Scenarios for Interactive Cross-Language Retrieval Systems Position paper

Julio Gonzalo
UNED, Madrid
julio@lsi.uned.es

Abstract

This position paper argues that a user perspective is needed to focus mid-term research in Cross-Language Information Retrieval (CLIR), and proposes different search scenarios that pose different challenges for CLIR researchers and represent plausible user needs.

1 Introduction

When designing an Information Retrieval (IR) system, it can be argued that considering the user's perspective is more crucial in a Cross-Language setting than in a monolingual one. While interactive features are meant to improve monolingual search tasks, they become necessary to enable accomplishment of cross-language search tasks.

In monolingual IR, users seem to find out quickly the optimal way to use, and interact with, a system:

- Some results of the Interactive TREC tasks [5] suggest that major improvements in precision/recall do not imply major improvements in user searching tasks. For instance, a simple boolean retrieval system performing poorly in the ad-hoc TREC task can be an excellent IR tool once the user learns to use conjunction, disjunction and term negation in large queries, adapting to the system behavior.

- Users seem able to solve web-based Question Answering tasks fast and accurately using Internet search engines and directories [4], which do not provide sophisticated interaction mechanisms beyond querying and navigating hierarchical directories.

- Interactive features seem necessary to improve searches when the search is not fully predefined (berry-picking model). This is the case of typical Internet searches, where queries are very short (e.g. "mp3", "digital cameras") and search engines return appropriate portals where the user begins (rather than ends) his browsing/searching exploration [3]. Of course, fuzzy user needs are the most difficult to incorporate in a systematic evaluation (such as the interactive TREC).

However, in a complete Cross-Language Information Retrieval (CLIR) system, no information need can be fulfilled without help from the system:

- documents cannot be selected (i.e recognized as relevant) without some sort of translation or indication about its contents [6].

- Queries cannot be enhanced without some information about the foreign language candidate translations of the user terms [8].

In spite of these facts, most CLIR research is yet focused towards non-interactive retrieval models, and potential user communities are not aware of state-of-the-art CLIR capabilities. The

CLIR user needs survey conducted as a part of the CLEF evaluation campaign [2] revealed that there is a community of potential CLIR users and deployers -e.g. developers of Digital Libraries (DL) initiatives- willing to incorporate and use cross-language search capabilities. For this community, however, such capabilities are seen as a part of integral solutions that permit querying, use of multilingual thesauri and metadata, browsing foreign-language documents, etc. Simple retrieval of foreign-language documents is not perceived as a useful feature per se.

A good example of the gap between CLIR research and potential users is multilingual thesauri. It seems that the use of multilingual thesauri to assist searching can be more useful than in a monolingual environment, because selecting adequate descriptors in the user's language can lead to accurate searches in all other languages included in the thesaurus. In addition, the DL community is employing significant efforts in building and merging such thesauri. But the CLIR research community has paid little attention to multilingual thesauri in CL searches, with a few exceptions such as [1].

In the remainder of the paper, we examine a few distinctions that may lead to the consideration of different cross-language retrieval scenarios that demand different types of interaction with the user. In our opinion, such (user-oriented) scenarios should be taken into account when detecting potential research opportunities in CLIR.

2 Types of Information need

Research in Information Retrieval now reflect the fact that different information needs demand different IR approaches. The TREC ad-hoc task, for instance, has disappeared in favor of web search or question answering tasks. Most CLIR evaluations (TREC, CLEF, NTCIR) are currently adopting the TREC ad-hoc methodology. Probably, research in CLIR (including, or even starting with evaluation) should perhaps evolve the same way. Some interesting possibilities are:

Question Answering. How can a system help a user to find the answer to a particular question, even if the answer is expressed in some foreign language? This in an interesting research challenge that differs substantially from the classical CLIR approach (e.g., the approach to data fusion from different languages should be quite different from current approaches), but represents a common, realistic search scenario for many uses of the Internet, especially for non-English language speakers. An example can be a traveler looking for updated bus schedules for a local trip in a foreign country, or a Spanish photographer checking out the latest advances in digital cameras (which will take months to be translated into Spanish).

Image (captions) searching. This is an interesting variant of CLIR, because the user, regardless of the image caption language, can immediately recognize and use retrieved images. This is a very common search scenario, for instance, for journalists using image databases from international press agencies. Query expansion and translation, as well as interaction with the user, has its own features in this retrieval scenario.

Bibliographic searches. Examples are collecting material for background articles, or looking for scientific papers dealing with some more or less specific issue. Such scenarios match better current CLIR evaluations, but the issue of integration with Digital Libraries (semi-structured or distributed data, use of multilingual thesauri) is largely unexplored. Concrete examples include, for instance, an European laws expert searching comparable legislation of some issue across EU countries. He will probably master one or more EU languages, have a passive knowledge of some others, and totally ignore the rest. The task is identifying all relevant documents for (possibly manual) translation. The system should use all information available, e.g. comparison to retrieved documents in the language(s) known to the user, mapping into available thesauri, search across different databases, use of semi-structured data, etc.

Web surfing. By web surfing we mean here web searches without a completely predefined information need. This is probably the most common way of searching the Internet, as proved

by the short, underspecified queries that are prevalent in the main search engines (e.g. "sex", "mp3", "britney spears"). Several Internet portals and search engines already offer the possibility of navigating on-line translations of URLs. Such facilities could be combined with interactive search/browsing systems capable to suggest search concepts across language boundaries. Of course, evaluating such approaches is extremely delicate, as they are meant to work with fuzzy information needs. An example in this direction can be [9], where a phrase browsing approach to interactive retrieval is enhanced with cross-language relevant phrases. The system is evaluated in an observational setting, studying the logs of interactive sessions of real (uncontrolled) users with real information needs in a universitary domain portal.

Speech Retrieval. Cross-Language Speech Retrieval also poses interesting challenges from an interactive point of view. For instance, full Machine Translation should not be used for document selection, because the input from speech recognition systems will be too noisy for standard MT systems.

3 Active vs. Passive language knowledge

There are, at least, two realistic assumptions about language abilities of searchers that should lead to different kinds of interactive CLIR: strictly monolingual (v.g. an average American user looking for Chinese documents) or with some passive knowledge of the target language (v.g. an Italian user looking for documents in French, Spanish or Portuguese).

- **Strictly monolingual**. For instance, an average European user looking for documents in Chinese.
- **Passive knowledge of target language**. For instance, an Italian user looking for documents in French, Spanish, Portuguese, Catalan, etc. Even if he is not able at all to pose queries in such languages, he will probably recognize translations and grasp the content of documents without translation.

In the first case, full translation assistance is required. For query refinement, possible translations for query words must be described, either with definitions in the user language or with inverse translations into the user language. For document selection, the documents must be translated or represented in a way that the user can make judgments about their relevance. A tradeoff between translation quality (for better judgments or even direct use of documents) and computational cost (documents should be quickly translated on the fly, or massively translated offline) must be taken into account. Other problems and questions are: How to facilitate searching in a fully multilingual setting (the user might be annoyed if having to consider translations and refine the query in several unknown languages)? Is it better to search the original documents with query term translations, or to search directly the space of translated documents (avoiding query translation issues)?

In the second case, design issues are rather different. The system must provide ways of crossing language boundaries and offering good candidate translations, but does not need special machinery to describe translations or show document contents. There remain issues such as granularity of candidate translations (Machine Translation of the query, word-by-word translation, intermediate steps such as phrase translation), presentation of results for multiple language searches, feedback, etc.

4 Conclusions

According to the TREC model, the design of increasingly challenging comparative evaluations is an optimal way of focusing community research. That is why, even if evaluating interactive systems is more elusive and more costly than current TREC, CLEF and NTCIR cross-language tasks, we believe that it has to be pursued to connect CLIR research with real search situations. Following

the ideas in [6], in CLEF 2001 we organized an experimental interactive task of (foreign-language) document selection [7]. The track had three participants, and produced interesting hints about the document selection task: Machine Translation proved enough for accurate document selection; word-by-word translation is a low-cost alternative, but far less accurate to discriminate relevant documents; and, finally, MT document selection can be improved using cross-language summaries that permit faster translation and faster judgment than full MT, without loss of precision.

This document selection track has had a continuation in CLEF 2001, where the emphasis has been made in comparing query translation and expansion strategies for broad (multi-faceted) queries. The results of this track will be presented at the CLEF 2002 workshop at Rome.

The next steps to follow are yet under discussion. According to the scenarios presented above, an interesting possibility could be a mid-term focus on Cross-Language Question Answering, for instance:

- CLEF 2002: query translation and document selection on narrow (single-faceted) queries, with few relevant documents in some of the target languages.
- CLEF 2003: Cross-Language Question Answering using Internet; perhaps establishing a common search engine (e.g. Google) to which participants add a layer to permit unrestricted, web-based cross-language searching.

References

[1] F. Gey, H. Jiang, V. Petras, and A. Chen. Cross-language retrieval for the clef collection: comparing multiple methods of retrieval. In *Proceedings of CLEF 2000*, 2001.

[2] J. Gonzalo, C. Peters, A. Peñas, and F. Verdejo. Clef user needs survey. Technical report, CLEF Deliverable 1.1.1, 2002.

[3] M. Hearst. Next generation web search: Setting our sites. *IEEE Data Engineering Bulletin*, 2000.

[4] W. Hersh, L. Sacherek, and D. Olson. Observations of searchers: Ohsu trec 2001 interactive track. In *Proceedings of TREC 2001*, 2001.

[5] W. Hersh, A. Turpin, S. Price, D. Kraemer, B. Chan, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? an analysis from the trec-8 interactive track. In *The Eighth Text REtrieval Conference (TREC-8)*, 2000.

[6] Douglas Oard. Evaluating cross-language information retrieval: Document selection. In Carol Peters, editor, *Cross-Language Information Retrieval and Evaluation: Proceedings of CLEF 2000*, Springer-Verlag Lecture Notes in Computer Science, 2001.

[7] Douglas W. Oard and Julio Gonzalo. The CLEF 2001 interactive track. In Carol Peters, editor, *Proceedings of CLEF 2001*, 2001.

[8] W. Ogden and M. Davis. Improving cross-language text retrieval with human interactions. In *Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS-33)*, 2000.

[9] Anselmo Peñas, Julio Gonzalo, and Felisa Verdejo. Cross-language information access through phrase browsing. In *Applications of Natural Language to Information Systems*, Lecture Notes in Informatics, pages 121–130, 2001.