documents as possible. This is akin to the "*building blocks*" technique that professional searchers are taught for use with systems that support Boolean query logic. Obviously, this tactic is recall oriented. However, although our searchers performing a CLQA task mentioned that they did pay attention to key concepts, they almost always used the exact words from the question in their initial queries. For example the common initial queries for the question "how many pandas are there in the wild in China" were "pandas china" and "pandas wild China." Only when no good search term was present in a question for a key concept did searchers introduce a new search term in their initial query. For example, the word "population" was included in the initial query "Bombay population" for the question "how many people live in Bombay." Synonyms or morphological variants were used in subsequent queries only when the initial query failed to return the answer. This is more similar to the professional searcher's "pearl growing" technique.

4.4 Other Factors

We did observe numerous interactions between the accuracy of the answers, the time used for search, the number of relevant documents in the collection, and the way that the answer are stated in those documents. For example, "who is the director of international monetary fund" was a question for which none of our searchers had previous knowledge, but it turned out to be an easy question (all 8 searchers marked this question as easy) because many returned documents directly stated the exact string "the director of international monetary fund" with the answer "Michel Camdessus." All eight searchers found the correct answer, in an average of 83 seconds, about half the 176 seconds average time for all 16 topics. However, although five searchers stated that the question "when did Lenin die?" sounded familiar, six searchers marked it as a difficult question, and only two searchers found the correct answer. This was probably because the only

relevant document they could find was one that indirectly implied the answer with "after 70 anniversary of Lenin's death" in an article from 1994. The average time that the searchers spent on that question was 267 seconds; six of them used all 5 minutes without finding the answer. Interestingly, these two questions suggest that a searcher's *pre-search familiarity* with a question does not always play an important role in finding the answer rapidly, or even correctly.

The suboptimal quality of machine translation was another factor that we observed could affect the accuracy of answers, but only for the more difficult or vague questions. For example, the searchers did not have any problem finding the answer for "who is the director of international monetary fund," but they did have trouble finding a correct answer for "who is the German Minister of Economic Affairs" because many machine-translated documents contained phrases such as "German Minister of Outer Subjects" and "German Minister of Economic Cooperation." Because the searchers knew that machine translation may not be perfect, they could mistakenly assume that the person associated with "German Minister of Outer Subjects" or "German Minister of Economic Cooperation" (and especially the latter) was the correct answer. As a result, only two searchers correctly found the answer to that question, whereas three other searchers gave an incorrect answer. That question also had the second longest average search time (264 seconds). Another example of the quality of machine translation affecting the searchers' judgments was that there were many returned documents mentioning "bogging bear." It took a while for the searchers to become convinced that "bogging bear" was a bad translation of "panda."

Summarization quality was also observed to affect the results, but only for the more difficult questions. Because of time pressure, the searchers made extensive use of summaries to find documents that potentially contained an answer. When the answer strings were present in the summaries, they could find them with ease, but they would miss the relevant documents if the answer strings were not in the summary. For example, although the question about Lenin's death was a difficult, two searchers just happened to use the a query that resulted in inclusion of the answer string "70 years anniversary of Lenin's death" in the displayed summary. Therefore, those two searchers found the answer fairly easily.

The clarity in expressing a question could also affect the results. Two questions asked about times; one was "When did the attack at the Saint-Michel underground station in Paris occur," and the other was "When did Latvia gain independence." The answer to the first one was "July 25, 1995," while the answer to the second one was less precise: "September 1991." Some searchers wondered whether the exact date of Latvia's independence was required. A more problematic question was "How many people were declared missing in the Philippines after the typhoon 'Angela'?" Of course, the immediate aftermath of a disaster (which can be expected to dominate the reporting) is typically somewhat chaotic, so data appearing in the media might initially be inaccurate. This naturally led to different interpretations by different searchers. Problems of that sort could be minimized by including clearer criteria in the question (e.g., by specifying a time frame "after four days," or a source "in the final government statistics.")

4.5 Subjective Evaluation

Overall, all the searchers thought that finding answers under both conditions was easy and that both types of summaries were effective in supporting their tasks. The searchers liked the display of additional text around the answers because it allowed them to judge the correctness of the answer. Five of the eight searchers preferred the passage summaries because the summaries typically offered more context information than the KWIC ones. The other three searchers preferred the KWIC summaries because the summaries allowed them to gets a sense of the content of the full documents and because took less time to read. They also felt that the passage summaries did not always give the information they needed, and sometimes the passage summaries were too long. Highlighting query terms in both summaries and full documents was appreciated because it helped the searchers to zoom in to the right text, a very useful feature in longer texts.

5 Conclusion and Future Work

In this experiment, we compared the effectiveness of an interactive cross-language information retrieval system enhanced with two alternative types of document summaries for supporting the task of finding answers in Spanish documents to questions expressed in English. We found that our MIRACLE system was moderately effective, with correct answers found in 62% of the 128 searches that were performed. Users achieved comparable accuracy with either type of summary, but they achieved that accuracy somewhat more rapidly (167 vs. 185 seconds), on average, when using single-passage summaries. Our experiment results revealed substantial differences among the eight users participating in our study, both in terms of the number of questions they answered correctly (accuracy) and the average time they spent answering a question. We also investigated question difficulty, finding that both the amount of time needed to answer a question and the number of queries that were posed increased as the questions became harder (i.e., as accuracy decreased).

Question answering is an attractive task for evaluation of interactive cross-language information retrieval systems because it is grounded in something that real users really do. Our initial results from these first experiments with interactive question answering are indeed promising, but there are many interesting questions that remain to be explored. The first and most obvious is how our systems might be tailored to better support this task. In our iCLEF 2004 experiments, we tried alternative types of summaries, but we used the same summary for every question type. Can we tailor the summary to the question type, either automatically or under the user's control? Are there other system functions (e.g., term highlighting) that might also be adapted based on the question type? Thinking more broadly, are there other important question types that would yield new insights? What functions might we provide to support inference across documents? Can we design experiments to model the more realistic case in which the user has partial knowledge of the answer that they seek?

Over the past six years, CLEF has become increasingly grounded in real tasks. In its first two years, CLEF focused on building ranked lists. The 2001 iCLEF evaluation introduced a focus on interactive selection of documents from those ranked lists. In 2002 and 2003, we expanded this focus to include iterative refinement of the queries from which those ranked lists were produced. And now, in 2004, we focus on a complete task that end users sometimes actually perform, seeking answers to factual questions. As we move closer to real tasks, we have learned more about the kind of system support that are needed. CLEF plays a unique and important role in the CLIR community by uniting this focus on the task with the challenge of building systems to support that task. We look forward to continuing this exploration, and to working with the CLEF community to identify the next directions for this important effort.

Acknowledgments

The authors would like to thank Julio Gonzalo for coordinating iCLEF and Nizar Habash for providing the Spanish-English bilingual term list. This work has been supported in part by DARPA cooperative agreements N660010028910.

References

1. Bonnie J. Dorr, Daqing He, Jun Luo, Douglas W. Oard, Richard Schwartz, Jianqiang Wang, and David Zajic. iCLEF 2003 at Maryland: Translation Selection and Document Selection. In *Proceeding of CLEF 2003*, 2003.
2. Nizar Y. Habash. *Generation-heavy Hybrid Machine Translation.* PhD thesis, Department of Computer Science, University of Maryland at College Park, 2003.
3. Daqing He and Dina Demner-Fushman. HARD Experiment at Maryland: from Need Negotiation to Automated HARD Process. In *Proceeding of TREC 2003*, 2003.
4. Daqing He, Douglas W. Oard, Jianqiang Wang, Jun Luo, Dina Demner-Fushman, Kareem Darwish, Philip Resnik, Sanjeev Khudanpur, Michael Nossal, Michael Subotin, and Anton Leuski. Making MIRACLEs: Interactive Translingual Search for Cebuano and Hindi. *ACM Transaction of Asian Language and Information Processing*, 2003.
5. Douglas W. Oard, Gina-Anne Levow, and Clara I. Cabezas. CLEF Experiments at Maryland: Statistical Stemming and backoff translation. In C. Peters, editor, *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum, CLEF 2000*, pages 176–187, Lisbon, Portugal, 2000.
6. Ari Pirkola. The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998. ACM.
7. Jianqiang Wang and Douglas W. Oard. iCLEF 2001 at Maryland: Comparing Word-for-Word Gloss and MT. In C. Peters, M. Braschler, J. Gonzalo, and Kluck M, editors, *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001*, pages 336–354, Darmstadt, Germany, 2001.