User-Assisted Query Translation for Interactive CLIR

Daqing He,[1] Jianqiang Wang, [1,2] Douglas W. Oard,[1,2] Michael Nossal[1]

[1] Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742

{daqing,nossal}@umiacs.umd.edu

[2] College of Information Studies
University of Maryland, College Park, MD 20742

{wangjq, oard}@glue.umd.edu

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Graphical user interfaces (GUI)*

General Terms

Performance, Experimentation, Human Factors, Languages.

Keywords

CLIR, User-assisted query translation, KWIC, MIRACLE.

We view interactive Cross-Language Information Retrieval (CLIR) as an iterative process in which the searcher and the retrieval system collaborate to find documents that satisfy the searcher's needs, regardless of the language in which those documents are written. Our motivation is that humans and machines can bring complementary strengths to this process. Machines are excellent at repetitive tasks that are well specified; humans bring creativity and exceptional pattern recognition capabilities. Properly coupling these capabilities can result in a synergy that greatly exceeds the ability of either human or machine alone. Designers of CLIR systems can select from a variety of fully automatic techniques to overcome problems with unknown terms and translation ambiguity, but automatic processing of this sort risks reducing the searcher's understanding of system operation. This, in turn, tends to work against the synergy that we seek to accomplish. We are therefore exploring more transparent approaches to support interactive CLIR.

The *user-assisted query translation* approach that we will demonstrate offers the searcher insight into translation alternatives made automatically by the system, allowing interactive refinement of those alternatives. Behind the scenes, we use Pirkola's structured query method [2] with the InQuery text retrieval system, for which batch retrieval experiments have shown that selecting all known translations of a query term is very effective (with measures that average over many topics). The system therefore starts with all possible translations selected, allowing the searcher to interactively deselect inappropriate translations. The user can, of course, perform an initial search immediately, only returning to perform interactive translation selection in the event that the initial results are unsatisfactory.

For searchers with some knowledge of the document language, we could think of this process as the system helping the searcher translate their query. But it is more natural to think about the searcher helping the system when the searcher does not have document-language skills. To support that process, we must find some way to explain the meaning of each translation alternative. Optimally, we would want to present dictionary definitions in the query language of each possible document-language translation. Unfortunately, relatively few bilingual dictionaries are structured in that way, and even fewer such dictionaries are easily available in electronic form for incorporating into a CLIR system. We have therefore developed two automatic techniques that together can often offer a monolingual searcher with a useful degree of insight into the meaning of a translation: 1) *back translation*, and 2) *Keyword In Context (KWIC)*. Back translations are query-language terms that share a common translation. For example, the English word "fly": might be translated as "bragueta," whose back translations can be "zipper." "Fly" also might be translated as "huir," which back translates as "flee." Back translations are easily found using a simple bilingual term list, and as the above example shows they can sometimes be informative. In other cases, it can be more helpful to have an example of usage. Our KWIC technique uses sentence-aligned parallel text to identify a brief query-language passage for each translation by finding a sentence-translation pair in which the query term appears on one side and the desired document language term appears on the other. The words around the query-language term can then be shown as an example of usage that is illustrative of this translation. More details on both techniques can be found in [1].

MIRACLE, the system that we will demonstrate was designed to support user studies for the Cross-Language Evaluation Forum's interactive track (iCLEF). For this reason, provisions for recording relevance judgments and automatic generation of event logs are also included. The system uses a client-server model to facilitate remote interaction and centralized data collection. The client is easily reconfigurable to support experiments with contrastive conditions (for example, without user-assisted query translation). The client and server (both of which are coded in Java) are freely available, although integration of corpora, term lists, and InQuery or some other back end system would likely require a nontrivial effort. Nonetheless, we believe that the availability of this system is a useful contribution, both for the insight it can provide into the utility of user-assisted query translation, and as a first step towards reducing barriers to entry in interactive CLIR evaluations. This work has been supported in part by DARPA cooperative agreement N660010028910.

REFERENCES

[1] He, D., Wang, J. Oard, D.W. and Nossal, M. Comparing User-assisted and Automatic Query Translation. In Peters, C. editor, proceedings of CLEF2002.

[2] Pirkola, A. The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. In Proceedings of SIGIR'98, Australia, 1998.