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 tests on applying 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 previous studies have shown the utility of such corpora for cross-language information retrieval. The goal of the current tests is to see how we can improve the quality of the translation models and make best uses of them. Several questions are considered: Is it useful to consider the IDF factor in addition to the translation probabilities? Is it useful to further clean the training corpora before model training or the translation models themselves? How could we combine the translation models with bilingual dictionaries? Although our test do not allow us to answer all these questions, they provide useful indication to several further research directions.

1. Introduction

In addition to the problems of monolingual IR, CLIR has some additional aspects to be considered, in particular, the selection of the appropriate translation terms/words, and the proper weighting of these terms [Grefenstette98]. The selection of translation words can be done either through a bilingual dictionary prepared manually, or through the use of a parallel corpus. A bilingual dictionary can be exploited directly by a translation process in a more or less complex manner (e.g. select all the possible translation words, or select some of them and assign a weight to them according to statistical information). It can also be incorporated in a machine translation system. In that case, the selection of translation words is usually done by using rules prepared manually. Whatever the manner, the translation by a dictionary is always limited to those stored in the bilingual dictionary. The question that we can ask is whether the dictionary can contain all the appropriate translations. In many cases, the answer is no for several reasons.

1. A dictionary of general domain is often covers the most frequently used meanings of words. They usually contain few meanings in specialized domains.
2. Languages are evolving. A word may be associated with a new meaning. The terminology used in computer science is a typical example (e.g. the word "web"). A manually prepared dictionary often cannot incorporate these new words or new meanings promptly.

A possible alternative to a dictionary is a statistical translation model. A statistical translation model (TM) is a probability function that is trained on a large set of parallel texts. It observe the uses of words and their co-occurrences in the corresponding sentences in two languages in order to extract the translation relationships between words in the two languages. Using a TM has several advantages over the use of a dictionary (either in a simple way of within a machine translation system):

1. The training of a TM can be completely automatic. No (or little) manual preparation is required.
2. The resulting TM reflects well the word usage in the training corpus. This offer the possibility to train specialized and up-to-date TMs.

There have been several experiments on using statistical translation models for Cross-Language Information Retrieval (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.

However, a critical aspect of using TM is the availability of the training corpora. For a few languages, there are parallel corpora ready to be used (e.g. for English-French). However, such corpora are not available to many languages pairs. Therefore, we conducted a project on automatically mining parallel web pages on the Web [Nie99, Chen00]. This has been very successful: using some simple criteria, we were able to gather large sets of parallel web pages for the following language pairs: Chinese-English, French-English, German-English, Italian-

English[1]. In our previous research, several approaches have been tested on the use of these parallel corpora for CLIR. However, a number of questions remain unanswered. In our tests in CLEF2001, we conducted several new alternatives in order to see how these corpora can be better used for CLIR. The following problems will be investigated in our tests:

- combination of translation probabilities with the IDF factor to improve term weighting;
- 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

For our bilingual IR tasks, we the parallel corpora of French, Italian and German with English are used. Their sizes are shown in the following table (where raw data contain HTML markups, while most of them are removed from the cleaned data):

	fr-en		de-en		it-en	
# Text Pairs	18 807		10 200		8 504	
Raw data (MB)	198	174	100	77	68	50
Cleaned data (MB)	155	145	66	50	50	35

Table 1. Size of the training corpora

Note: Through out this report, we will use fr, de, it, en to designate French, German, Italian and English.

Let us briefly describe the model training process.

1. Sentence alignment

Given a set of parallel texts in two languages, it is first aligned into parallel sentences. The goal of this process is to limit the translation relationships between words within the corresponding sentence. The criteria used in sentence alignment are basically the order of the sentence in the text (parallel sentences are in similar orders in two parallel texts), the length of the sentence (their length ratio is comparable with the standard length ratio of the two languages) [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.

2. Model training with EM

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 iteratively the expectation of probability with respect to the aligned sentences. The details of EM algorithm is described in [Brown93]. The final result is a probability function $P(t|s)$ which gives the probability that a source s is translated to a target word t.

Using this function, we can determine a set of probable word translations in the target language for each source word of the query, or for a complete query in the source language. In our tests reported here, we use the second approach.

Our previous research has tested the number of translation words per query to be retained. It turned out that a number between 20-50 is reasonable for queries in TREC and CLEF. In our current tests, we use 50 translation words (or less, if there are no 50 translation words). A threshold of 0.0001 on the translation probabilities is also

[1] The mining of parallel corpora for German-English and Italian-English has been done in collaboration with Wessel Kraaij of TNO.

used: translations with a probability lower than this threshold are not retained. These turned out to be some effective ways to remove unreliable translations.

3. Translation probabilities alone v.s. prob*idf?

The translation models provide a set of most probable words and their probabilities. A straightforward use would to consider this list as a new query without further treatment.

However, the probability P(*t*|*s*) provided by a TM can only tells how probable *t* is a translation to *s* with respect to the training corpus. It does not tell how this translation fits the text collection searched, or how specific a translation word is with respect to the specific text collection. This second factor is important in IR. In fact, in the weighting scheme tf*idf, we explicitly incorporate the specificity (i.e. IDF) of index terms to some documents. We need to proceed in a similar way in weighting translation words.

Therefoe, the following comparison was made: using only the translation probabilities as the weights of translation words in the new query, or combining them with the IDF factor. The incorporation of IDF has been simple in the SMART system that we use for out experiments. The two weightings correspond respectively to "nnn" and "ntn" in SMART.

The following table and figures show the comparison for Clef2000 and Clef2001 queries:

	de-en		fr-en		it-en	
	nnn	ntn	nnn	ntn	nnn	ntn
Clef 2000	0.1589	0.1613	0.2534	0.2463	0.2349	0.2201
Clef 2001	0.2097	0.2360	0.2572	0.2635	0.3301	0.3577

Table ??. Comparison of weighting of translation words.

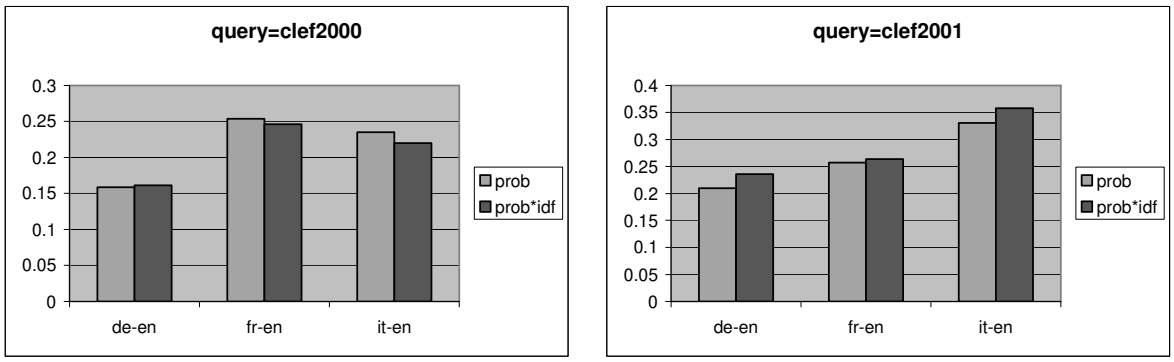


Figure 1. Comparison of weighting of translation words.

We can see that the impact of adding IDF is not always the same. For Clef2000 queries, we observe decreases in retrieval effectiveness (average precision), while for Clef2001, there are improvements. Note that the two sets of queries are evaluated against the same set of documents. So the impact of IDF only depends on the queries.

A possible factor is the correspondence between the queries and the document collection. A hypothesis is that, if the queries correspond well to the document collection (i.e. there are more relevant documents for them), then the IDF factor seems to have a positive impact. This hypothesis seems to correspond well to the results we have: We have better retrieval effectiveness for Clef2001 queries than for Clef2000 queries. Therefore, the IDF factors may be more reasonable for the words in the translations of Clef2001. However, we do not have stronger evidence now because we only have two sets of queries to compare. Nevertheless, this hypothesis seems to be reasonable in the following sense. The IDF factor essentially tries to adapt the original weighting to the document collection. If the translation words of queries are

*****

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 low-probability translations are usually not reliable. Although we used the number of translation words retained (50) and the probability threshold as means to filter out some unreliable translation words, this is not enough. Many inaccurate translation words still appear in the translations. Therefore, a stricter filtering is necessary.

In addition, the storage of a huge translation model requires a large space. It is not always 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.

Using the first strategy, we created models of the following sizes: 1 million (1M), 100 thousands (100K), 10 thousands (10K), 5 thousands (5K) and 1 thousand (1K). The thresholds 0.05, 0.1 and 0.25 are used in the second strategy.

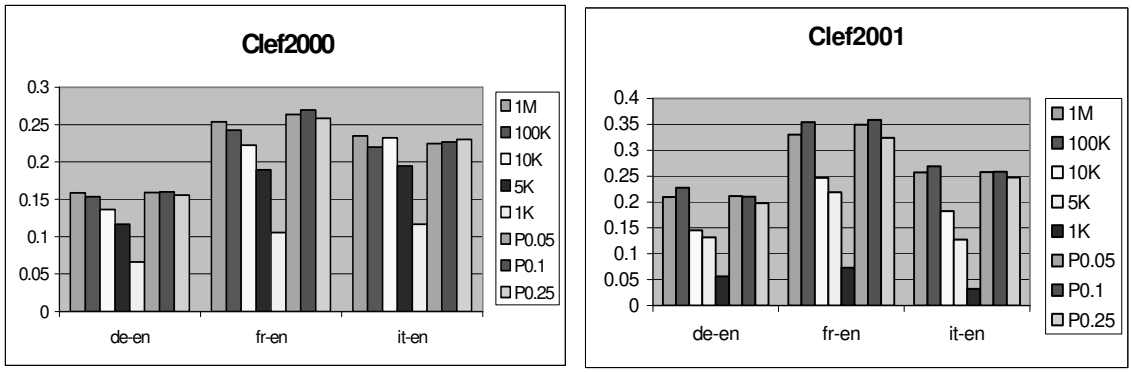


Figure ??. Effects of model cutoffs.

We can see that the cut-off by probability is a better solution. When cutting with model size, except for the size 100K in Clef2001, it leads to decreases in effectiveness in the other cases. When using a cutoff with a probability threshold, in several cases, an improvement is observed, in particular, for P0.1.

The experiments shown in above figures use TMs in a straightforward way. ***

4. Cleaning up of parallel corpora

The parallel corpora gathered from the Web contain a certain proportion of non-parallel texts. If we can filter out this part of non-parallel texts, the translation models, and the CLIR effectiveness, could be improved. However, it is impossible to eliminate all the non-parallel texts in such a corpus automatically. While we remove some non-parallel texts, there will also be some truly parallel texts removed. In addition, the filtering process could also result in a smaller corpus, maybe too small for model training. So the question we raise is whether it is helpful to filter the corpora in some manner.

In our sentence alignment process, we observed that a non-parallel text pair usually has more difficulty to be aligned. In particular, there will be likely a large proportion of empty alignments, i.e. sentences that failed to get aligned with anything. Therefore, the principle we use for the filtering is essentially to test the proportion of the empty alignments. If the empty alignments are higher than a certain proportion, then we consider the text pair to be non-parallel, and it is removed from the corpus [Nie01]. In addition, we also notice that sentence alignment relies only on sentence length. Even in the case where two sentences have very different lengths, the length-based alignment algorithm would always end up with an alignment which is the best among all the possible ones. This does not mean that the alignments are always reasonable. In our analysis, we observed that a certain number of sentences of very different lengths have been unduly forced to get aligned. To remedy this problem, we try to favor alignment of sentences of truly comparable lengths. That is, if two sentences have a length ratio that is similar to the standard length ratio of the two languages, then a bonus value is added to the alignment score of the sentences. This score is the one used in the sentence alignment algorithm.

In addition, we also observe that "known translation" is another useful criterion for sentence alignment. By known translation, we mean the translations that are known by the machine, e.g. those translations stored in a

bilingual dictionary. This factor is used by human beings. In fact, for a human being who does not know every word in a pair of sentences, but only a part of them, one possible approach he would use to judge if they are parallel is to see how many of the known words can be translated mutually. In other words, he would use the translations known to him in his judgment. We exploit the same principle in our alignment process. The score of sentence alignment is increased by a factor that is proportional to the amount of known translations we can find in the sentences.

The above approach has been recently tested on a Chinese-English corpus of web pages, mined with the same mining tool [Nie01]. We tested both the quality of resulting translation models (with 200 randomly selected words) and CLIR performance. The following tables describe the results we obtained.

Translation accuracy	Combination	
	No filter	Best filtering
E-C	80.50%	91.50% (+13.66%)
C-E	77.00%	86.50% (+12.34%)

Table 4. Accuracy of translation models and improvements with the filtered corpora for ch-en.

	Combination	
Direction	No filter	Best filtering
E-C	0.1843	0.2013 (+9.22%)
C-E	0.1898	0.2063 (+8.69%)

Table 5. CLIR effectiveness on two sets of TREC documents

For the de-en, fr-en and it-en corpora, we applied the same filtering criteria as for ch-en. The sizes of corpora have been reduced after the filtering, as shown in the following table:

	de-en		fr-en		it-en	
	de	en	fr	en	it	en
Original	66M	50M	155M	145M	50M	35M
Filtered	42M	42M	147M	139M	34M	33M

Sizes of original and filtered corpora

Despite the differences between the languages, we still expected to have some degree of improvement after the filtering. However, as shown in the following figures, the effectiveness is generally reduced in every case (except for Clef2000, fr-en case).

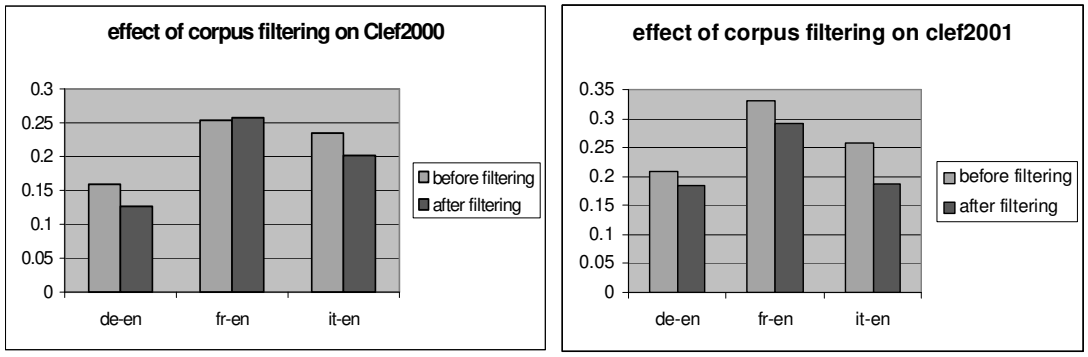


Figure ??. Effect of corpus filtering for 1M model.

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;
- In particular, the filtering process may rely too much on the comparable-length factor. In fact, if we observe the sizes of the resulting corpora after the filtering, we see that the sizes for two languages become surprisingly similar. Even in the case of manually constructed high-quality parallel corpora, there is much larger difference in size. This means that the filtering process is not perfect, although it worked well for Chinese-English.
- A certain number of researchers in computational linguistics believe that "more data is better data" whatever the quality of it is. It is too early to say that our decreased results with smaller (and possibly better) corpora confirm this belief, but this is another possibility to examine closely.

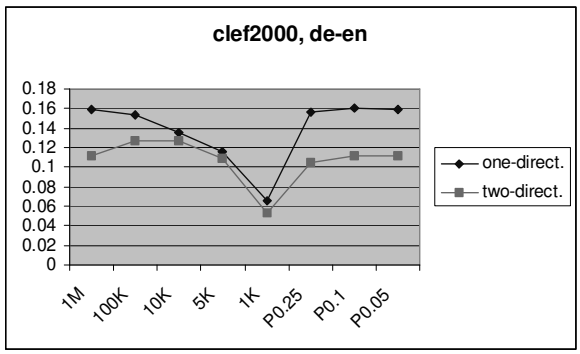
We will investigate other filtering method in order to arrive at better results. For the following test, we will only use the unfiltered corpora.

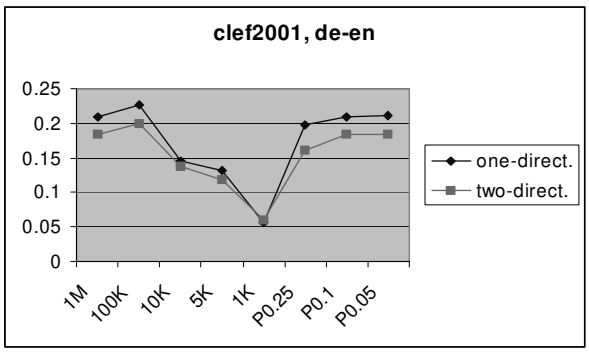
5. Two-direction translation of queries

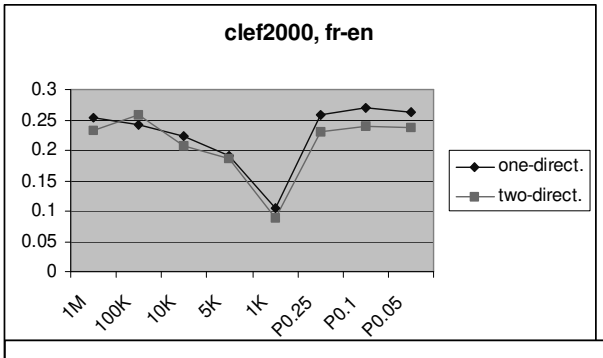
It is observed that certain common English words often appear as one of the top-raked translations of queries (e.g. make, provide, due, etc.). This is because these words frequently appear in the training corpora, and they are not considered as stopwords. So they are considered to be highly probable 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.

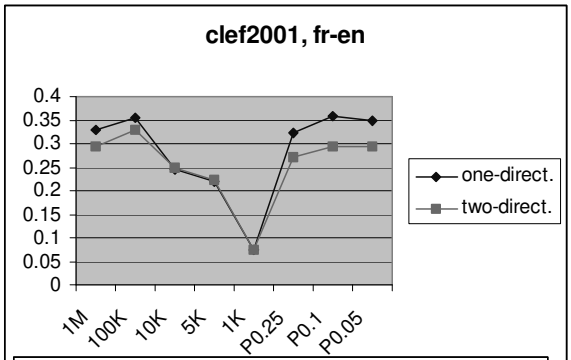
In our implementation, the translation probabilities of both directions are multiplied, i.e. for a pair (e, f) of English and French, the probability of $P(e|f)$ is multiplied by the probabilities of $\Sigma_{f'} P(f'|e)$ where f' is all the words in the original French query, then normalized.

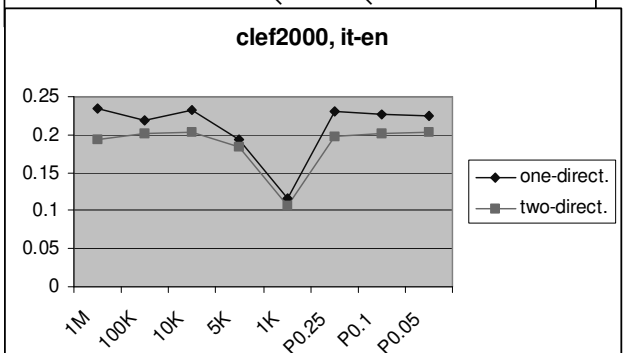
However, our test did not show that this idea or this implementation works well. The following figures show the results:

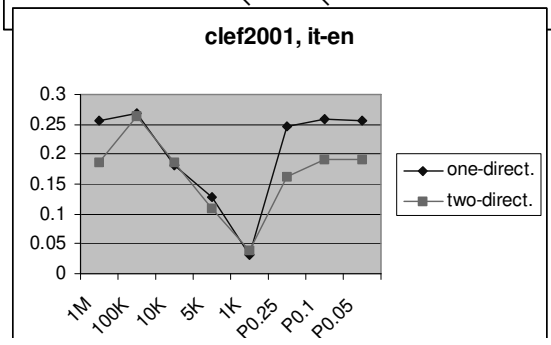












We observe a large decrease in performance when two-direction translation is used. A possible reason is the filtering of the translations: after the translation of each query, we only keep the translations whose probability is higher than a certain threshold (0.0001). So in many cases, there are much less than 50 translation words that we want to retain. The following table compares the number of translation words per query in the two translation results. We can see that the two-direction translation retains much less than one-direction translation.

	Clef2000			Clef2001		
	de-en	fr-en	it-en	de-en	fr-en	it-en
One-direction	28.4	43.3	39.2	30.2	43.7	38.9
Two-direction	23.0	36.7	35.7	22.4	36.5	35.5

Table ??. Number of translation words per query.

Another fact that we observe is that, although there is a relative concentration of probability value to the top-ranked translation words in the two-direction translation (i.e. the top-raked words have a higher probability), many undesirable translation words still remain in the query translation. Their probability values are often less, but their rank in the translation is not necessarily decreased. As an example, we can examine the translations of the first Italian query in Clef2001("reconstruction of Berlin"):

One-direction translation:

```
0.10420 berlin
0.08136 architecture
0.05469 wall
0.04925 part
0.04849 cad
0.04206 find
0.03932 general
0.03665 aspect
0.03635 city
0.03364 architectural
0.02988 rebuild
0.02512 documentation
0.02395 sedan
0.02060 relate
0.01870 town
0.01773 reconstruction
0.01694 detail
0.01268 wait
0.01218 trovato
0.00981 special
0.00873 saloon
0.00829 regard
0.00780 fall
0.00683 reconstruct
0.00672 library
0.00647 architectures
0.00577 generali
0.00483 parte
0.00480 muro
0.00468 expect
0.00422 architectonic
0.00410 compartment
```

Two-direction translation

```
0.13189 berlin
0.11845 architecture
0.07879 cad
0.05806 part
0.05322 general
0.04290 rebuild
0.04267 city
0.04002 documentation
0.03464 find
0.03423 wall
0.03319 aspect
0.02720 architectural
0.02080 reconstruction
0.01464 town
0.01242 sedan
0.00862 generali
0.00772 trovato
0.00646 relate
0.00584 reconstruct
0.00564 parte
0.00471 trova
0.00462 architectures
0.00429 wait
0.00363 regard
0.00326 detail
0.00212 architectonic
0.00161 expect
0.00129 library
0.00123 fall
0.00111 murano
0.00109 special
0.00088 start
0.00079 generally
0.00074 informations
```

We observe that some good translation words (e.g. berlin, architecture, rebuild) have been enhanced. Some undesirable words (e.g. cad, part, general, documentation) have also been enhanced. So the enhancement is not only for the desired words. This shows that the two-direction translation may not be as effective as we expected (at least in the way we did it).

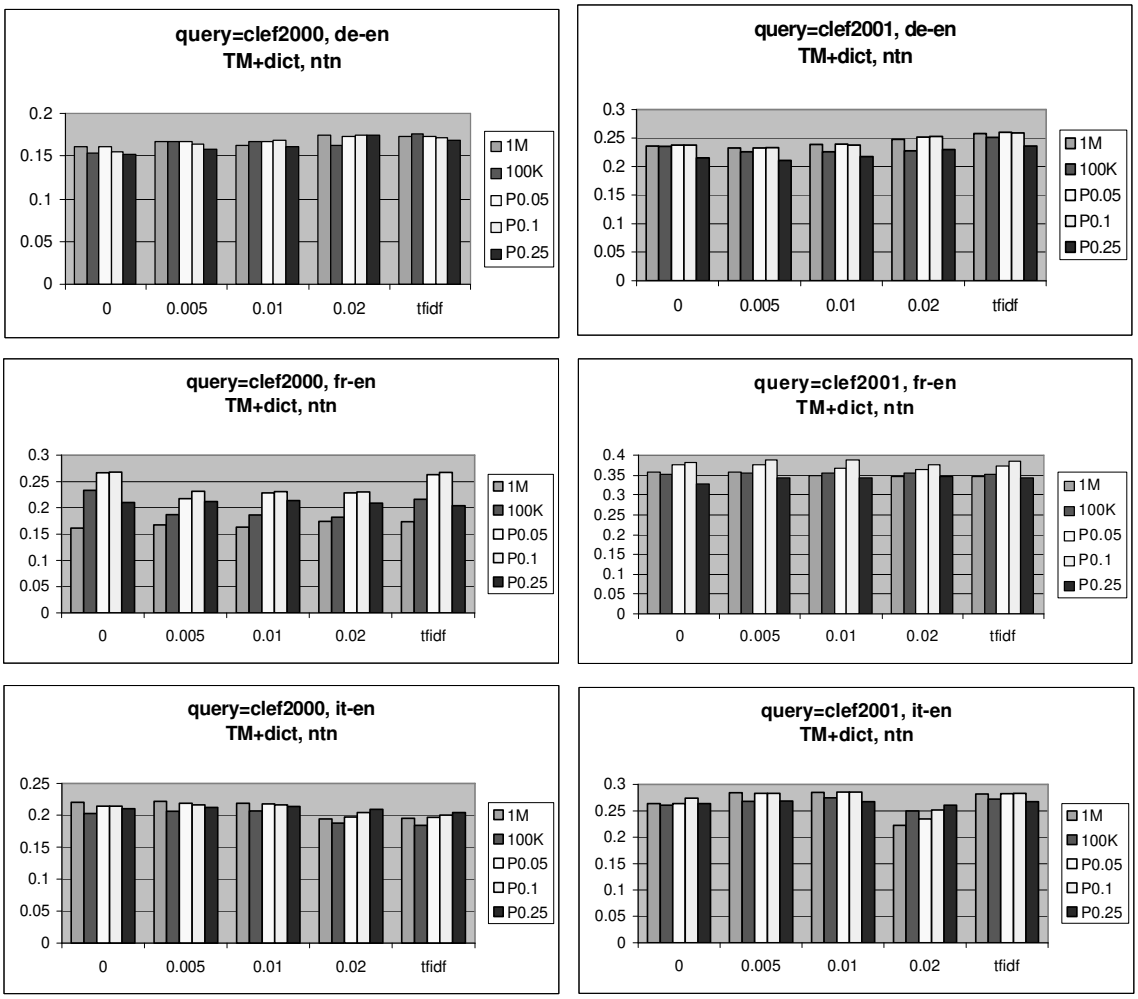
Another phenomenon we observe is that untranslated words (e.g. generali, trovato) have higher ranks in the two-direction translation. This is expectable: These words only appear in some untranslated sentences in the training corpus. So their translation by themselves is strong in both directions. They are thus enhanced in the two-direction translations. The inclusion of these untranslated words is harmful (they do appear in some of the English documents used in CLEF experiments).

These problems point us to some possible improvements of the two-direction translation in the future.

6. Combination with dictionaries

In our previous research, we found that we can greatly improve the CLIR effectiveness by combining a translation model with a bilingual dictionary. The method we used 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 increased by a fixed value (0.01 seemed to be a good value).

We observed that common words often have more translations than specialized 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 to lower the probabilities of common translation words. The following fiigures show the experiments of combining TMs with bilingual dictionaries. The dictionaries are FreeDict that we found on the Web (http://www.freedict.com/dictionary/index.html). In particular, we used the dictionaries of French, German and Italian to English.



Figures ??. Combination of TM with dictionaries.

Globally, we see that the combination with a dictionary (in one of the two ways) leads to some improvements. For de-en, a combination with dictionary is generally helpful. In particular, when the dictionary translations are weighted with tfi*df scheme, we obtain the best results.

For fr-en, the results are more variable. The combinations with fixed weighting do not seem to help in Clef2000, while these combinations have little impact on Clef2001. However, again, the combination with tf*idf weighting leads to some improvements in both Clef2000 and Clef2001.

For it-en, the best combinations are those with fixed values 0.05 and 0.1. These combinations bring higher improvements in Clef2001 than in Clef2000.

Considering the combination of TM filtering and dictionary combination, it seems that the best model filtering is with a threshold of 0.1 on translation probability. The dictionary combination with tf*idf weighting seems to be a good choice in most cases.

8. Final remarks

In this series of tests on bilingual IR in CLEF, 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 cutoffs. 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.

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).

[Grefenstette98] G. Grefenstette. The Problem of Cross-Language Information Retrieval. In *Cross-language Information Retrieval*. Kluwer Academic Publishers. pages 1-9, 1998

[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, pp. 453-458, (2001).

[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).