Interactive Cross-Language Information Retrieval

Douglas W. Oard

Over the past year, there has been increasing interest in interactive applications of Cross-Language Information Retrieval (CLIR). There have now been two workshops devoted to that topic, and the European Cross-Language Evaluation Forum (CLEF) has started an experimental track on evaluation of interactive CLIR. Some might ask "why all the fuss?" In part the answer is simple - CLIR has been uncharitably called "the problem of finding documents that you can't read." But it's actually worse than that. It may actually be "the problem of finding documents that you can't even recognize!" The purpose of this brief contribution is to recap the genesis of the present interest in this topic, to point out some of the key questions, and to suggest some ways to get involved.

Two years ago, at the Association for Computational Linguistics (ACL) conference, an unexpected question planted a seed. The setting was a theme session on Natural Language Processing for CLIR, and the question was "Is the cross-language information retrieval problem now solved?" Reports of cross-language mean average precision exceeding 75% of monolingual results were becoming commonplace, well designed dictionary-based techniques were reported at SIGIR 98, and sophisticated corpus-based techniques were now being reported at ACL 99. So perhaps the answer should have been "yes." But if the problem was solved, why weren't the solutions appearing in the marketplace? A curious situation had emerged—Web search engines were adding machine translation capabilities to translate pages that they found, but none of the major search engines had implemented a cross-language search capability. Amazingly, that situation is unchanged two years later! Perhaps the marketplace also knew something about CLIR that we were not yet able to see.

Two months later, I found myself at a SIGIR 99 workshop that Carol Peters and I had organized on "Multilingual Information Discovery and Access." Our principal goal was to think more broadly about the contexts within which technologies such as CLIR would be employed. A second seed was planted when the discussion there turned to the role of the user in interactive cross-language searching. There was considerable interest in that topic among the participants, a remarkably diverse group with backgrounds in CLIR, multilingual metadata, machine translation, and cross-language document summarization. Bill Ogden, Judith Klavans, Chin-Yew Lin and I decided that further discussion was warranted, so we set out to organize a follow-on workshop on what we called "Interactive Searching in Foreign-Language Collections."

The Human-Computer Interaction Laboratory (HCIL) at the University of Maryland hosts a set of one-day workshops each summer that have proven to be a convenient venue for the topic that we had in mind, so we scheduled the workshop for June 1, 2000 at Maryland. We organized the day in four parts: demonstrations, user needs, research results, and future

directions. Our discussions were broad-ranging, but they led to one inescapable conclusion—we knew much more about how to build interactive CLIR systems than we knew about how to evaluate them. Five working systems were demonstrated that day, but we had little more to go on than our own intuition when evaluating about which of the ideas were most promising. Simone Teufel gave a well-received presentation on evaluation techniques, and Bill Ogden presented some interesting user study results, but it was clear that many of us had a lot to learn before we would even be in a position to ask the right questions.

With this background, the limitations of measures such as mean average precision began to come into sharper focus. Mean uninterpolated average precision involves expectations over two random variables: the query that the user will pose, and the number of relevant documents that the user will choose to examine. We take it as an article of faith that a system with better mean average precision can better serve interactive users. A closer look reveals that this faith is based on an implicit assumption that the user can understand the documents well enough to recognize the relevant ones—if for no other reason than they will need to decide when to stop examining documents! If they are not able to read the language in which the documents are written, an interactive searcher would need some help from the system. Machine translation researchers now see this as one of the central tasks on which their system should be evaluated, but no agreed evaluation metrics have yet emerged.

Document selection is only one part of the story, however. Interactive query translation, included in two of the systems demonstrated at the HCIL workshop, poses even more complex evaluation challenges. And the situation is further complicated by the process of interactive query refinement that we see in monolingual interactive systems and would expect to see in interactive CLIR systems as well. Over the course of the last year, we and others have explored some of these questions, and this year there are two events focused on evaluation of interactive CLIR systems: an interactive track at the European Cross-Language Evaluation Forum (CLEF), and a second HCIL workshop.

The CLEF interactive track builds on the methodology developed over the course of several evaluations of interactive monolingual interactive retrieval at the Text Retrieval Conference (TREC). The principal goals of this first interactive track at CLEF are to foster the development of a research community around this challenge, and to work on the development of a practical experiment design that can offer insight into questions of importance to interactive CLIR. We have chosen to focus on the document selection problem, using an unbalanced F measure as a measure of the system effect and a Latin square design to block topic and searcher effects.

For this year's HCIL workshop on interactive CLIR, we chose to focus on assembling a broad view of the evaluation issues that should be considered. Clare Voss of the Army Research Laboratory began the day with a demonstration of their Forward Area Language Converter (FALCON), a portable system designed to produce translations of scanned documents. Although FALCON presently includes only a limited grep-like search capability, the presentation provided a great deal of insight into user needs and document display issues that also arise in interactive CLIR applications. Jianqiang Wang then described ARCTOS, an interactive CLIR system developed at New Mexico State University, and presented the results of a pilot study that suggested that transition analysis could provide a useful source of insight into aspects of the query reformulation process that are unique to interactive CLIR.

Examples of two other evaluation frameworks were included. Eiman Hamday presented the results of a qualitative study with a design that was similar in broad outline to that of the CLEF interactive track. One interesting finding of that study was a clear trend indicating that presenting a greater number of translation alternatives resulted in fewer relevance assessments in a fixed time period, although the extent of individual differences precluded statistical significance. Fred Gey of the University of California at Berkeley shifted the focus from document selection to query formulation, suggesting that query terms that are automatically recommended to the user can be evaluated by leveraging the existence of test collections that have been manually indexed.

After I described the design of the CLEF interactive track evaluation, Nizar Habash presented an overview of machine translation evaluation. The participants found particular interest in the distinction between accuracy and fluency measures, and considerable discussion of the relative importance of those factors for interactive CLIR ensued. Gina Levow completed the formal presentations with a description of available resources that could reduce barriers to entry for groups interested in evaluation of interactive CLIR systems.

The most valuable part of any workshop is the interaction between the participants, so we devoted the last session to a reflective discussion of our reactions to the panel. Irene Tseng of Galudet University, Paul McNamee from the Johns Hopkins University Applied Physics Laboratory, and Bonnie Dorr initiated the discussion by offering their assessment of the issues that had been raised. One important theme that emerged was the nearly universal recognition among those that had run user studies that these types of experiments require a far greater effort than was initially estimated. One person suggested that a useful rule of thumb was to plan on a net investment of $1,000 per participant. Two participants in the workshop, Dagobert Soergel and Jianqiang Wang, contributed their notes from the discussions throughout the day for inclusion on the workshop's Web page. Those notes, together with the slides from many of the presentations, can be found at the URL provided below.

Perhaps the most significant result of the extended collaboration described in this brief contribution is the realization that evaluation of interactive CLIR is an important and interesting problem to which participants from many disciplines can productively contribute their perspective. Quantitative and qualitative studies both offer valuable insights, and researchers with backgrounds in information retrieval, natural language processing, and human-computer interaction have all made valuable contributions. Important work clearly still remains to be done, however, before we will be able to state that the interactive CLIR problem has been solved.

Links to Web Resources

- MIDAS Workshop: http://www.clis.umd.edu/conferences/midas/
- HCIL 2000 Workshop: http://www.clis.umd.edu/conferences/hcil00/
- HCIL 2001 Workshop: http://www.clis.umd.edu/conferences/hcil01/
- CLEF Interactive Track: http://ibuzuki.lsi.uned.es/iCLEF/