Using a Probabilistic Translation Model for Cross-Language Information Retrieval

Jian-Yun Nie, Pierre Isabelle, Pierre Plamondon, George Foster
Laboratoire RALI,
Département d'Informatique et Recherche opérationnelle,
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec, H3C 3J7 Canada
{nie, isabelle, plamondo, foster}@iro.umontreal.ca

There is an increasing need for document search mechanisms capable of matching a natural language query with documents written in a different language. Recently, we conducted several experiments aimed at comparing various methods of incorporating a cross-linguistic capability to existing information retrieval (IR) systems. Our results indicate that translating queries with off-the-shelf machine translation systems can result in relatively good performance. But the results also indicate that other methods can perform even better. More specifically, we tested a probabilistic translation model of the kind proposed by Brown & al. [2]. The parameters of that system had been estimated automatically on a different, unrelated, corpus of parallel texts. After we augmented it with a small bilingual dictionary, this probabilistic translation model outperformed machine translation systems on our cross-language IR task.

1. Introduction

Adequate text processing systems have become widely available for most natural languages. While English remains the dominant language on the Internet, the relative share of other languages now appears to be on the rise. The network has become truly multilingual. This situation has created an acute need for tools capable of performing language-sensitive search in multilingual databases. In particular, there is a need for tools capable of performing cross-language information retrieval (CLIR), that is, of matching an information query written in one particular language with documents that may be written in one or several different languages.

Given such a need, the solution that immediately comes to mind is to translate the information query using a machine translation (MT) system, and to feed the resulting translation into a classical monolingual IR system.

However, it should be stressed that MT and IR have widely divergent concerns. First, observe that MT systems are expected to produce syntactically correct translations and that they tend to spend a lot of effort trying to attain that rather elusive goal. On the other hand, current IR systems tend not to care about grammar : for them texts are mostly viewed as vectors of content words. Second, note that 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. Here again what the MT system is expected to do turns out to be unnecessary and maybe undesirable from an IR point of view. As a case in point, classical IR systems often perform a *query expansion* process by which certain query terms/words are mapped onto several equivalent or related index terms. Not surprisingly, such a process could well make provision for mapping the query word "organique" onto the two index terms "organique" and "biologique" so as to account for (partial) synonymy between these words. In other words, MT systems attempt to systematically eradicate translational ambiguity instead of taking advantage of it to capture synonymy relations.

At the opposite end of the spectrum, MT is replaced with a simple bilingual dictionary lookup. To that end, one can use either an ordinary general-purpose dictionary, a technical terminology database, or both. Because of the fact that in any sizable dictionary most words receive many translations, the dictionary approach will in effect subject the query to a rather massive expansion process. The resulting target language query is likely engender a lot of noise (irrelevant documents that get retrieved), mostly due to the fact that in each dictionary entry some of the translations can correspond to different meanings of the source language word. For example, the English word "drug" is translated in French as "drogue" (an illegal substance) or as "médicament" (a legal medicine) depending on the context. There is most often no explicit clue in the query that would allow one to choose the appropriate meaning.

Yet another 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. There are now well-known methods for aligning the sentences of parallel corpora (Gale & Church [6], Simard, Foster & Isabelle [10]). Then, the translational equivalence of a given pair of words can be estimated by their degree of co-occurrence in parallel sentences. 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 the next section, we describe the structure of a probabilistic translation model that can calculate $p(f|\mathbf{e})$, the probability of observing word *f* as part of the translation of sentence **e**. Given a query **e**, we can then select the *n* best-scoring values of *f* as the set of index terms in the target language. This method will be compared to the other two mentioned above.

2. A Probabilistic Translation Model

Any source language input **e** can usually be translated in a great many different ways. Machine translation systems are expected to select but one particular translation **f** for each input. In the current state of the art, unaided MT is generally unable to produce high-quality translations: human translators remain mostly unchallenged. Moreover, it has been shown repeatedly that human translators seldom find it practical to post-edit MT output: the machine has just made too many wrong or questionable decisions.

If the goal is to help human translators, it is advisable to stop short of producing a full-blown automatic translation. There is no point in having the machine spontaneously propose a detailed target language syntactic structure unless there is at least a reasonably good chance that the translator will want to use it. Similarly, there is no point in having the machine select target language equivalents for all source language words unless most of these equivalents are likely to be retained by the translator.

In recent years it has been shown that existing MT techniques can produce useful results when they are applied to tasks that amount to somewhat less than translation proper. In previous work, we have shown that probabilistic translation models such as those of Brown et al. [2] could be used as the key component of various translation support tools. Specifically, our work on the TransTalk project [1, 4] has established that such models could become instrumental in improving the process of automatically transcribing a spoken translation. And our ongoing work on the TransType project [5] indicates that models of the same kind can drive typing aids for translators.

A key feature of such applications is that they do not expect the machine to volunteer a full-fledged translation on its own. Rather, the machine is only expected to restrict the range of possible translations so as to make it easier to guess what the intentions of the human translator are.

For example, in certain incarnations of our TransTalk system, the translation model is used as a means of answering the following question: given a source language sentence **e**, what is the likelihood of observing the word *f* in any target language sentence **f** that constitutes a valid translation of **e**? If **e** is an English sentence that contains the word ôhorsesö, then the likelihood that ô chevauxö (the most direct equivalent for ôhorsesö) will appear in a French translation **f** of **e** is much greater than the a priori likelihood of observing ô chevauxö in a random French sentence. In contrast, there is no reason to expect that the likelihood of observing ôcheveuxö (an acoustically close word that means ôhairö) in **f** will be significantly altered:

$$p(chevaux \in \mathbf{f} \mid horses \in \mathbf{e}) > p(chevaux \in \mathbf{f})$$
$$p(cheveux \in \mathbf{f} \mid horses \in \mathbf{e}) \approx p(cheveux \in \mathbf{f})$$

TransTalk makes use of this fact to help resolve the acoustic ambiguity between the French words ôchevauxö and ôcheveuxö.

From the point of view of translation support, doing somewhat less than full-blown MT is likely to achieve more.

In this paper, we want to argue that CLIR is facing a similar situation in that subjecting the source language query to a process that stops short of producing a full-blown target language query can result in a better retrieval performance. The prima facie case for such a hypothesis appears reasonable for (at least) the following reasons:

- Given a source language query **e**, there does not appear to be any necessity to produce a target language query **f** that is fully specified with respect to ôsuperficialö grammatical and lexical features of the target language: IR is not interested in superficial form but in meaning.

- As a case in point, there is no reason to spell out in detail the target language syntactic structures and word order in **f** because IR systems tend to disregard such features anyway.
- As another case in point, whenever the target language allows for several equivalent terms (ôsynonymsö) there does not seem to be any reason to force one particular choice in **f**. In fact, many monolingual IR systems subject the query to a process known as ôquery expansionö which amounts to adding synonyms and other related words to the query.

As we have seen above, probabilistic translation models of the kind used in our TransTalk system can stop short of fully specifying a target language translation **f**, producing instead a representation that is in a sense compatible with many alternative translations.

By translation model, we mean a mechanism which associates to each source language sentence (or query) **e** a probability distribution $p(\mathbf{f}|\mathbf{e})$ on the sentences (or queries) **f** of the target language. A precise description of a family of such models can be found in Brown & al. [2]. The model we will be using for the experiments reported here is basically their ôModel 1ö. In this model, a source **e** and its translation **f** are connected through an alignment **a**, that is a mapping of the words of **e** onto those of **f**. If $\mathbf{e} = e_1, e_2$, à, e_l and $\mathbf{f} = f_1, f_2$, à, f_m then $\mathbf{a}_j$ will be used to refer to the particular position in **e** th at is connected with position j in **f** (for example, $\mathbf{a}_2 = 4$ expresses the fact that f_2 is connected with e_4) and e_{aj} will be used to refer to the word in **e** at position $\mathbf{a}_j$.

The probability $p(\mathbf{f}|\mathbf{e})$ is decomposed as a sum over all possible alignments:

$$p(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}\in \mathbf{A}} p(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

The conditional probability of **f** under alignment **a** given **e** can be analysed as follows:

$$p(\mathbf{f},\mathbf{a}|\mathbf{e}) = p(\mathbf{f}|\mathbf{a},\mathbf{e})\, p(\mathbf{a}|\mathbf{e}) = K_{\mathbf{e},\mathbf{f}}\, p(\mathbf{f}|\mathbf{a},\mathbf{e})$$

The latter equality stems from the fact that in model 1, all alignments are considered equiprobable (see below). Consequently $p(\mathbf{a}|\mathbf{e})$ is a constant $K_{\mathbf{e},\mathbf{f}}$ equal to 1 over the total number of alignements.

The core of the model is $t(f|e)$, the lexical probability that some word e is translated as word f. The value of $p(\mathbf{f}|\mathbf{a,e})$ depends mostly on the product of the lexical probabilities of each word pair connected by the alignment:

$$p(\mathbf{f}|\mathbf{a},\mathbf{e}) = C_{\mathrm{f,e}} \prod_{j=1,m} t(f_j|e_{aj})$$

where $C_{\mathbf{f,e}}$ is a constant that accounts for certain dependencies between the respective lengths of sentences **e** and **f** (mostly irrelevant here).

The probability of observing word f_j in **f** under a particular alignment **a** is:

$$p(f_j|\mathbf{a},\mathbf{e}) = t(f_j|e_{aj})$$

And the probability of observing word f_j in **f** under any alignment is:

$$p(f_j|\mathbf{e}) = \sum_{i=1,l} t(f_j|e_i)$$

Since all alignments are considered equiprobable, we can simply sum up the values obtained by connecting f_j to each word e_1, e_2, à, e_l of **e**. In other words, the probability of observing a particular word in a given position in **f** is established as the total of the lexical contributions of each word of **e**.

The parameters of our translation model are estimated from a bilingual parallel corpus in which each sentence has been aligned with the corresponding sentence(s) of the other language. Such alignments can be produced using algorithms such as the one described in [10]. Given such alignments we can estimate reasonable values for the parameters $t(f|e)$ using the Expectation Maximization algorithm, as described in [2]. The model used in the experiments reported here has been trained using 8 years of the Canadian Hansard (parliamentary debates), that is, approximately 50 million words in each language.

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 $t(f|e)$ are independent from the positions of f and e. As a result, for any j, jÆ the model assigns the same value to $p(f_j|\mathbf{e})$ and to $p(f_{j\text{Æ}}|\mathbf{e})$. 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. At the very least, one would need to use it in tandem with a language model $p(\mathbf{f})$ capable of capturing some constraints on acceptable sequences of words in the target language.

Notwithstanding its weaknesses Model 1 does capture some non trivial aspects of the translation relationship as we observe it across natural languages. For example, it is indeed a property of that model that a relatively unmabiguous source language word (say, the English "chimney") will reinforce its equivalents in a stronger way than a very ambiguous word. An anmbiguous word like ôdrugö will reinforce each of its equivalents ("médicament" and "drogue") according to a translation probability estimated from the training corpus. While the model only operates at the level of simple word (as opposed to *complex terms*), it should be observed that it nonetheless captures some non-trivial contextual effects. For example, 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").

And given the fact that the intended application is not MT but CLIR, the use of a ôweakö translation model turns out to be, in some respects, an advantage. As mentioned above, for CLIR it is not particularly useful to produce a full-blown translation of the source-language query. The IR model to which the translated query will be passed has no use for detailed syntactic descriptions: queries and documents are uniformly reperesented as vectors of weighted terms. Given any query **e**, our translation model will calculate a value for $p(f|\mathbf{e})$, the probability of observing word f in the translation of **e**. It turns out to be straightforward to reinterpret this probability distribution as a vector of weighted terms.

3. Cross-Language information retrieval

After a brief description of the principal functions of an IR system, we report our experiments on CLIR.

3.1. The tasks of an IR system

An IR system perform three main tasks[9]:

- document indexing
- query indexing
- matching the query and the documents

Document indexing creates an internal representation (for example, a vector) for each document. Before indexation can be accomplished, some pre-processing is often necessary. In our case, pre-processing involves the following steps :

- Morphological analysis : each word is transformed into a canonical, citation form. For example, nouns and (French) adjectives are transformed into their masculine singular form, andverbs are transformedinto their infinitive forms. This neutralization of irrelevant differences in form often reduces retrieval silence.
- Elimination of grammatical words: words that are more or less semantically empty are useless for IR. Such words are eliminated in order to reduce the size of the index and speed up the search process.

For the indexing process, each document is represented as a set or a vector of weighted terms (words in canonical form). Term weight is determined by the following two factors :

- *tf* (term frequency): the relative frequency of the term in the document; and
- *idf* (inverse document frequency): a measure of the nonuniformity of the distribution of term across documents of the collection.

The terms that rank best within a document d are those that are at the same time frequent within d and distributed unevenly in the collection of documents. The *tf*idf* weigthing schema combines these two criteria [3, 9] . To determine the weight w_{t_i} of term t_i in document d we used the following variant of *tf*idf*:

$$w_{t_i} = [\log(f(t_i, d)) + 1] * \log(N/n)$$

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.

The indexation process maps each document onto a vector of weights within the vector space of the indexes of the corpus. Each weight represents the relative importance of the term in the document. For example,

Vector space: $< t_1, t_2, ..., t_n >$

$d \rightarrow \quad < w_{d_1}, w_{d_2}, ..., w_{d_n} >$

where w_{d_i} is the weight of t_i in document d.

The indexing process for queries is the same:

$q \rightarrow \quad < w_{q_1}, w_{q_2}, ..., w_{q_n} >$

Query matching involves measuring the degree of similarity sim(d, q) between the query vector q and each document vector d. In our case, sim(d, q) is calculated as follows:

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

The IR system then produces a list of documents sorted by order of similarity with the query.

3.2. Experiments

Our experiments are conducted on a French corpus used in TREC-6 (Texte Retrieval Conference) [8]. The corpus contains a collection of articles from a Swiss newspaper - SDA (Schweizerische Depeschen Agentur) û French edition, published between 1988 and 1990. There are 141,656 documents, for a total size of 87 megabytes. TREC-6 data includes 25 queries, each written in English, French and German versions. Manual evaluations for 22 of these have been made available by NIST (National Institute of Standards and Technology). Our evaluations are based on this data: the French documents and the French and English queries.

We compared five different approaches:

1. Monolingual French query-French documents IR. This is not CLIR, but is used as a reference point with which CLIRperformance is compared.

In the other approaches, the English query is translated into a French query using various tools. The translated queries are then used to retrieve French documents in the same way as in monolingual IR. We tested the following translation approaches :

2. Using MT systems (two of them: LOGOS and SYSTRAN) ;
3. Using a bilingual dictionary only;
4. Using a probabilistic translation model;
5. Combining 3 and 4.

Each approach is now described in detail.

Monolingual IR

We used a classical IR approach to match the original French queries with the French documents, namely the vector space model described in Section 3.1. System performance is assessed by a standard IR method: average precision over 11 points of recall. We use the term *IR effectiveness* to refer to this particular measure.

In this monolingual task, our average precisionfor the 22 queries was 37.31%.

CLIR using MT

Two MT systems - LOGOS and SYSTRAN were tested. Globally, the resulting translations were reasonably good, maybe in part because of the relative linguistic simplicity of our English queries. The first three test queries are reproduced here:

English queries :

- Reasons for controversy surrounding Waldheim’s World War II actions.
- Are marriages increasing worldwide?
- What measures are being taken to stem international drug traffic?

LOGOS translations:

- Raisons pour les actions de deuxième guerre mondiale d'entourer de controverse ?Waldheim's.
- Les mariages augmentent-ils dans le monde entier ?
- Quelles mesures sont prises pour contenir la circulation de médicament internationale ?

SYSTRAN translations:

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

LOGOS flags the words missing from its dictionary with a question mark. In the case of the first query, the missing word `Waldheim` will still be considered during indexing because there are French documents that happen to contain it (fortunately, proper names tend to be preserved intact in translations). In other cases, words that the MT system did not know will end up being ignored at indexing. For example, one ofour queries contained the rare word ô reusageö which none of our MT systems knew.

As stated earlier, the (sometimes questionable) quality of translations with respect to syntactic structure has little effect on IR effectiveness. What is important is the choice of correct target language equivalents. Both LOGOS and SYSTRAN produced several instances of inappropriate choice. For example, one of our queries contained ôdrug trafficö; while SYSTRAN correctly translated this term as "trafic de drogue", LOGOS incorrectly translated it as ôcirculation de médicamentö. The same query contained the word `stem` used as a verb and SYSTRAN mistranslated it as the noun ô tigeö (ôtree stemö). Such errors lead to retrieving irrelevant documents.

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 meaning of ôdrugö in the sense of Query 3 may be expressed as ôdrogueö or ôstupéfiantö. By choosing to translate ôdrugö only by ôdrogueö, documents describing ôstupéfiantö cannot be retrieved.

Despite such errors, the translations produced by LOGOS and SYSTRAN scored relatively high: an average precision of 28.66% with LOGOS and 27.63% with SYSTRAN. These results appear very good in comparison with comparable tests conducted in TREC-6 [6, 7]: typically, the average precision of this method was only about 1/2 - 2/3 as high as monolingual IR.

CLIR using a bilingual dictionary

We obtained from the Ergane project a bilingual dictionary which contains 7898 citation forms in English. Each English word is translated into one or more French words. For example:

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

We tested a very simple approach: each word of an English query was replaced by all the French equivalents listed in the dictionary. For the first 3 queries, this resulted in the following word lists:

Query #1
cause, motif, raison.
polémique.
entourer.
?waldheim
monde.
guerre.
ii
activité, action.

Query #2
, trafic.
mariage.
accroître, agrandir, amplifier, augmenter, étendre, accroissement, grossir, s'accoître, redoubler, accroissement.
?worldwide

Query #3
quoi.
mesurer,mesure, taille.
tige, queue, tronc.
international.
remède, médicament, drogue, stupéfiant.
circulation

where ?waldheim and ?worldwide are unknown words. During indexing, the word worldwide will be ignored whereas waldheim will be indexed.

From the above examples, we can observe the following facts:

In some cases, our dictionary lookup only produces inappropriate translations. For example, ôstemö is used as a verb in the third query, but the dictionary only gives equivalents for the nominal sense as in ôtree stemö (tige, queue, tronc). In many other cases, inappropriate translations are given along with some correct ones. Thus, ôdrugö receives the correct equivalents drogue and stupéfiant, but also the inappropriate remède and médicament. On one hand, in failing to choose between distinct meanings of a source language word (drogue ≠ médicament) the dictionary method will produce additional retrieval noise; on the other hand, in refraining from arbitrarily selecting between target language synonyms (drogue ≅ stupéfiant) the method performs a natural query expansion which will reduce retrieval silence.

We also observe that the dictionary is not well distributed in the sense that less important words (from the IR point of view) may have more translations than more important ones. For example, in query 2, the word ômarriageö has only one translation, whereas the word ôincreaseö has 10 translations. As a consequence, documents containing a word meaning ôincreaseö will have a higher chance to be retrieved than a document about ômarriageö. Bilingual dictionaries do not seem to reflect the notion of importance that is relevant for IR.

Our test queries contained few words that were missing from our dictionary, despite its limited size. No doubt, this is because the queries were mostly about general topics.

Our dictionary-translated queries scored an average precision of 18.33%, that is, about 50% of our monolingual score.

A variant of this approach consists in using a bilingual terminology database instead of a bilingual dictionary. In contrast with dictionaries, terminology databases tend to contain a lot of complex terms. Moreover, the terms are usually classified into domains. Consequently, one would expect terminology databases to provide a better basis on which to choose accurate indices for IR queries.

We tested this approach using the ôBanque de Terminologie du Québecö (Terminology database of Quebec û BTQ). This database contains over 500 000 terms in English and French, classified into about 160 domains. Most terms are highly specialized. Thus, the database is very rich in domain-specific information. On the other hand, words and expressions of everyday language are often missing. For example, in Query 1 "Reasons for controversy surrounding Waldheim's World War II actions", only the following words are found in BTQ: surround, ii, action. In addition, matched words are often assigned very idiosyncratic meanings in different specialized domains. In Query 2 "are marriages increasing worldwide ?", none of the words is found. Replacing the original query with BTQ matches does not result in anything close to a reasonable translation. As a result, our average precision was only about 8%, a performance well below our dictionary approach. We conclude that a highly specialized terminology database such as BTQ is not appropriate for general CLIR.

CLIR using a probabilistic translation model

Query translation is performed as follows. An English query **e** is submitted to the probabilistic model as a single sentence so as to calculate $p(f|\mathbf{e})$, the probability that word f will occur in any translation **f** 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: probability estimations are then notoriously unreliable. By limiting the resulting list to an appropriate length, the amount of noise may be reduced.

Thus, our ôtranslationö of a query **e** will be simply made up of the *n* words *f* for which $p(f|\mathbf{e})$ is highest. We will experiment with several values of *n* in order to assess how this parameter affects IR effectiveness.

The resulting list of words is used to retrieve documents in the same language (French). The following lists show the first 20 words in the translations of the first 3 queries of our test corpus and their probabilities.

Query #1

```
.=0.117685
affaire=0.069960
waldheim=0.067383
guerre=0.062125
,=0.059158
raison=0.048319
ii=0.047925
monde=0.043656
controverse=0.038537
entourer=0.036864
mesure=0.022972
mondial=0.019244
prendre=0.018364
second=0.015948
suite=0.013105
action=0.011012
susciter=0.006899
donner=0.006639
pouvoir=0.006223
cause=0.005515
```

Query #2

```
mariage=0.172244
?=0.152780
.=0.088396
augmenter=0.056274
mondial=0.044127
augmentation=0.042161
,=0.034191
monde=0.030830
accroître=0.017007
hausse=0.016589
entier=0.016356
pouvoir=0.011882
union=0.011524
marier=0.007423
fait=0.005743
international=0.00551
6
parenté=0.005454
séparation=0.005454
connaître=0.005317
apparenté=0.005290
```

Query #3

```
médicament=0.110892
?=0.103753
mesure=0.091091
international=0.08650
5
.=0.067732
trafic=0.052353
drogue=0.041383
,=0.040058
découler=0.024199
circulation=0.019576
pharmaceutique=0.01872
8
pouvoir=0.013451
prendre=0.012588
extérieur=0.011669
passer=0.007799
demander=0.007422
endiguer=0.006685
nouveau=0.006016
stupéfiant=0.005265
produit=0.004789
```

Punctuation symbols are treated as ordinary words because we failed to remove them from consideration in our training. This has little impact because they are ignored during query indexing. We plan to remove them altogether in our future experiments.

Some interesting facts may be observed in these lists:

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

2).This dependence on the training corpus sometimes leads to odd translations. For example, the word ô billeö is considered as a French translation of ôloggingö in the English query ôeffects of logging on desertificationö. 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 ôacupunctureö in the training corpus. Because of that, the model fails to assign a higher probability to the French ô acuponctureö than to other semantically unrelated words that appeared in the same sentence.

4). The model sometimes fail to distinguish the real translation from noise induced by simple statistical associations. For example, the word ô prendreö appears in the translations of queries 1 and 3. It is attributed with even higher probabilities than the true translation words of the query such as "second", "action" and "stupéfiant". Statistics alone may prove insufficient for tackling this problem correctly.

Despite these problems, we observe that real translations and associated words tend to score relatively high and appear at the top of the list. When the probabilities are incorporated into the query vector used to retrieve documents, the documents containing these words will be retrieved in priority.

What use should we make of the probabilities that our translation model associates to each word? Should we use them directly as the weights appearing in our query vector? Should we rather combine them with other information?

Notice that the probabilities assigned by the translation model are related to the *tf* (term frequency) criterion of IR: our definition of $p(f|\mathbf{e})$ is such that each individual occurrence of a word *e* in the query **e** will reinforce the *f*Æs that are likely translations for *e*.

However, our translation model has little to say about the other criterion that is so important in IR: *idf* (inverse document frequency). One possible way to derive a *tf*idf*-like weighting is to use the following transformed weight in the query vector:

$$w_{f_i} = p(f_i | \mathbf{e}) * \log(N/n)$$

where $p(f_i| \mathbf{e})$is the probability obtained by the probabilistic translation model, and $\log(N/n)$ represents the *idf* criterion as described in section 3.1.

In our experiments, we tested different lengths of the list of translation words, as well as the two weighting methods in query vectors. The following table shows the IR effectiveness obtained in different cases.

Length of the list of translation words	Using the probability as weight	Using the transformed weight
10	23,45%	25,46%
20	24,15%	26,35%
30	24,28%	26,60%
40	24,33%	26,64%
50	24,38%	26,71%
100	22,51%	25,06%

We observe that when the length of the translation word list increases from 10 to 50, the retrieval effectiveness increases slightly. However, when the length becomes too high (100), the effectiveness declines. This phenomenon may be explained as follows: the more wordswe retain in the translation: 1) the more related words get to be included; but 2) the more unrelated words get tobe included as well. A good compromise is needed. Comparing lists of length 100 with shorten ones confirms our intuition that ignoring words with low probabilities reduces the risk of incorrect word associations, thus the risk of retrieving irrelevant documents.

It is also evident that the transformed weighting which takes into account the *idf*-criterion produces better results than translation probabilities alone.This is just another confirmation of the importance of the *idf*-criterion in IR.

We mentioned above that our probabilistic translation model is sometimes unable to distinguish true translations from accidental statistical associations. We thought it might help to incorporate additional evidence of a true translation relationship if any such evidence was available. It is often the case in IR that combining different sources of evidence increases IR effectiveness. This is why we tried combining our probabilistic translation model with the bilingual dictionary mentioned above.

Combining the probabilistic translation model with a bilingual 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 French translation is stored in the bilingual dictionary, its probability is increased by a *default value*, a constant determined manually. The new ôprobabilityö is used to obtain the transformed weight for the query vector as before. We tested several default values, ranging from 0.005 to 0.05. The following table shows the IR effectiveness obtained in each case.

Default value	Length of the list oftrans			
	10	20	30	40
0.005	26,71	27,87	28,12	28,13
0.01	27,55	28,73	28,91	28,96
0.02	28,73	29,59	29,62	29,67
0.03	28,11	29,06	28,98	28,97
0.04	27,51	28,42	28,27	28,26
0.05	26,87	27,61	27,29	27,29

First and foremost, note that in all cases the combined resources yield better retrieval effectiveness than either the probabilistic model alone or the bilingual dictionary alone. This strongly confirms our intuition that combining two sources of information should produce better results.

In many of the tested cases the combined approach outperform the MT systems. In the case where the default value is 0,02, and 50 translation words are retained, we obtained the best effectiveness 29,85% (among all the tested cases). It may be claimed here that there are better tools for CLIR than MT systems.

Although the improvements in effectiveness of the combined approach over MT systems obtained so far are still small, we think that this approach may be further improved by 1) using a better training corpus; 2) using a more complete bilingual dictionary; and 3) a better method of combination. It is also possible to combine our probabilistic translation model with an MT system. As these two methods are based on different knowledge sources, the results could well prove superior too. We plan to examine this combination in the future.

We are currently doing experiments on two other test corpora. We expect to report on the results in the final version of this paper.

4. Conclusions

MT systems are considered by many as appropriate tools for CLIR. In this paper, we showed that there are better tools for CLIR than MT. We investigated the possibility of using a probabilistic translation model built automatically from a parallel corpus. 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.

When applied to CLIR, MT systems (LOGOS and SYSTRAN) can give a relatively good performance. Simpler approaches based only on bilingual dictionaries or terminology databases like BTQ lead to much poorer performance. Our probabilistic translation model almost rivals the performance of the MT systems, despite the fact that our training corpus is not closely related to the test corpus.

In our experiments, we observed different advantages and disadvantages for different approaches to translate queries from a language to another. They often have complementary properties, and may be successfully combined. In this study, we combined our probabilistic translation model with a bilingual dictionary. This combination outperformed the MT systems, leading us to the conclusion that there are better approaches to CLIR than MT.

In all cases, the performance of CLIR remains substantially lower than that of monolingual IR. Thus there is still a lot of room for further improvement. There may not be any single translation method that will fill the bill. We believe that progress is likely to come from combining various sources of translation knowledge and we intend to continue testing such methods in our future research.

References

1. J. Brousseau, C. Drouin, G. Foster, P. Isabelle, R. Kuhn, Y. Normandin, and P. Plamondon., French speech recognition in an automatic dictation system for translators: the TransTalk project. *Eurospeech 95*, Madrid, Spain, 193-196 (1995).
2. 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).
3. C. Buckley, Implementation of the SMART information retrieval system. Cornell University, Technical report 85-686, (1985).
4. M. Dymetman, J. Brousseau, G. Foster, P. Isabelle, Y. Normandin, and P. Plamondon, Towards an automatic dictation system for translators: the TransTalk project. *ICSLP 94*, Yokohama, Japan, 691-694 (1994).
5. G. Foster, P. Isabelle, and P. Plamondon, Target-text Mediated Interactive Machine Translation. *Machine Translation*, vol. 12, pp. 175-194 (1997).
6. W.A. Gale, K.W. Church, A program for aligning sentences in bilingual corpora, *Computational Linguistics*, 19 :1, 75-102, (1993).
7. G. Grefenstette, Cross-Language Information Retrieval. : Kluwer Academic Publisher, (1998).
8. D. K. Harman and E. M. Voorhees, Text REtrieval Conference (TREC-6). . Gaithersburg, (1997).
9. G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*: McGraw-Hill (1983).
10. 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).