Using Statistical Translation Models for Bilingual IR

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

Abstract: This report describes our test on using statistical translation models for bilingual IR tasks in CLEF-2001. These translation models have been trained on a set of parallel web pages automatically mined from the Web. Our goal is to compare the following approaches:

- using the original parallel corpora or a cleaned corpora to train translation models;
- using the raw translation probabilities to weigh query words or combine the probabilities with IDF;
- using different cut-off probability values in the translation models (i.e. delete the translations lower than a threshold).

Our results show that:

- the models trained on the original parallel corpus work better than those on the cleaned corpora;
- the combination of the probabilities with IDF is beneficial;
- and it is better to cut-off the translation models at a certain value (0.01 in our case) than not cut them.

1. Introduction

There have been several experiments on using statistical translation models for CLIR [Franz98, Nie99]. It has been shown that with a proper use of the translation models, we can obtain effectiveness comparable to that using a good machine translation system. In our previous studies we have successfully mined several large sets of parallel web pages, and used them to train translation models for the following language pairs: French-English, German-English, Italian-English and Chinese-English. The model training on these corpora has been the same as that using a manually prepared parallel corpus. However, there are several differences between these corpora and a manually prepared corpus. In particular, the corpora with parallel web pages contain much more noise (non-parallel pairs).

Therefore, some special means seem to be necessary to explore the best uses of the parallel web pages in the CLIR tasks. In this report, we describe our attempts to make better translation models from the parallel corpora, and to make better use of the trained translation models in CLEF bilingual IR tasks. The following problems will be investigated in our tests:

- the clean up of the parallel corpora in order to create neater corpora;
- the cut-off of translation models in order to eliminate unreliable translations;
- the use of two-directional query translation;
- and the combination of translation models with bilingual dictionaries to increase the coverage.

In the following sections, let us first recall briefly the corpora and the training of translation models. Then we will report each of the above experiments.

2. Parallel text mining and translation model building

We used the same sets of parallel web pages as the last year. The three corpora of French, Italian and German with English have been constituted automatically [Chen00]. Their sizes are as follows:

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

Table 1. Size of the training corpora

The training of translation models is as follows:

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.

3. Cut-off of translation models

A translation model contains translations for every lexical item encountered in a training corpus, even if the item appeared only once. In the case where an item does not appear often, its translation is often sparse, i.e. it is translated by many different words with low probabilities. In fact, if a source word occurs only once in the training corpus, it may be translated by every word in the target sentence with quite comparable probabilities. Such translations are not reliable. In addition, storing the translations for every word requires a large space. It is not practical to use such a translation model in a real application. For example, the original French-English model contains more than 13 millions entries. Therefore, we exploited the possibility of reducing the translation models by cutting off the translations that we judge unreliable.

Two ways to cut off a model are investigated:

- We only keep a fixed number of lexical items in the model;
- We remove all the translations below a certain threshold of translation probability.

We tested several numbers in the first cut-off method: 1 million (1M), 100 thousands (100K), 10 thousands (10K), 5 thousands (5K) and 1 thousand (1K), and the thresholds tested are 0.05, 0.1 and 0.25. Below is a summary of the results with the CLEF-2000 data[1].

	1M	100K	10K	5K	1K
de-en	0.1684	0.1559	0.1403	0.1212	0.0714
it-en	0.2442	0.2237	0.2426	0.2059	0.0989
fr-en					

Table 2. Cut-off of models by size

[1] Note: The experimental results on fr-en are missing. We are redoing the experiments to complete the missing numbers and will report them at the workshop.

	P≥0.05	P≥0.1	P≥0.25
de-en	0.1693	0.1651	0.1640
it-en	0.2444	0.2524	0.2393
fr-en			

Table 3. Cut-off of models by probabilities

Note: All the queries are translated by the 50 strongest translation words.

We can see that the cut-off by probability is a better solution. In particular, when the threshold is set at 0.1, we obtain generally good results. In fact, the unreliable translations usually have low probabilities. So by eliminating low-probability translations, a large part of unreliable translations is removed.

Given that the words used in the CLEF queries are not rare words, we did not encounter the problem of increasing unknown words by cutting off the models.

4. Cleaning up of parallel corpora

The parallel corpora gathered from the Web contain a certain proportion of non-parallel texts. In fact, it is impossible to eliminate all the non-parallel texts in such a corpus automatically. The question we raised is whether it is helpful to use additional criteria to filter the corpora so that certain non-parallel texts can be eliminated. We did this filtering first on the Chinese-English corpus, also gathered with the same mining tool. The filtering criteria are the following [Nie01]:

- Tighten the control on text length: If the lengths of two presumably parallel texts are too different from the typical length ratio of the two language, then the texts are considered to be non-parallel.
- Control on empty alignment: Once an automatic alignment algorithm is applied on a pair of texts, if the proportion of empty alignments (a sentence is aligned with nothing, or n:0 and 0:n alignments) is higher than a threshold, then the pair is eliminated.
- Using a bilingual dictionary in alignment: In order to increase the accuracy of sentence alignment, we also examine the proportion of words in the sentence that have known translations (stored in the dictionary) in the corresponding sentence. This proportion is taken into account and used in combination with the length criterion in a length-based alignment algorithm [Gale93].

In our tests on Chinese-English corpus, we found that after the filtering, both the translation accuracy judged with 200 randomly selected words and the effectiveness of CLIR between English and Chinese have been increased significantly. Below are the best improvements we have been able to obtain:

Direction	Combination	
	No filter	Best filtering
E-C	161 (80.50%)	183 (91.50%)
C-E	154 (77.00%)	173 (86.50%)

Table 4. Numbers and proportions of correct first translations for 200 words

Direction	Combination	
	No filter	Best filtering
E-C	0.1843 (47.11%)	0.2013 (50.63%)
C-E	0.1898 (49.16%)	0.2063 (53.43%)

Table 5. CLIR effectiveness on two sets of TREC documents

We expected to have similar results with the same filtering on the corpora of European languages. However, our tests showed a decrease in performances. That we can see by comparing the following table with Tables 2 and 3.

	1M	100K	P≥0.05	P≥0.1	P≥0.25
de-en	0.0764	0.0745	0.0777	0.0751	0.0669
it-en	0.2209	0.2418	0.2453	0.2448	0.2363
fr-en					

Table 6. With filtered training corpora.

The results show a large drop of effectiveness in the de-en case. In the case of it-en, it is quite comparable to that without filtering. This counter-performance may be due to several factors:

- The parameters determined for the Chinese-English corpus (that are used during the filtering) are not necessarily suitable to the corpora of European languages;
- There may be a flaw in the training of the translation models, as this was done in a hurry in the last minutes.

We are now investigating both factors, and hope to find an explanation soon.

5. Two-direction translation of queries

It is observed that certain common English words often appear as one of the top-raked translations of queries (e.g. make, provide, due, etc.). This is because these words frequently appear in the training corpora, and they are not considered as stopwords. So they are considered to be highly probable translations for many words in French, German and Italian. However, if we also consider translations of these English words back to the source languages, it would likely be translated to many different words, i.e. their translations would not be concentrated on the same source words for which they have been suggested as translations. Therefore, a combination of both directions in translation could eliminate such common words in query translations. This is the intuition of using the two-direction translation of queries.

In our implementation, the translation probabilities of both directions are multiplied, i.e. for a pair (e, f) of English and French, the probability of P(e|f) is multiplied by the probabilities of $\Sigma_{f'} P(f'|e)$ where f' is all the words in the original French query.

However, our test did not show that this idea or this implementation works well. The following table summarises the results:

	1M	100K	10K	5K	1K	P≥0.05	P≥0.1
de-en	0.1026	0.1337	0.1339	0.1138	0.0545	0.1259	0.1257
it-en	0.2116	0.2149	0.2182	0.1971	0.0945	0.2185	0.2181
fr-en							

Table 7. Two-direction translations

We observe a large decrease in performance compared with Tables 2 and 3. A possible reason is the filtering of the translations: after the translation of each query, we only keep the translations whose probability is higher than a certain threshold (0.001). So in many cases, there are much less than 50 translation words (as in the case of one-directional translations). As a matter of fact, the average numbers of English translation words per query are about 14 for French queries, less than 17 for German queries and less than 16 for Italian queries. This shows that the way we used to keep translation words is too selective. As an example, the first Italian query (on "reconstruction of Berlin") is translated by the following words:

```
berlin          0.242987        architectural    0.048981
cad             0.147230        reconstruction   0.038330
architecture    0.110667        architettura     0.027761
part            0.100185        town             0.026326
rebuild         0.080169        sedan            0.022886
city            0.079734        reconstruct      0.010781
wall            0.063963
```

While we observe that many false translation words have been eliminated, some new ones have been introduced (e.g. cad, part, and sedan), and the word "architettura" is translated by itself, probably due to the appearances of this word in English texts. So globally, we did not obtain any benefit from this two-directional

translation. We will experiments some variances of this method in order to know if it is the idea or our particular implementation method that did not work well.

6. Combination with dictionaries

In the past, we have found that we can greatly improve the CLIR effectiveness by combining a translation model with a bilingual dictionary. The method we tested is to assign a fixed value to those translations that are found in the dictionary. In other words, if a translation is stored in the dictionary, then the weight of this translation will be incremented by a fixed value.

We observed that common words often have more translations than specialised words. As a result, common words will be represented by more translations, and indirectly, their importance in the query is increased unduly. In order to take into account the commonness of a translation word, we combine the translation probabilities with the *idf* value of the translation word. The intuition is to increase the probabilities of the uncommon translation words and lower the probabilities of common translation words.

7. Submitted runs

We submitted 9 bilingual runs, three for each of the language pairs: fr-en, de-en and it-en.

The first group of runs uses only the 50 first translation words and the translation probabilities. The translation models used are cut-off by a threshold of 0.1 (P≥0.1). These runs are prefixed by RaliP01.

The second group combines the models used in the previous group with dictionaries - FreeDict that we found on the Web (http://www.freedict.com/dictionary/index.html). In particular, we used the dictionaries of French to English, German to English and Italian to English. All the translations in the dictionaries are assigned a fixed value (0.001). These runs are prefixed by RaliM001.

The third group also combines with the dictionaries, but the weights attributed to the translations are their *idf*. These runs are prefixed by RaliMidf.

The effectiveness (non-interpolated average precision) of these runs are summarised in the following table:

	RaliP01	RaliM001	RaliMidf
fr-en	0.3499	0.3564	**0.3685**
de-en	0.2124	0.2188	**0.2565**
it-en	0.2731	**0.2742**	0.2562

Table 8. Summary of the effectiveness of the submitted runs

Globally, we see that the combination with a dictionary (in one of the two ways) leads to some improvement. In particular, for fr-en case, the RaliMidf case is 5.3% better than the RaliP01 case. For de-en, the improvement is even greater (20.7%). This is probably because of the relatively poor coverage of German-English corpus (this corpus is the smallest among all we have mined). On the other hand, there is no observable advantage for the it-en case to combine with a dictionary.

Our focus in this CLEF was on fr-en case, because for both French and English, we have a good lemmatizer (in fact, we transform each word in its form of citation). For the two other language, a simple stemmer found from the Web (http://www.muscat.com, however, the site does not seem to contain the stemmers) was used. For German, this is certainly not enough because of agglutination.

The three best submitted runs are compared with the medium effectiveness in the following table:

	RaliMidfF2E	RaliMidfD2E	RaliM001D2E
≥ medium	41	27	27
< medium	6	20	20

Table 9. Comparison with medium effectiveness

From this first comparison, we see that the RaliMidfF2E run (the run using a translation model mixed with dictionary translations whose weights are determined by *idf*) is quite good and well above the medium performance. On the other hand, for Italian and German queries, the performances are quite comparable to the medium performances. We think that the main reason is that we do not have a good stemmer for both languages.

The good effectiveness of the fr-en run shows that it is a good idea to merge a translation model with a bilingual dictionary and to weight the dictionary translations by *idf*. In so doing, the statistical translation model can be completed by the dictionary, and the *idf* factor assigns the dictionary translations with a reasonable weight.

8. Final remarks

In this series of tests on bilingual IR, we wanted to test several ideas:

- the use of cut-offs of statistical translation models,
- the clean-up of the training corpora,
- the use of two-directional query translation,
- and the combination of translation models with bilingual dictionaries.

Model cut-off was shown to be not only a way to reduce the space required to store the translation models, but also a way to increase the quality of query translation. This is due to the elimination of many unreliable translations during the cut-offs. In addition, we found that it is better to cut off the models according to the translation probabilities than by size of the model.

However, the clean up of initial training corpora did not show any improvement in the CLIR tasks. This seems to be contradictory with our recent experiments on Chinese-English CLIR. We would like to further examine the processes of our clean up and model training in order to find the reasons for this.

The use of two-directional query translation did not improve the CLIR performance either. Several factors in our implementation may have affected this test, in particular, the setting of translation probability threshold of 0.001 that leads to few translation words per query. Additional tests are undertaken in order to find a more satisfactory explanation.

Finally, the combination of translation models with bilingual dictionaries is, in general, beneficial. In particular, for the case of French-English CLIR that we focused on, we observed a certain degree of improvement. In this case, we found it a good idea to assign a weight to the dictionary translations according to their *idf* values. This would assign a common translation word with a low weight, and a fairly rare word with high weight. This approach seems to be very intuitive.

We are doing additional test. We hope to be able to report a more thorough account of this series of tests at the workshop.

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).

[Chen00] J. Chen, J.Y. Nie. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. Proc. ANLP, pp. 21-28, Seattle, 2000.

[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)

[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).

[Nie01] J.Y. Nie, J. Cai, Filtering noisy parallel corpora of web pages, IEEE symposium on NLP and Knowledge Engineering, 2001 (to appear)

[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).