Towards a Unified Approach to CLIR and Multilingual IR

(Position paper)

Jian-Yun Nie
DIRO
University of Montreal
Email: nie@iro.umontreal.ca

1. Current CLIR approaches

Most current approaches to CLIR make a clear separation between different languages and between the translation and the retrieval steps. For example, the following schema has been used in most of the previous studies:

1. Query translation: Translate a query from one language into several target languages;
2. Document retrieval: Using a translated query to retrieve documents in the same language as the translation;
3. Merging of the results: The results produced in different languages are merged to produce a unique result list.

We can notice following two facts in this approach:

1. The translation has been usually considered as a separate step. In some of the cases (usually when a statistical approach is used for query translation), translation probability is used in the retrieval step; however, the integration is loose. Few studies put the translation step in a whole CLIR process.
2. Different languages are processed separately. It is assumed that the documents in each language form a separate document collection. This makes it necessary to carry out a merging step.

2. Possible problems

Both the translation and the retrieval steps are uncertain. When the two steps are separate, their uncertainties are be integrated in a principled way. In addition, we notice that the measure f uncertainty I the translation step is usually produced from a general translation perspective. It is independent from the final goal of retrieval. For retrieval, the goal of translation may be different from general translation. For example, in general translation, it is usually the case that one prefers to produce a translation that contains commonly used words. For example, if one has to choose between “chaise” (a furniture) and “chaire” (research chair) as French translation of English word “chair”, the first translation will likely dominate if a general translation approach is used. This is because the first meaning is much more used for the English word “chair” than for the second one. O the other hand, the documents dealt with may talk about programs of research grants. In this case, the second translation is the correct one. However, it is difficult to find it with a general translation tool.

In some previous studies (e.g. [Gao et al. 2001] and [Xu et al. 2001]) the translation tool is related to the document collection. For example, one can train a language model with the document collection, which help select the appropriate translation words. Commercial automatic translation systems also offer translation in several specialization areas. However, it is still the case that this

step, although assisted somehow by the document collection, is relatively independent from the retrieval step.

The clear separation between different languages makes it difficult to compare the results in different languages. The previous studies [Rasolofo et al. 2001] on result merging clearly showed that it is difficult to arrive at the same level of effectiveness in a unique collection with a retrieval-and-merging approach. So a better approach is to deal with all the documents as forming a unique document collection, whatever the language is.

3. Proposal: toward a unified CLIR approach

We suggest a unified approach to CLIR in which the translation step is integrated with the retrieval process, and different languages are mixed in the same document collection.

Integration of translation with retrieval

The integration of translation and retrieval in a unified framework may be compared with an inferential approach to IR. In fact, inferential IR tries to infer related expressions of the query (or query expansion) before the retrieval process is carried out. CLIR is similar to an inferential IR in that the translation step can be seen as an inference step that tries to infer the translations of the query.

For example, if one uses a Bayesian network to model inferential IR [Turtle and Croft 1990], then the translation step can be integrated in the network as an additional transition level. Different from a separate approach, the inference step associated with this transition is part of the whole process. One can also model the CLIR process with a language model. In recent years, the use of language models in IR has been a great success [Lafferty and Zhai 2001, Lavrenko and Croft 2001]. It is possible to extend the approach to CLIR by integrating a translation model. This is the approach proposed in [Kraaij et al 2002].

At this level, we can also make a parallel with the statistical translation method in which there are a language model and a translation model. In a similar way, we can model a statistical CLIR approach by integrating a translation model and a retrieval model. This approach is different from current use of statistical translation methods to CLIR in which the translation model is train independently from the document collection. The translation model is only loosely connected with the retrieval process (e.g. by using the translation probability in term weighting). However, the connection is not tuned for the retrieval purposes. In an integrated approach, one can train the whole process statistically for the purposes of CLIR. In other words, the integration of translation within the whole process is also part of the optimization process. Notice that some of the previous approaches work (e.g. [Gao et al. 2001]) in a very similar way, however, the integrated approach has not been identified clearly.

Integration of different languages in multilingual IR

As the retrieval on a unique document collection usually performs better than an approach with separate retrieval then merging, we can consider putting all the documents in a unique collection (if the whole volume can be treated by a centralized IR system). The difference between different languages can be marked with a language tagger associated with each index term. For example, we can identify the French index “chaise” as “chaise_f”. When a query is translated into different languages, then a large query “translation” is created that contains all the index terms in different languages. For example, we may have “chaise_f”, “chaire_f”, “chair_e”, … in a single query. Then the CLIR problem is no longer different from a monolingual IR problem.

One advantage of this approach is that the weights of index terms in different languages may be more comparable, because they are determined in the same way (although the weights may still be unbalanced because of the unbalanced occurrences of index term I the document collection).

Another advantage is due to the removal of the problematic merging process. The retrieval result naturally contains answers in different languages. One may expect a higher effectiveness (if we compare with the previous experiments on result merging).

Finally, we believe that this approach contributes in lower the barrier between different languages. In fact, documents in different languages often co-exist in the same collection. By separating them, we artificially enhance the difference between them. In fact, the difference between languages is not more (likely less) than the difference between areas. In monolingual IR, documents in different areas are often grouped into the same collection. Then why not group documents in different languages in the same collection? Especially, in some of the cases (e.g. on the Web), they appear naturally together.

4. Roadmap of CLIR

In conclusion, we believe that in the coming years, the CLIR community has to work on integrated models for CLIR. Until now, the approaches are rather a combination of translation + monolingual IR model. There is no specific model for CLIR. We believe that an appropriate CLIR model could greatly improve the quality of CLIR, and contribute to a better understanding of the CLIR problem.

To the multilingual IR problem, we believe that we ca no longer work o each language separately. Language should only be considered as just an additional parameter in a unique document collection, rather than a division of document collection. We believe that buy dividing a document collection according to the languages, we are amplifying the difference between languages, and making the retrieval results less comparable. A better solution would be to simply mark the difference between languages with an additional parameter associated to indexes. This unified approach has to be tested in the coming years.

References

Gao, J. Nie, J.-Y., Xun, E. Zhang, J., Zhou, M., Huang, C., Improving query translation for cross-language information retrieval using statistical models, SIGIR 2001, pp. 96 – 104.

Kraaij, W., Nie, J.-Y., Simard, M., Embedding Web-based Statistical Translation Models in Cross-Language Information Retrieval, Submitted to Computational Linguistics, 2002.

Lafferty, J., Zhai, C., Document language models, query models, and risk minimization for information retrieval, SIGIR 2001, pp. 111 – 119.

Lavrenko, V., Croft, W. B., Relevance based language models, SIGIR 2001, pp.120 – 127.

Rasolofo, Y., Abbaci, F., Savoy, J., Approaches to Collection Selection and Results Merging for Distributed Information Retrieval. CIKM'2001, Atlanta, November 2001, 191-198

Turtle, H., Croft, W. B., Inference networks for document retrieval, SIGIR 1990, pp. 1 – 24.

Xu, J., Weischedel, R., Nguyen, C., Evaluating a probabilistic model for cross-lingual information retrieval, SIGIR 2001, pp.105 – 110.