Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web

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

ABSTRACT

This paper describes the use of a probabilistic translation model to cross-language IR (CLIR). The performance of this approach is compared with that using machine translation (MT). It is shown that using a probabilistic model, we are able to obtain performances close to those using an MT system. In addition, we also investigated the possibility of automatically gather parallel texts from the Web in an attempt to construct a reasonable training corpus. The result is very encouraging. We showed that in several tests, such a training corpus is as good as a manually constructed one for CLIR purposes.

Keywords: probabilistic translation model, cross-language information retrieval, text mining.

1. Introduction

In addition to the classical IR tasks, cross-language IR (CLIR) also requires that the query (or the documents [7]) be translated from a language into another. In this paper, we investigate several approaches to translate an IR query into a different language.

There are three groups of possible approaches: using a machine translation (MT) system, using a bilingual dictionary or terminology base, and using a statistical/probabilistic model based on parallel texts.

At first glance, MT seems to be the ideal tool for CLIR. However, it should be stressed that MT and IR have widely divergent concerns. First, observe that MT systems tend to spend a lot of effort trying to produce syntactically correct sentences. This effort has little, if any, incidence on current IR approaches which are usually based on single words. Second, MT systems are expected to select one of the many translations that words may have. For example, in translating the English word "organic" the MT process will be led to select between the French words "organique" and "biologique". Generally speaking, this selection process is very difficult and MT systems often end up selecting the wrong target language equivalent. In addition, in many cases the multiple possible choices are indeed synonyms or closely related words. By limiting the selection to only one word, the MT process prevents the IR system from expanding the original query by synonyms or related words. Finally, another major obstacle to using MT in CLIR is the unavailability of MT systems for many language pairs, and developing them would take enormous time and human resources.

At the opposite end of the spectrum, one can use either an ordinary general-purpose dictionary or a technical terminology database for query translation. Note that in any sizable dictionary most words receive many translations that correspond to different meanings. Despite many studies to disambiguate word meanings (e.g. [1]), it is still very difficult to determine the correct meaning, thus translations, of a word in a query. This is particularly true in the Internet application in which most user's queries are 2-3 words long. Therefore, the resulting target language query is likely to engender a lot of noise.

The third approach is to determine translational equivalence automatically, on the basis of a corpus of parallel texts (that is, a corpus made up of source texts and their translations). One way of doing this is to start by establishing translation correspondences between units larger than words, typically sentences using now well-known methods [10, 17]. Then, given a sentence S in a source language, it is possible to determine the probability $P(t|S)$ of having the word t of the target language in its translation. Using this probability, we can determine a set of most probable words as the translation of an IR query S. Compared to the previous approaches, this has the following advantages:

- There is no need to acquire or to compile a bilingual dictionary or a complete MT system.
- Word translations are made sensitive to the domain, as embodied by the training corpus.
- As we will see below, it is relatively easy to obtain a suitable degree of query expansion based on translational ambiguity.

In this paper, we will describe the construction of a probabilistic translation model using parallel texts and its use in CLIR. Our experimental results will show that the probabilistic model may achieve comparable performances to the best MT systems.

Parallel texts have been used in several studies on CLIR [2, 6, 19]. Although the principle of using parallel texts in CLIR is similar, the approaches used may be very different. In [19], for example, an IR-like technique is used to find statistical association between words in two languages. [2] used the same training parallel corpus as [19]. However, he first tries to construct a thesaurus from parallel texts using co-occurrence information. A word in an original query is translated by the corresponding words in the target language stored in the thesaurus, together with the co-occurrence information. In [6], a source language query is first used to retrieve documents in a parallel corpus, then a new query is derived from the corresponding documents in the target language. In our case, the approach is based on one of the theoretical models for machine translation described in [3]. The core of the model is the probability $p(t|s)$, the probability of having a word t in the translation of a sentence containing a word s. The construction of the model follows a more strict method than the approaches mentioned above.

The critics one often raises with regard to the use of parallel texts concerns the availability of reliable parallel text corpora. It is true that there are only a few parallel text corpora available. However, we note that it is easier to build a parallel corpus than to build an MT system. In fact, there are many parallel texts in translation services that may be exploited for this end. In addition, the Internet is also a source of parallel texts, as many sites provide documents in two or more different languages. To prove the feasibility of using Web documents as training data, we implemented a simple automatic mechanism to gather parallel texts from the Web. This provides us with a huge amount of parallel texts for English-French. Our tests using a model trained on the first 5000 documents showed that such a parallel corpus may be as good as a manually constructed corpus in several cases. This shows that the availability of parallel texts for CLIR is not an unsolvable problem, at least for French and English.

2. Building a Probabilistic Translation Model

The history of MT showed clearly that building a full-fledged machine translation system to replace human translators is, if not impossible, extremely difficult. An alternative that emerged (particularly in recent years) is to construct automatic tools to help human translators in their translation task [9]. One research direction aims to take advantage of previous translation examples in order to predict possible translations for a new sentence. Most work in this direction uses a parallel text corpus to build a probabilistic translation model.

By a probabilistic translation model, we mean a mechanism which associates to each source language sentence (or query) S a probability distribution $p(T|S)$ on the set of sentences T of the target language. Given such a model, we are able to determine the most probable translations Ts of S to suggest to the user. A precise description of a family of such models can be found in [3]. Roughly speaking, the principle of model training is as follows: given a set of parallel texts in the source and the target languages, if two elements often co-occur in the parallel texts, then they have a high chance to be the translation of one another. In most of the models in [3], linguistic constraints of the two languages (such as word-order constraints) are taken into account. This is necessary to produce a correct translation of the source sentence. Our requirements are much less; we only need a set of translation words to feed an IR system. This set of words is the best translation candidates of the words in the source query S. So the goal of the probabilistic model in this study is to provide the probability $p(t|S)$ – the probability of having the word t of the target language in the translation of the source sentence S. As our goal is not to provide a correct translation, we do not consider the constraints of the two languages. This model corresponds roughly to “model 1” of [3].

Let us give a simple description of the training of this model. Given a single alignment a_k between a source sentence S and its translation T, both sentences are considered as sets (because positions are not considered) of words: $S = s_1, s_2, \ldots, s_l$ and $T = t_1, t_2, \ldots, t_m$. From this alignment, each word s_i in S is considered related to each word t_j in T. In addition, for a word in the source sentence, all the words in the target sentence T are assumed to be equiprobable translations. So, we have

$$p(t_j| s_i, a_k) = C_T / l$$

where C_T is a parameter set to account for the length of the target sentence.

Now, given a set of parallel sentence alignments A, the probability $p(t_j| s_i, A)$ is determined by the sum of all $p(t_j| s_i, a_k)$:

$$p(t_j| s_i, A) = C_A \sum_k p(t_j| s_i, a_k)$$

where C_A is a normalization factor over all the alignments A. Finally, the probability $p(t_j| s_i)$ is determined from $p(t_j| s_i, A)$ using the Expectation Maximization algorithm, as described in [3].

Given a source query S, the probability of having t_j in its translation should be determined by the mutual contribution of the words in S, that is:

$$p(t_j| S) = C_S \sum_i p(t_j| s_i)$$

where C_S is another normalization parameter related to S’s length.

More specifically, our construction of the model from a corpus of parallel texts follows the following steps:

- Texts in the parallel corpus are submitted to a word transformation process. The purpose of this process is to transform each word into its citation form (e.g. singular form for nouns, singular-masculine form for adjectives, and infinitive for verbs in French). This process is based on a statistical tagging [8].
- Texts in the corpus are then aligned into parallel sentences. This alignment may be 1-1, 1-n or n-1. The alignment algorithm used is a variant of the system described in [17].
- The probability $p(t|s)$ is estimated from the sentence alignment as described above.

Obviously, a translation model in which all alignments are considered equiprobable, like Model 1, can only be a very coarse model. The lexical translation probabilities $p(t|s)$ are independent from the positions of t and s. In other words, the model is completely blind to syntax. This means that it is much too weak to generate full-blown translations on its own. Notwithstanding its weaknesses, Model 1 does capture some non-trivial aspects of the translation relationship as we observe it across natural languages. For example, an ambiguous word like “drug” will reinforce each of its equivalents ("médicament" and "drogue") according to a translation probability estimated from the training corpus. However, if the training corpus contains many occurrences of the expression "drug traffic" translated as "trafic de drogue", the presence of the English word "traffic" will thereafter tend to reinforce the French word "drogue" (in this instance, more than the French word "médicament").

3. Mining the Web for parallel texts

A model such as that described above was built using as training material the Hansard corpus – a collection of English-French parallel texts made up of 8 years of Canadian Parliament debates.

This corpus contains approximately 50 million words in each language.

The question one may raise is to what degree a probabilistic model may be constructed without a clean training corpus such as the Hansard. To answer this question, we developed an approach to automatically extract parallel texts from the Web. Our goal is to examine whether it is possible to construct a reasonable parallel text corpus from the Web to replace Hansard.

The Internet is a new source of translation examples. In fact, many sites are bilingual, mostly English and another language. Automatically extracting good parallel texts from the Web is an interesting scientific challenge because:

1. There is a huge number of sites to explored;

2. Useful documents are mixed up with garbage;

3. High-quality translations are mixed up with poor translations.

Our investigation is limited to the English-French pair for the moment. The approach may be separated into the following steps:

- Selection of candidate sites
- Selection of candidate documents from candidate sites

3.1. Selection of candidate web sites

We noticed that a parallel document is usually linked to the version in another language, and the link's anchor text often indicates the language of the linked text. For example, from an English text, there is often a link with "en français", "French", or "French version" … as the anchor text. This link points to the French version of the text. On the opposite direction, we usually have links anchored by "in English", "version anglaise", and so on. This phenomenon is used as our selection criterion: if at a web site, there are documents containing links to a document in the same site with one of these anchors in both direction, then the site is a candidate site of parallel texts.

Because of the availability of big search engines which index a large number of documents in the Internet, we take advantage of them in our selection of candidate web sites. A query is sent respectively to AltaVista and Northern Light, asking them for English (or French) documents containing an anchor text indicating a French (English) version. These engines will return lists of documents from which a set of candidate web sites are extracted.

3.2. Selection of candidate documents

For a candidate site, it is possible to know all the documents accessible from the Internet. The question now is how to pair them up as parallel texts. Of course, one may compare each document with all the others in that site. This would be extremely time-consuming.

Our selection makes use of the following heuristic: Parallel texts usually have similar names. The difference between their names is often a segment indicating the language. For example, "file-fr.html" vs. "file-en.html", "f-file.html" vs. "e-file.html", and so on. Therefore, for all the documents retrieved from a candidate site, we compare names to determine the first list of candidate text pairs using this criterion. Some flexibility is allowed: for example, parallel texts may be stored in different language-specific directories, say "eng/" v.s. "fr/". At this stage, documents other than texts are also eliminated.

It is interesting, however very difficult, to evaluate the precision and recall ratios of this pairing. We only did a preliminary evaluation on a set of samples. From the first 60 candidate sites, we obtained about 4000 possible parallel document pairs. There are in total 8000 French documents on these sites. So the recall ratio is at least 50% (if we consider that every French document has an English translation in these sites). For precision, we examined 164 randomly selected document pairs. 162 of them are indeed parallel. This gives us a precision of over 95%. This result is very encouraging. It shows that the simple name criterion is indeed very effective, and that this naming principle is widely used by web sites.

To further improve the selection, we then used the HTML structure of the texts to confirm the parallelism of the texts. Note, however, that true translation texts do not always have identical but similar structures. So, small variations are allowed.

A second possible improvement is to try to align the candidate texts. If they may be aligned, their chance of being parallel is very high. However, this process would take much time. In our present implementation, it is not used. Instead, we use the text length as an additional criterion: parallel texts should have similar lengths. Our preliminary test using 1000 randomly selected candidate pairs showed that the results using the length only differ by 2% from those obtained using an alignment algorithm. So we consider this criterion as a good replacement for alignment.

The above approach selected 14198 parallel document URLs after 75 hours, which correspond to 135 Mbytes French texts and 118Mbytes English texts. The process was stopped manually, after exploring about 30% of 5474 candidate sites selected at the first step. This result shows that the Internet contains a great number of parallel texts, and it is possible to automatically gather them using simple heuristics. The question that remained is about the quality of such a corpus. We will examine this in our experiments.

This work is similar to that of [15] and is carried out in parallel. The criteria we used seem to be more effective.

4. Experiments

The goal of the experiments is first to compare the effectiveness of the CLIR approach using a probabilistic model, then to examine how good the automatically constructed parallel corpus is for CLIR.

4.1. System description

For IR in a single language, we used a modified version of the SMART system [4] with *mtc* weighting scheme for both documents and queries, that is:

$$w_{t_i} = \frac{f(t_i, d)}{\max_j f(t_j, d)} * \log (N/n)$$

where $f(t_i, d)$ is the frequency of t_i in d, N is the total number of documents in the collection, and n is the number of documents including t_i.

Given a document vector and a query vector, their similarity is estimated as follows:

$$\mathrm{sim}(d, q) = \frac{\sum_{i=1,n} (w_{d_i} * w_{q_i})}{\sum_{i=1,n} (w_{d_i}^2) * \sum_{i=1,n} (w_{q_i}^2)]^{1/2}}$$

4.2. Data

Our experiments have been conducted on the two CLIR corpora used in TREC6 and TREC7 ([12]) – English AP and French SDA document collections with 25 queries in TREC6 and 28 in TREC7 written in both French and English.

In SDA, there are 141,656 documents, and in AP, 242,918 documents. We conducted the following experiments:

1. Monolingual French and English IR on SDA and AP respectively.

This is not CLIR, but is used as a reference point with which CLIR performance is compared.

In the other experiments, the English queries are translated into French and French queries are translated into English using various tools:

2. Using an MT system (Systran);
3. Using a bilingual dictionary only;
4. Using the probabilistic translation model estimated with the Hansard corpus;
5. Using the probabilistic translation model estimated with the Web corpus;
6. Using a combination of a probabilistic model and a bilingual dictionary.

4.3. Results

We used query titles and descriptions in all our tests. System performance is assessed by the average precision over 11 points of recall. The first 1000 documents retrieved by the system are used for performance evaluation.

First, let us show the number of evaluated queries in each collection:

	AP	SDA
Trec6	21	21
Trec7	26	28

Table 1. Number of evaluated queries

Notice that although 21 Trec6 queries are evaluated in both AP and SDA, they are not the same. This difference may partly explain the difference in performances described below.

4.3.1. Monolingual IR

The following table shows the performances obtained in each case of monolingual retrieval.

	E-E	F-F
Trec6	0.2895	0.3686
Trec7	0.3202	0.2764

Table 2. Average precision for monolingual IR

We observe a clear difference between E-E and F-F runs in Trec6. This difference may be explained by the difference between the document collections. It is also partly due to the difference of the query sets used in the two cases. After an examination of the queries, we found that several English and French queries are different in difficulty and in concept coverage. In English Query 3 (on drug traffic), the confusion verb "stem" is used, and in Query 4 (on reusage of garbage), the unusual word "reusage" is used. In Query 7 (on sex education) the concept of school is not mentioned while it is in the French version.

In the case of Trec7, the queries seem to be more similar in English and in French, but still, the evaluated queries are different.

4.3.2. CLIR using MT (Systran)

We used one of the best MT systems available in the Internet – Systran [18] to do query translation. The following table shows the average precisions obtained for each case (where F-E means translating French queries into English:

	F-E (%mono)	E-F (%mono)
Trec6	0.3098 (107.0%)	0.2727 (74.0%)
Trec7	0.3293 (102.8%)	0.2327 (84.2%)

Table 3. Average precision using MT

These performances represent well what we can achieve now with MT systems. For F-E runs, the results are surprisingly good. They are even slightly higher than the monolingual runs (the percentages show the comparisons with the respective monolingual runs). For E-F runs, on the other hand, we notice important degradations in both Trec6 and Trec7. The even higher performance in F-E runs may reflect the good quality of Systran's French to English translation; but it is also partly due to the higher quality of some French queries of Trec6, as we mentioned above.

Below are some query translations obtained with Systran:

Original English queries (the description field):

```
1: Reasons for controversy surrounding
   Waldheim's World War II actions.
2: Are marriages increasing worldwide?
3: What measures are being taken to stem
   international drug traffic?
```

French translations:

```
1: Raisons pour la polémique entourant des
   actions de la deuxième guerre mondiale de
   Waldheim.
2: Sont des mariages augmentant dans le
   monde entier ?
3: Quelles mesures sont prises au trafic de
   stupéfiants international de tige?
```

English translations from the original French queries:

```
1: Reasons of the controversy with regard to
   the intrigues of Waldheim during the
   Second World War.
2: Does the rate of the marriages increase
   in the world?
3: Which are measurements taken to control
   the smuggling of narcotics?
```

As we can notice from these examples, the English translations are reasonably good. However, sentences in French translations are often incorrectly structured. In the case of Query 3, due to the use of the word "stem", the French translation include a noise word "tige" (meaning "tree stem").

If we only consider the translations at the word level, we observe that in most cases, the word choice is good or acceptable. This may explain the good performances obtained using MT.

Because MT systems choose a unique equivalent for each source language term, the resulting query sometimes misses documents containing different but related words. For example, the English translation of Query 3 chooses to use the alternative expression "smuggling of narcotics" instead of "drug traffic". Documents retrieved using this translation will be different from those using the original English Query 3.

4.3.3. CLIR using bilingual dictionaries

We obtained from the Web a small bilingual dictionary which contains about 7900 citation forms in English and in French. This dictionary is used for query translation. For an English word in a query, we use all the corresponding French words stored in the dictionary as its translations. For example, the words “drug” and “increase” will be translated by

drug:	remède, médicament, drogue, stupéfiant.
increase:	accroître, agrandir, amplifier, augmenter, étendre, accroissement, grossir, s’accroître, redoubler, accroissement.

The performances obtained with these translations are shown in the following table:

	F-E (% mono)	E-F (%mono)
Trec6	0.1276 (44.1%)	0.1740 (47.2%)
Trec7	0.1048 (36.2%)	0.0785 (28.4%)

Table 4. Average precision using a small bilingual dictionary

As we can see, the performances are merely about 40% of the monolingual performances. There may be several explanations to this result:

1. Word translations are ambiguous in the dictionary. In fact, all the senses of a word are mixed up in its translations. Although we can find some synonyms in the translations, there is a quite amount of noise words that are not related to the query.
2. Common words (e.g. increase) tend to have much more translations than specific words. As a result, the translation of a query will be flooded by these common words.

We also obtained the “Banque de Terminologie du Québec” (Terminology database of Quebec – BTQ) from the "Office de la Langue Française" of the Quebec government, and several other bilingual dictionaries from the Internet. They were combined into a big dictionary of over 1 million entries (most of them are compounds). The IR effectiveness using this new dictionary is shown in the following table:

	F-E (%mono)	E-F (mono)
Trec6	0.1707 (59.0%)	0.2305 (62.5%)
Trec7	0.1701 (53.1%)	0.1352 (48.9%)

Table 5. Average precision using a big bilingual dictionary

Although we observe some improvement from the small dictionary, the performances are still lower than the 2/3 of the monolingual performances that we can usually obtain [11].

4.3.4. CLIR using a probabilistic translation model estimated from Hansard

In these tests, we use the probabilistic model trained with the Hansard corpus (called the Hansard model). The translation is performed as follows. An English query E is submitted to the probabilistic model as a single sentence so as to calculate $p(f|E)$, the probability that word f will occur in any translation of E. Since f ranges over a very large vocabulary (all the French words observed in our training corpus), we want to retain only the best scoring words. This is because:

1) The longer the word list, the longer the time for the retrieval process. So a restriction in length leads to an increase in retrieval speed.

2) As the translation model is not perfect, the list is sometimes noisy. This is especially true when the source language query contains words whose frequency was low in our training corpus. In this case, probability estimations are notoriously unreliable. By limiting the resulting list to an appropriate length, the amount of noise may be reduced.

Thus, our “translation” of a query will be simply made up of the n words t for which $p(t|S)$ is highest. We will experiment with several values of n in order to assess how this parameter affects IR effectiveness.

The following lists show some of the first words in the translations of 2 queries (see section 4.3.2 for original queries):

Translation of Query 1

English to French	French to English
affaire=0.0700	war=0.0840
waldheim=0.0674	waldheim=0.0789
guerre=0.0621	world=0.0665
raison=0.0483	reason=0.0558
ii=0.0479	controversy=0.0442
monde=0.0436	affair=0.0427
controverse=0.0385	action=0.0243
entourer=0.0368	business=0.0181
mesure=0.0230	global=0.0137
mondial=0.0192	controversial=0.0111
prendre=0.0184	what=0.0110
second=0.0159	matter=0.0091
suite=0.0131	serve=0.0089
action=0.0110	activity=0.0078
susciter=0.0069	president=0.0070
donner=0.0066	deal=0.0064
pouvoir=0.0062	case=0.0062
cause=0.0055	credential=0.0062

Translation of Query 3

English to French	French to English
médicament=0.1109	control=0.1114
mesure=0.0911	what=0.0940
international=0.0865	drug=0.0833
trafic=0.0524	smuggle=0.0670
drogue=0.0414	narcotic=0.0422
découler=0.0242	measure=0.0399
circulation=0.0196	action=0.0198
pharmaceutique=0.0187	make=0.0196
pouvoir=0.0135	legislation=0.0126
prendre=0.0126	stagger=0.0115
extérieur=0.0117	amazing=0.0093
passer=0.0078	step=0.0084
demander=0.0074	illicit=0.0079
endiguer=0.0067	bill=0.0072
nouveau=0.0060	astound=0.0063
stupéfiant=0.0053	monitor=0.0054

Some interesting facts may be observed from query translations:

1) The word translations obtained reflect the peculiarities of our training corpus. For example, the word "drug" is translated by, among others, "médicament" and "drogue", and a higher probability is attributed to "médicament". This is because in the Hansard corpus, the English “drug” refers more often to the sense “médicament” than to “drogue”.

2) This dependence on the training corpus sometimes leads to odd translations. For example, the word “bille” is considered as a French translation of “logging” in the English query “effects of logging on desertification”. This translation comes from the fact that in the Hansard corpus "log" in English is often translated as "bille de bois" in French.

3) Some words are rare or even absent in our training corpus, and this leads to unreliable translations. For example, there was only one occurrence of “acupuncture” in the training corpus. Because of that, the model fails to assign a higher probability to the French “acuponcture” than to other semantically unrelated words that appeared in the same sentence.

4) The model sometimes fails to distinguish the real translation from noise induced by simple statistical associations. For example, the word “pouvoir” appears in the translations of queries 1 and 3 with a quite high probability, and “donner” in Query 1.

Despite these problems, we observe that real translations and associated words tend to score relatively high and appear at the top of the list.

The obtained probabilities are further combined with the *idf* factor in the final query vectors. It has been shown [14] that the combination improves the retrieval effectiveness.

The following table shows the performances obtained using this probabilistic model (where the length of the translation word lists changes from 25 to 100).

	Length	F-E (%mono)	E-F (%mono)
Trec6	25	0.2166 (74.8%)	0.2501 (67.9%)
	50	0.2058 (71.1%)	0.2514 (68.2%)
	75	0.2063 (71.3%)	0.2347 (63.7%)
	100	0.1983 (68.5%)	0.2350 (63.8%)
Trec7	25	0.3124 (97.6%)	0.2587 (93.6%)
	50	0.3401 (106.2%)	0.2030 (73.4%)
	75	0.2699 (84.3%)	0.2037 (73.7%)
	100	0.2589 (80.9%)	0.2030 (73.4%)

Table 6. Average precision using Hansard model

First, observe how length affects the results. In Trec6 F-E run and Trec7 E-F run, a short translation word list (25) seems to be appropriate, while in the two other runs, better results are obtained with 50 words.

In comparison with monolingual runs, the best performances in the four cases vary from 68% to 106% of the corresponding monolingual runs. As for MT, the Trec7 F-E run outperformed the monolingual run in the best case.

In comparison with MT, in the Trec7 F-E, Trec6 E-F and F-E cases, the performances are comparable. In the Trec6 F-E case, however, MT approach performed much better. This result may be explained by the dependency of the probabilistic model on the training data. As our training data are rather particular, they may, or may not fit the document collection to which we will apply the model. This led to several other translation problems of the probabilistic model:

1. Translation by statistically related wrong words

This is the most important problem we observed in query translation. Many concepts are translated by these wrong words. The probabilistic model is unable to distinguish a statistical association from semantic association. For example, in the “reusage of garbage” query, the French word “recyclage” has been first translated as “retraining” with a probability of 0.159, whereas the correct translation “recycling” only received a probability 0.026. The word “British” is often translated by “Colombie britanique” (British Colombia), and the concept “west” is often translated as “Ottawa-ouest” and “Calgary-centre-ouest”, because the latter appeared very often in the parallel sentences which contain “west”.

2. Translating a compound term word by word

Probabilities are estimated on a word to word basis. This makes it difficult to translate compounds. For example, the French compound “pomme de terre” (potato) is first translated by “land”, then “potato”. A few other wrong words are also included at the top of the translation list, such as “apple” and “earth”. This problem seems difficult to solve in the current model. A possible solution lies in a correct identification of compounds in texts. These compounds can then be considered as inseparable entities during model construction. This is one of our future research projects.

3. Unknown words

Unknown words are included in the translation list with a fixed “probability” value (0.05 in our tests). This setting may correctly deal with proper nouns such as “Banco Ambriosiano”, “Ustica”. However, if a common word is unknown, the concept may not often be recovered in this way. For example, in the query on “child abuse”, the French word for “abuse” – “maltraitance” is an unusual word, and is unknown by Hansard. The French word is added directly in the resulting word list, leading to no interesting documents. The solution to this problem is to increase the size of the training corpus. Unfortunately, this solution does not seem to be achievable at present time for the Hansard.

It is interesting to observe that synonyms and related words have been included in many query translations. For example, the French word “parfum” is translated to both “perfume” and “fragrance”. For the query on “organic farming” and “organic cotton”, both “organique” and “biologique” have been included in the French translation at top level. This produces a natural query expansion effect.

Globally, we can conclude that the probabilistic model performed reasonably well. Its performance is close to the best MT systems.

4.3.5. Using the probabilistic model estimated from the web documents

The corpus of parallel texts obtained from the Web is big. It takes several days to train a probabilistic model. Unfortunately, we now only have a model trained with the first 5000 documents of the Web corpus. The results described below are obtained with this limited model. We will call the current model the Web model. The following table show the average precision obtained in the different cases.

	Length	F-E (%mono)	E-F (%mono)
Trec6	25	0.2103 (72.6%)	0.2595 (70.4%)
	50	0.2103 (72.6%)	0.2595 (70.4%)
	75	0.2102 (72.6%)	0.2600 (70.5%)
	100	0.2108 (72.8%)	0.2640 (71.6%)
Trec7	25	0.2380 (74.3%)	0.1975 (71.5%)
	50	0.2379 (74.3%)	0.1972 (71.3%)
	75	0.2382 (74.4%)	0.1974 (71.4%)
	100	0.2382 (74.4%)	0.1977 (71.5%)

Table 7. Average precision using Web model

We first observe that the length factor has almost no impact on the performance. This is surprising. We are still analyzing the causes of this phenomenon.

In comparison with the Hansard model, we observe similar performances in Trec6 runs, and much lower performances for the Web model in the Trec7 runs. After analyzing the resulting word lists, we found that the difference was mainly created by the great number of country and region names (e.g. Germany, France, Switzerland, Sudan, and so on) in Trec7 queries. In fact, 18 queries out of 28 in Trec7 contain a country or region name, versus 5 among 25 in Trec6. The translation of country and region names is particularly problematic for the Web model. For example, the name "Sudan" is not only translated by "Soudan", but also by "Singapour" and "Royaumes unis" with quite high probabilities. For the query on "Swiss Confederation's public debt", the countries "United Kingdoms", "Canada", "Uzbekistan", "Ukraine" and "Turkey" are also included among the 25 first translation words. A possible explanation of this phenomenon is that a number of documents in this training corpus contain lists or descriptions of different countries in a single sentence. Thus, different countries may not be separated by sentence alignment algorithm. As a result, any country in such a list is a possible translation of any country in the aligned sentence. This kind of text does not appear often in the Hansard corpus.

Apart from this particular problem, we also observed that French functional words often appear in English translations of French queries. This is because of the noisy parallel documents gathered and poor translation quality of some documents in the Web. In fact, in a number of cases, the documents in French are not translated, but they are labeled as "English version". It is more common that a title of document (for example a French novel) is left non-translated in the English version and vice versa. From such "parallel" texts, the French functional words may appear in the English translations of queries in the same language. As they are not considered as functional words in English, their appearance is harmful. In order to avoid this problem, we can apply a language recognition mechanism to filter out non-translated documents or sentences. A language identification system like SILC [13] is the most appropriate. We plan to use this mechanism to further filter the parallel Web corpus in the near future.

Once these particular problems have been solved, we believe the Web model will be able to reach at a similar level of performance to the Hansard model. This is very encouraging: not only does it prove that we can rival a well controlled parallel corpus like the Hansard with an automatically constructed training corpus, but also, it offers the possibility for CLIR between other language pairs for which there is no Hansard-like parallel corpora. In fact, the same approach may be applied for English-Spanish, English-German, English-Chinese, English-Japanese, and so on. These languages are all active on the Internet and many sites contain bilingual documents.

4.3.6. Combining the probabilistic translation model with a bilingual dictionary

We noticed the problem that a probabilistic translation model is unable to distinguish true translation words from statistically associated words. One way to distinguish them is to use a bilingual dictionary to increase the probability of the translation words that are stored in the dictionary.

A problem arises in such a combination due to the different nature of each element: one is weighted and the other is not. In other words, the question is the following: if a French word is a translation of an English word in the bilingual dictionary, how much should we increase the weight (probability) of this translation in the probabilistic model? Our goal was not to provide a theoretically well-founded answer to that question but simply to see if a simple-minded solution would prove useful in practice. We tested the following approach: when a translation is stored in the bilingual dictionary, its probability is increased by a *default value*. We tested several default values, ranging from 0.005 to 0.03. The following tables report the IR effectiveness obtained once the Hansard model with length=25 is combined with the small and big dictionaries respectively.

	Default *p*	F-E (%mono)	E-F (%mono)
Trec6	0.005	0.2233 (77.1%)	0.2671 (72.4%)
	0.01	0.2388 (78.8%)	0.2754 (74.7%)
	0.02	0.2337 (80.7%)	0.2816 (76.4%)
	0.03	0.2322 (80.2%)	0.2784 (75.5%)
Trec7	0.005	0.3135 (97.9%)	0.2650 (95.9%)
	0.01	0.3148 (98.3%)	0.2665 (96.4%)
	0.02	0.3123 (97.5%)	0.2619 (94.8%)
	0.03	0.2995 (93.5%)	0.2480 (89.7%)

Table 8. Combining Hansard model with the small dictionary

	Default *p*	F-E (%mono)	E-F (%mono)
Trec6	0.005	0.2312 (79.9%)	0.2794 (75.8%)
	0.01	0.2425 (83.8%)	0.2908 (78.9%)
	0.02	0.2526 (87.3%)	0.3037 (82.4%)
	0.03	0.2560 (88.4%)	0.3053 (82.8%)
Trec7	0.005	0.3245 (101.3%)	0.2649 (95.8%)
	0.01	0.3244 (101.3%)	0.2628 (95.1%)
	0.02	0.2810 (87.8%)	0.2580 (93.3%)
	0.03	0.2708 (84.6%)	0.2443 (88.4%)

Table 9. Combining Hansard model with the big dictionary

The results clearly show the advantages of such a combination, even in the case of a very small bilingual dictionary. In fact, in all the cases, the effectiveness obtained after combination is generally higher than the probabilistic model alone.

The quality of the dictionary also has a significant impact. Using the big dictionary, more (and better) translations have been added than in the case of the small dictionary. As a consequence, the IR effectiveness is increased.

In some cases (Trec6), the default "probability" value may be set at a quite high level. In Trec7, when this value increases, the effectiveness decreases. So we cannot observe a general rule on the setting of the default probability value. It is strongly query- and corpus-dependent.

Compared with the performances using MT (Table 3), we can see that the combination with the big dictionary performed better in 2 cases, worse in 1 case, and equivalently well in 1 case. Note further that the cases of length 25 in Hansard model are not always the best. In Trec7 F-E cases, in particular, if we combine the length 50 cases of Hansard model with the dictionaries, we obtained average precision of 0.3422 and 0.3545 respectively. These values are higher than those obtained with MT (table 3 Trec7 F-E cases).

In conclusion, globally, the Hansard model, together with a bilingual dictionary, gives comparable performances to (or slightly better than) those obtained with MT.

In the case of Web model, we also observed a general increase in performance once a bilingual dictionary is added. The following table shows the results for the combination of the Web model - length 25 with the big bilingual dictionary.

	Default p	F-E (%mono)	E-F (%mono)
Trec6	0.005	0.2297 (79.3%)	0.2702 (73.3%)
	0.01	0.2425 (83.8%)	0.2789 (75.6%)
	0.02	0.2528 (87.3%)	0.2983 (80.9%)
	0.03	0.2590 (89.5%)	0.3041 (82.5%)
Trec7	0.005	0.2598 (81.1%)	0.2230 (80.7%)
	0.01	0.2610 (81.5%)	0.2296 (83.1%)
	0.02	0.2483 (77.5%)	0.2290 (82.8%)
	0.03	0.2447 (76.4%)	0.2220 (80.3%)

Table 10. Combining Web model and the big dictionary

These performances are comparable to MT in the two E-F cases, and worse in the two F-E cases. However, the global performances are still reasonably good. They are slightly higher than 80% of the monolingual performances. We expect that the model trained with all the parallel documents from the Web will perform better. Already, the current results indicate that an automatically constructed parallel corpus may be a reasonable resource for CLIR.

5. Conclusions

A good MT system, if available, may perform query translation of reasonable quality for CLIR purposes. However, MT systems are available for only a few pairs of languages. It is difficult to construct more good MT systems to cover other languages.

In this paper, we investigated the possibility of replacing MT with a probabilistic model for CLIR. In comparison with MT, this approach is more flexible. It may be used for any pair of languages for which an appropriate parallel corpus is available. The results we obtained using such a model are globally comparable to those obtained with an MT system.

One often mentioned the unavailability of parallel texts as a major obstacle to a probabilistic approach to MT. For MT purposes, the training corpus should be tightly controlled; otherwise, wrong or poor-quality translations will be produced. For CLIR, the requirements are much less: It only requires the model to provide a list of the most probable translation words without taking into account syntactic aspects. For this, a parallel corpus of lower quality still can provide reasonably good query translations. Based on this hypothesis, we investigated the automatic gathering of parallel texts in French and English from the Web. The number of parallel texts obtained is surprisingly high. We used a part of the parallel texts to train a small model, and used the model for CLIR. The results we have obtained already showed clearly the feasibility of using Web parallel documents for model training. We can now envision to apply the same technique to other pairs of languages for which there is no readily available sizable parallel corpora, for example, Chinese-English, Italian-English, Japanese-English, and so on.

There are several possible improvements on the approach presented in this paper. 1) The estimation of probabilistic model may be improved with regard to very common words and compounds. 2) For the Web corpus, a language identification system [13] may be added in order to filter out the documents or parts of documents that are not translated. 3) Finally, there are still rooms to improve the utilization of a probabilistic model for CLIR. Many questions need to be answered. For example, is it possible to determine the appropriate number of translation words automatically for a query? Is this number related to the length of the original query? Is it possible to combine two probabilistic models to build a transitive model (i.e. from a model on A-B and another on B-C to construct a model for A-C)? These are some of the questions we will address in our future research.

References

[1] L. Ballasteros, W.B. Croft, Resolving ambiguity for cross-language retrieval, *ACM-SIGIR*, pp.64-71, 1998.

[2] R.D Brown, Automatically-extracted thesauri for cross-language IR: When better is worse, *1st Workshop on Computational Terminology (Computerm)*, pp.15-21, 1998.

[3] P. F. Brown, S.A.D. Pietra, V. D. J. Pietra, and R. L. Mercer, The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19: 263-312, 1992.

[4] C. Buckley, Implementation of the SMART information retrieval system. Cornell University, Tech. report 85-686, 1985.

[5] M. Davis, T. Dunning, Query translation using evolutionary programming for multilingual information retrieval, *Proc. Of the 4th Annual Conf. on Evolutionary Programming,* 1995.

[6] M.W. Davis, W.C. Ogden, QUILT, Implementing a large-scale cross-language text retrieval system, *ACM-SIGIR,* pp.92-98, 1997.

[7] S.T. Dumais, T.K. Landauer, M.L. Littman, Automatic cross-linguistic information retrieval using latent semantic indexing, *SIGIR'96 Workshop on Cross-Linguistic Information Retrieval,* 1996.

[8] George F. Foster, *Statistical Lexical Disambiguation*, M.Sc thesis, McGill University, School of Computer Science, 1991.

[9] G. Foster, P. Isabelle, and P. Plamondon, Target-text Mediated Interactive Machine Translation. *Machine Translation*, 12: 175-194, 1997.

[10] W. A. Gale, K.W. Church, A program for aligning sentences in bilingual corpora, *Computational Linguistics*, 19 (1): 75-102, 1993.

[11] G. Grefenstette (ed.), *Cross-Language Information Retrieval*. Kluwer Academic Publisher, 1998.

[12] D. K. Harman and E. M. Voorhees (eds.), *Text REtrieval Conference (TREC-6).* Gaithersburg, 1997.

[13] P. Isabelle, G. Foster et P. Plamondon, *SILC : un système d'identification de la langue et du codage*, http://www-rali.iro.umontreal.ca/ProjetSILC.en.html, 1997.

[14] J.Y. Nie, P. Isabelle, P. Plamondon, G. Foster, Using a probabilistic translation model for cross-language information retrieval, *6th Workshop on Very Large Corpora*, Montreal, pp.18-27, 1998.

[15] P. Resnik, Parallel Stands: A preliminary investigation into mining the Web for bilingual text, *AMTA'98,* 1998.

[16] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*: McGraw-Hill, 1983.

[17] M. Simard, G. Foster, P. Isabelle, Using Cognates to Align Sentences in Parallel Corpora, *Proc. of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, 1992.

[18] Systran-reference, http://babelfish.altavista.digital.com/

[19] Y. Yang, J.G. Carbonell, R.D. Brown, R.E. Frederking, Translingual information retrieval: learning from bilingual corpora, *Artificial Intelligence*, 103: 323-345, 1998.