

# Using Parallel Web Pages for Multi-lingual IR

**Jian-Yun Nie, Michel Simard, Goerge Foster**

Laboratoire RALI,

Département d'Informatique et Recherche opérationnelle,

Université de Montréal

C.P. 6128, succursale Centre-ville

Montréal, Québec, H3C 3J7 Canada

{nie, simardm, foster}@iro.umontreal.ca

In this report, we describe the approach we used in CLEF Cross-Language IR (CLIR) tasks. In our experiments, we used statistical models estimated from parallel texts automatically mined from the Web. In our previous experiments, we tested CLIR for English-French and English-Chinese. Our goal of this series of experiments is to see if the approach may be extended to multi-lingual IR (with other languages). In particular, we compare models trained from the Web documents with models that also combine other resources such as dictionaries.

## 1. Introduction

In the past two years, we have tested the possibility of using parallel texts mined from the Web for CLIR [Nie99]. We were able to build models for French-English and Chinese-English translations. In TREC8, the results we obtained were difficult to compare with other systems as the bilingual French-English runs are unofficial, and there is no report on similar experiments in TREC8 runs. This year, we successfully found several sets of parallel Web pages for the following language pairs: English-Italian, English-German, in addition to the English-French corpus we found previously. So our goal in this year's CLEF is to see if the parallel Web documents can also apply to multi-lingual IR.

In our previous experiments, we observed that a certain combination of the translation models with a dictionary could improve IR effectiveness. However, the combination remained empirical: a dictionary translation is attributed a certain "default probability" and it is added to that from a translation model. This year, we combined several resources together in the model-training step: the documents from the Web, the Hansard parallel corpus, and two bilingual dictionaries. The combination is set in such a way that maximizes the translation probability of set of texts (held-out).

In this report, we will first describe the mining system we used to gather parallel texts from the Web. Then a brief description of the training process will be described. Finally, the CLIR experimental results will be summarized.

## 2. Mining parallel texts from the Web

Statistical models have often been used in computational linguistics for building MT systems or constructing translation assistance tools. The problem we often have is the unavailability of parallel texts for many language pairs. The Hansard corpus is one of the few existing corpora for English and French. For other languages, such a corpus is less (or not at all) available. In order to solve this problem, we conducted a text-mining project on the Web in order to find parallel texts automatically. The first experiments with the mined documents have been described in [Nie99]. The experiments were done with a subset (5000) of the mined documents. However, they showed that the approach is feasible. Later on, we trained another translation model with all the Web documents found, and the CLIR effectiveness obtained with it is close to that with a good MT system (Systran).

The mining process is devised into the following three steps:

- selection of candidate Web sites
- finding all the documents from the candidate sites
- paring the texts using simple heuristic criteria

The first step aims to determine the possible web sites where there may be parallel texts for the given language pair. The way we did this is to send requests to some search engines, asking for French documents containing an anchor text such as "English version", "english", and so on; and similarly for English documents. The idea is, if a French document contains such an anchor text, the link to which the anchor is associated usually points to the parallel text in English.

From the set of documents returned by the search engines, we extract the addresses of web sites, which are considered as candidate sites.

The second step also uses the search engines. In this step, a series of requests are sent to the search engines to obtain the URLs of all the documents in each site.

The last step consists of paring up the URLs. We used some heuristic rules to determine quickly if an URL may be parallel to another:

- First, parallel texts usually have similar URLs. The only difference between them is often a segment denoting the language of the document. For example, "-en", "-e", and so on for English documents. Their corresponding segments for French are "-fr", "-f", and so on. Therefore, by examining the URLs of the documents, we can quickly determine which files may be a pair.
- We then use other criteria such as the length of the file to further confirm or reject a pair.
- The above criteria do not require to downloading the files actually. Once a set of possible pairs is determined, the paired files are downloaded. Then we can perform some checking of the document contents. For example, are their HTML structures similar? Do they contain enough text? Can we align them into parallel sentences?

The French-English parallel corpus has been constructed last year at RALI laboratory. This year, we cooperated with Twenty-One (W. Kraaij) to construct English-Italian and English-German parallel corpora, using the same mining system - PTMiner. The following table shows the number of text pairs as well as volume of the corpora for different language pairs.

	E-F		E-G		E-I	
Pairs	18 807		10 200		8 504	
Volume (Mb)	174	198	77	100	50	68

Table 1. Training corpora

The corpora found from the Web will be called WAC corpora (Web Aligned Corpora). The models trained with these corpora will be called the WAC models.

### 3. Principle of building a probabilistic translation model

Given a set of parallel texts in two languages, it is first aligned into parallel sentences. The criteria used in sentence alignment are the position of the sentence in the text (parallel sentences have similar positions in two parallel texts), the length of the sentence (they are also similar in length), and so on [Gale93]. In [Simard92], it is proposed that cognates may be used as an additional criterion. Cognates refers to the words (e.g. proper names) or symbols (e.g. numbers) that are identical (or very similar in form) in two languages. If two sentences contain such cognates, it provides additional evidence that they are parallel. It has been shown that the approach using cognates performs better than the one without cognates. Before the training of models, each corpus is aligned into parallel sentences using cognate-based alignment algorithm.

Once a set of parallel sentences is obtained, word translation relations are estimated. First, it is assumed that every word in a sentence may be the translation of every word in its parallel sentence. Therefore, the more two words appear often in parallel sentences, the more they are thought of to be translation of one another. In this way, we obtain the initial probabilities of word translation.

At the second step, the probabilities are submitted to a process of Expectation Maximization (EM) in order to maximize the probabilities with respect to the given parallel sentences. The algorithm of EM is described in [Brown93]. The final result is a probability function  $P(f|e)$  which gives the probability that  $f$  is the translation of  $e$ . Using this function, we can determine a set of probable word translations in the target language for each source word, or for a complete query in the source language.

#### 4. The training of multiple models and their combination

For English and French, we also have other resources: the Hansard corpus (a set of parallel French and English texts from the Canadian parliament debates), a big terminology database (Termium) and a small bilingual dictionary (Ergane). A translation model is trained from the Hansard data, in the same way as for the Web document (WAC).

In both the terminology database and the bilingual dictionary, we have English words/terms, and their French translations (words/terms). In some way, we can also think of these two resources as two sets of special parallel "sentences". Therefore, the translation probability between words can also be estimated with the same statistical training process. Therefore, two additional translation models are estimated from them. In total, we obtain 4 different translation models between English and French from four different resources (in each direction). The question now is how we can combine them in a reasonable way.

We choose a linear combination of the models. Each model is assigned a coefficient denoting our confidence on it. The coefficient is tuned according to a set of "held-out" data - a set of parallel sentences (about 100K words). This set is selected from different resources (however different from the resources we are using for model training) so that it gives a good balance of different kinds of texts. Finally, the following coefficients are assigned to each model:

Ergane	0.0413916
Hansard	0.300517
Termium	0.413493
Wac	0.244598

As we can see, the combination seems to favor the model which contains more vocabulary. Termium is attributed the highest coefficient because it contains about 1 million words/terms in each language. The Hansard corpus and the Wac corpus contain about the same volume of texts. So their coefficients are comparable. The Ergane dictionary is a small dictionary which only contains 9000 words in each language, its coefficient is very low. Although these coefficients are the best for the held-out data, they may not be suitable to our data in CLEF.

#### 5. Experiments

We used a modified version of SMART system [Buckley85] for monolingual document indexing and retrieval. The *ltn* weighting scheme is used for documents. For queries, we used the probabilities provided by the probabilistic model, multiplied by the *idf* factor. From the translation words obtained, we retained the top 50 words for each query. The value of 50 seemed to be a reasonable number on TREC6 and TREC7 data.

## 5.1. Monolingual IR

Monolingual IR results have been submitted for the following languages: French, Italian and German. This series of experiments uses the SMART *ltn* weighting scheme for queries as well. In addition, a pseudo-relevance feedback is applied, which uses the 100 most important terms among the top 30 documents retrieved to revise the original queries. The parameters used for this process is:  $\alpha = 0.75$ , and  $\beta = 0.25$ . The results obtained are shown below:

	French	Italian	German
$\geq$ medium	18	18	12
$<$ medium	16	16	25
Average precision	0.4026	0.4334	0.2301

Table 2. Monolingual IR

As we can see, a great difference can be observed in effectiveness in the above runs. Several factors have contributed to this.

### 1. The use of stoplist

In the case of French, a set of stopwords is set up carefully by French speaking people. In the case of Italian and German, we used two stoplists found from the Web [Stoplists]. In addition, a small set of additional stopwords was added for Italian.

### 2. The use of a lemmatizer or a stemmer

For French, we used a lemmatizer developed in the RALI laboratory that first uses a statistical tagger, then transforms a word to its citation form according to its part-of-speech category. For Italian and German, two simple stemmers obtained from the Web [Stemmers] are used. There is no particular processing for compound words in German. This may be an important factor that affected the effectiveness of German IR.

Overall, the French and Italian monolingual runs seem to be comparable to the medium performance of the participants; but the German run is well below the medium performance. We think the main reason is due to the lack of special processing on German (e.g. compound words).

## 5.2. Tests on bilingual IR

The bilingual task consists of finding documents in a language different from that of the queries. We tested the following bilingual IR: E-F (i.e. English queries for French documents), E-I and E-G. For this series of test, we first used the translation models to obtain a set of 50 weighted translation words for each query. Unknown words are not translated, and they are added into the translation words with the default probability of 0.05. The same pseudo-relevance feedback process is used.

Between English and Italian, English and German, we only have the Web parallel documents to train our translation models. For French and English, we have multiple translation resources: the Web documents, the Hansard corpus, and two bilingual dictionaries. So we also compare the model with only the Web documents (the WAC model) and the model with all the resources combined (the Mixed model). The following table summarizes the results we obtained for the official submissions in bilingual IR.

	F-E		I-E (WAC)	G-E (WAC)
	WAC	Mixed		
≥ medium	20	16	21	13
< medium	13	17	13	21
Average precision	0.2197	0.1722	0.2032	0.1437

Table 3. Bilingual IR with different models

For F-E and I-E cases, the WAC models lead to an effectiveness that is better than the medium. The Mixed model of F-E gives a medium performance. The comparison between the two translation models for French to English is particularly interesting. We expected that the Mixed model could perform better because it is trained with more data from difference sources. Surprisingly, its effectiveness is worse than the Wac model. There may be several reasons to this:

- The combination of different resources is tailored for a set of held-out data that does not come from the CLEF document set. So there may be a bias in the combination.
- During the combination, we observed that the combination results tend to favor dictionary translations. A high priority is attributed to dictionary translations. This may also be attributed to the biased tuning of combination.

The unreasonable combination may be further illustrated by the following examination of the bilingual IR effectiveness with each individual translation model. The following table shows the effectiveness over all the 40 queries for bilingual IR between French and English:

Model	Wac	Hansard	Termium	Mixed
Avg. precision	0.1989	0.2367	0.1800	0.1426

Table 4. Comparison of bilingual IR with different individual models.

The 0.1426 case over 40 queries corresponds to 0.1722 over 33 queries in the previous table (for the official evaluations). We can see that the mixed model performed worse than any of the individual models. This indicates clearly that the combination of the models is not suitable for the CLEF data.

### 5.3. Multilingual runs

In our case, the multilingual runs are only possible from English to all the languages (English, French, Italian and German). In these experiments, the following three steps have been done:

1. Translate English queries to French, Italian and German, respectively;
2. Retrieve document from different document sets;
3. Merge the retrieval results.

The translation of English queries to German and Italian was done by the WAC translation model (trained from the Web documents). For English to French, we also have the alternative of using the Mixed model. The translation words are submitted to the *mtc* transformation of SMART. This scheme is chosen because it leads to comparable similarities between different data sets, therefore, makes the result merging easier. The merging is done according to the similarity scores. The top 1000 retrieved are selected as the final results and submitted for evaluation.

The following table describes the results of different runs. In the Wac column, all the models used to translate English queries are WAC models. In the Mixed case, only the English to French translation uses the Mixed model, whereas the other translations still use the Wac models.

	Wac	Mixed
$\geq$ medium	14	12
$<$ medium	26	28
Average precision	0.1531	0.1293

Table 5. Multilingual IR

As we can see, these performances are all below the medium performance. One of the main reasons may be that the German monolingual retrieval does not use any linguistic preprocessing, and has a very poor effectiveness. This may greatly affected the multilingual runs. Another possible reason may be the over-simplified merging method we used. In fact, in order to render the English monolingual runs compatible (in terms of similarity values) with other bilingual runs, we had to choose the *mtc* weighting scheme as for the other cases. In our tests, we observe that this weighting scheme is not as good as *ltc*. Therefore, the ease of result merge has been obtained to the detriment of English effectiveness.

We observe again the negative impact of the Mixed model in this task. When the Wac model for English-French is replaced by the Mixed model, the effectiveness decreases. This shows once again that the coefficients we set for different models are not suitable for the CLEF data.

## 6. Final remarks

In this CLEF, we successfully used parallel Web pages to train translation models for language pairs other than English and French.

For monolingual IR, the effectiveness for French and Italian are similar to the medium performance. The German monolingual run is well below the medium. We think the main reason is that we did not carry out any particular processing on German morphology, which is an important problem for German IR.

For bilingual IR between English and French, and between English and Italian, the effectiveness seems to be reasonable. It is better than the medium effectiveness. Between English and German, however, the effectiveness is well below the medium effectiveness. The reason may be the same as for the German monolingual run.

For multilingual runs, the performance is below the medium. We believe the reason is once again the low effectiveness for German. In addition, result merging may also have affected the global effectiveness.

Between English and French, we also tried to combine different resources in our translation models. We used a linear combination of the models trained with different data, and the coefficients are determined by using a small set of held-out data. However, to our surprise, the mixed model performed worse than the model trained with the Web documents only. This clearly indicates that the combination is not reasonable to the CLEF data.

In our future work, we will try to determine a better way to combine different translation models between English and French. For German, we will use more linguistic processing, in particular, a more sophisticated stemmer.

Overall, we are still encouraged by this CLEF because we showed for the first time that the Web parallel documents could be used for multilingual IR.

## References

[Brown93] P. F. Brown, S. A. D. Pietra, V. D. J. Pietra, and R. L. Mercer, The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, vol. 19, pp. 263-312 (1993).

- [Gale93] W. A. Gale, K.W. Church, A program for aligning sentences in bilingual corpora, *Computational Linguistics*, 19 :1, 75-102 (1993).
- [Franz98] M. Franz, J.S. McCarley, S. Roukos, Ad hoc and multilingual information retrieval at IBM, *The Seventh Text Retrieval Conference (TREC-7)*, NIST SP 500-242, pp. 157-168 (1998)
- [Nie98] J.Y. Nie, TREC-7 CLIR using a probabilistic translation model, *The Seventh Text Retrieval Conference (TREC-7)*, NIST SP 500-242, pp. 547-553 (1998).
- [Nie99] J.Y. Nie, P. Isabelle, M. Simard, R. Durand, Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web, *ACM-SIGIR conference*, Berkeley, CA, pp. 74-81(1999).
- [Simard92] M. Simard, G. Foster, P. Isabelle, Using Cognates to Align Sentences in Parallel Corpora, *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal (1992).
- [Stoplists] <http://www.cs.ualberta.ca/~oracle8/oradoc/DOC/cartridg.804/a58165/appa.htm>
- [Stemmers] <http://www.muscat.com>