

# Merging Different Languages in a Single Document Collection

Jian-Yun Nie, Fuman Jin

Laboratoire RALI  
Département d'Informatique et Recherche opérationnelle,  
Université de Montréal  
C.P. 6128, succursale Centre-ville  
Montréal, Québec, H3C 3J7 Canada  
{nie, jinf}@iro.umontreal.ca

**Abstract.** Multilingual IR is usually carried out with separate collections, each for a language. Once a set of answers have been found in each language, all the sets have to be merged to produce a unique answer list. In our experiments of CLEF2002, we try to implement a different approach, in which the documents in different languages are mixed in the same collection. Indexes are associated with a language tag so as to distinguish homographs in different languages. The indexing and retrieval processes can then be done once for all the documents. No result merging is required. This report describes our first tests in CLEF2002.

## 1. Introduction

Most current approaches to CLIR make a clear separation between different languages. 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 two following fact in this approach: 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.

The clear separation between different languages makes it difficult to compare the results in different languages. The previous studies [Rasolofoa 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. In so doing, we can avoid the result-merging step, which seems to generate additional problems.

## 2. Our approach

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", "chaise\_e", "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.

The approach we implemented follows the following steps:

- Language identification:
  - In CLEF experiments, as the language of each sub-collection is clearly identified, we do not need to use an automatic language identifier. The language is indicated manually in our experiments.
- Language dependent preprocessing:
  - Stop words in each language are removed separately;
  - Each word is stemmed using the appropriate stemmer of the language,
  - Stems are associated with the language tags, *\_f*, *\_e*, *\_i*, *\_g*, and *\_s*, respectively.
- Indexing of the mixed document collection:
  - All the documents are then indexed using the SMART system.
  - The indexes are weighted according to the usual  $tf*idf$  schema:  
$$tf(t, d) = \log(\text{freq}(t, d)+1);$$
$$idf(t) = \log(N/n(t)), \text{ where } N \text{ is the total number of documents in the mixed document collection.}$$
- Query translation:
  - The original query (in English) is translated separately into French, Italian, German and Spanish. The translation words are stemmed and associated with the appropriate language tag as for document indexes. All the translation words are then put together to form a unique multilingual translation, together with the original query.
  - Each translation word is associated with its translation probability, which is then considered as the weight of that word. The problem we encountered is the weighting of the original query words with respect to the translation words. Several alternatives are tried: the weight of the original words is 1,  $1/n$  (where  $n$  is the number of words in the query). We also tried the following solution: attribute the weight 1 to the original query words, while normalize the weights of translation words so that the maximal translation probability for each language is 1.
- Retrieval
  - The retrieval is performed exactly in the same way as in monolingual retrieval.
  - The output is a list of documents in different languages.

Query translation is performed by statistical translation models trained on parallel Web pages mined with PTMiner [Chen and Nie 2000]. For en-fr, en-it, en-ge, we use the same models as last year [Nie and Simard 2002]. We add en-sp model this year.

### 3. Results of our preliminary experiments

The weighting methods we tried do not seem to work well. In fact, with all the three solutions, we observed that very often, one language dominates in the mixed result list: the first 100 documents retrieved are almost only in that language. This shows that we did not reach at a reasonable balance between languages as we expected. Our intention to mix the documents in a unique collection is to create a better balance between languages, and this may be achieved with the use of the same weighting scheme. However, the result is disappointing. Several reasons may be possible:

- The translation models are trained with different parallel corpora. The size and coverage of the corpora are different. This may result in translation of different quality in different languages. For example, the translation of an English word may be quite concentrated on a few words in one language, but not in another language. This makes the translation probabilities often incomparable.
- The weights we attributed to the original query words are not reasonable. In fact, in our experiments, we only tested a few simple solutions. They may not be the most appropriate ones.
- Finally, in our current approach, query translation is still made independently from the retrieval process. We believe that translation and retrieval should be considered together, as we suggested in [Nie 2002]: Both the translation and the retrieval steps are uncertain. When the two steps are separate, their uncertainties are to be integrated in a principled way. However, we only used the translation probabilities as the initial weights of the query words, which are then combined with the idf factor to

arrive at the final weights. This simplistic method may be greatly improved in order to arrive at a better integration of both such as in [Gao et al. 2001] and [Xu et al. 2001].

Among the three weighting schemes we tested, the one with  $1/n$  for the original query words seems to produce the best results. However, these results are still far below the average performance of the CLEF participants, as we can see in the following figure.

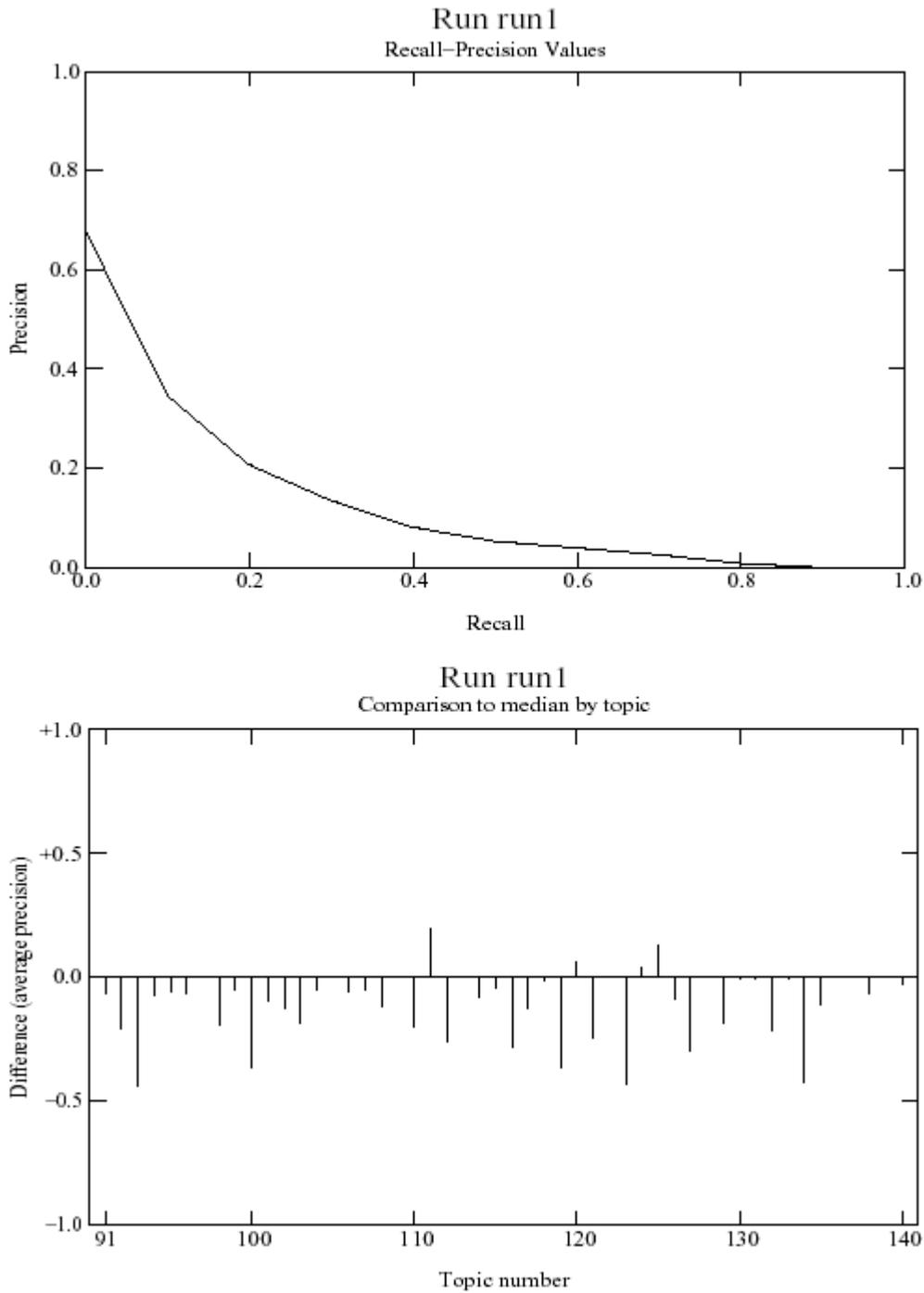


Fig. 1. The best performance we obtained in CLEF 2002.

## 4. Remarks

Despite the disappointing results of our tests, we still believe that our basic idea of mixing documents in the same collection is reasonable. Our current implementation is too simple to show the true potential of this method. In our future experiments, we will try to implement the idea more carefully. The translation probabilities will also be integrated more tightly with the retrieval process.

## References

- [Chen and Nie 2000] J. Chen, J.Y. Nie. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. Proc. ANLP, pp. 21-28, Seattle (2000).
- [Gao et al. 2001] 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.
- [Nie and Simard 2002] J-Y. Nie, M. Simard, Using statistical translation models for bilingual IR, Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001, LNCS 2406, ed. C. Peters, M. Braschler, J. Gonzalo and M. Kluck, Springer, 2002, pp. 137-150.
- [Nie 2002] J.-Y. Nie, Towards a unified approach to CLIR and multilingual IR, *Workshop on Cross-language information retrieval: A research roadmap, 25<sup>th</sup> ACM-SIGIR*, Tampere, Finland, August 2002, pp. 7-14.
- [Rasolofo et al. 2001] Rasolofo, Y., Abbaci, F., Savoy, J., Approaches to Collection Selection and Results Merging for Distributed Information Retrieval. CIKM'2001, Atlanta, November 2001, 191-198
- [Xu et al. 2001] Xu, J., Weischedel, R., Nguyen, C., Evaluating a probabilistic model for cross-lingual information retrieval, SIGIR 2001, pp.105 – 110.