CLIR using a Probabilistic Translation Model based on Web Documents

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

In this report, we describe the approach we used in TREC-8 Cross-Language IR (CLIR) track. The approach is based on probabilistic translation models estimated from two parallel training corpora: one established manually, and the other built automatically with the documents mined from the Web. We describe the principle of model building, the mining of parallel texts, as well as some preliminary evaluations.

1. Introduction

Last year, in TREC7, we compared three possible approaches to CLIR (for French and English), namely, the approach based on a bilingual dictionary, the approach based on a machine translation (MT) system, and the approach based on a probabilistic translation model using parallel texts. It has been shown that the dictionary-based approach did not give satisfactory performance. The approach using an MT system gave a good performance. In the case of the probabilistic model, the performance was close to that of MT approach.

In TREC7, the IBM group [Franz98] used a similar approach, but for document translation (instead of query translation as in our case) and using long queries (instead of short queries in our case for TREC7). Their system was one of the bests in TREC7 CLIR runs. This is an encouraging result that shows the approach based on a probabilistic model may perform very well.

In TREC8, our goal is to continue using our approach based on parallel texts, but we want to test the performance of a probabilistic model that is estimated from a set of parallel texts automatically mined from the Web. The purpose of these tests is to see if automatic mining of parallel texts may be a possible solution to the problem of unavailability of parallel texts for several language pairs. For the moment, we only have a model estimated for English-French. So our submitted runs only concern English and French documents (AP and SDA collections) using either English or French queries. Two sets of runs have been submitted: one with a probabilistic model trained with a manually established corpus - the Hansard; and the other with a model trained by the Web texts.

In the following sections, we will first recall the principle of building a probabilistic translation model from parallel texts. Then we will describe briefly the way in which parallel texts are mined from the Web. Finally we will give a description of some experimental results.

2. Principle of building a probabilistic translation model

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

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 some initial probabilities of word translation.

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

3. Mining parallel texts from the Web

The problem we often have with probabilistic models is the unavailability of parallel texts for many language pairs. The Hansard corpus is one of the only existing corpora for English and French. For other languages (e.g. Chinese and English), such a corpus is less (or not at all) available. In order to solve this problem, we conducted a text-mining project in the Web in order to find parallel texts automatically. The first experiments with the mined documents have been described in [Nie99]. The experiments were done with a subset (5000) of the mined documents. However, they showed that the approach is feasible. In TREC8, we intend to evaluate the performance of a probabilistic model trained with all the parallel documents we found (about 20 000 pairs).

The mining process is devised into several steps:

- selection of candidate web sites
- finding all the documents from the candidate sites
- paring the texts using simple or sophisticated criteria

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

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

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

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

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

The above process was launched and stopped after 75 hours. We obtained about 20 000 pairs that amount to 135 Mbytes French texts and 118Mbytes English texts. It is to be noticed that only 30% of 5474 candidate sites have been explored.

4. Experiments

We used a modified version of SMART system [Buckley85] for monolingual document indexing and retrieval. The *ltn* weighting scheme is used for documents. For queries, we used the

probabilities provided by the probabilistic model, multiplied by the *idf* factor. From the translation words obtained, we retained the top n words. The value of n is determined using TREC6 and TREC7 data.

4.1. Tests with TREC6 and TREC7 data

The purpose is to determine the optimal value of n (the number of translation words kept for each query) for each direction (E to F or F to E) and each model. The test runs gave the following performances (measured in average precision) using the long queries:

English to French

n	**Hansard**		**Web**	
	TREC6	TREC7	TREC6	TREC7
10	0.2745	0.2685	0.2642	0.2554
15	0.2842	0.3102	0.3193	0.2641
20	0.2861	0.3215	0.3146	0.2918
25	0.2932	0.3184	0.3160	0.2963
30	0.2930	0.3193	0.3242	0.3043
35	0.2930	0.3219	0.3239	0.3076
40	0.2932	0.3241	0.3242	0.3076
45	0.2937	0.3238	0.3258	0.3078
50	0.2938	0.3246	0.3277	0.3083
60	**0.2950**	**0.3249**	0.3278	0.3124
70	0.2943	0.3248	**0.3288**	**0.3125**
80	0.2894	0.3244	0.3279	0.3124
90	0.2893	0.3238	0.3279	0.3131
100	0.2900	0.3242	0.3274	0.3127

French to English

n	**Hansard**		**Web**	
	TREC6	TREC7	TREC6	TREC7
10	0.2675	0.3855	0.2857	0.3584
15	0.2959	0.3879	0.2992	0.3606
20	0.2944	0.3898	0.3047	0.3665
25	**0.2943**	0.3918	**0.3105**	0.3721
30	0.2936	**0.3978**	0.3102	0.3732
35	0.2929	0.3721	0.3095	0.3738
40	0.2929	0.3699	0.3099	0.3741
45	0.2884	0.3666	0.3097	0.3746
50	0.2690	0.3669	0.3086	0.3740
60	0.2697	0.3371	0.3089	0.3744
70	0.2696	0.3250	0.3097	**0.3748**
80	0.2696	0.2987	0.3097	0.3743
90	0.2692	0.2982	0.3092	0.3744
100	0.2688	0.2981	0.3090	0.3742

Fig. 1. Tests of the models on TREC6 and TREC7

As we can see in these tables, in the case of the Hansard model, the optimal number of translation words is 60 for English to French translation, and about 30 for French to English translation. In the case of the Web model, the number of 70 seems to be quite good for all the cases. Therefore, these numbers have been chosen.

Each translated query, a list of weighted words, is further transformed by the mtn weighting scheme of SMART. It is then run against the documents of the target language. In parallel, the documents in the target language are retrieved using the original query in the target language. The two sets of results are merged and ordered according to their similarity with the queries. The following four runs have been submitted (all for only English AP and French SDA collections):

RaliHanE2EF: Using English queries and the Hansard model
RaliHanF2EF: Using French queries and the Hansard model
RaliWebE2EF: Using English queries and the Web model
RaliWebF2EF: Using French queries and the Web model

What we can also observe in the above table is that the Web model performs generally slightly better than the Hansard model. In [Nie99], with the limited web model trained with 5000 pairs of parallel texts, the performance was not as good as that of the Hansard model. The above tables show that with enough parallel texts from the Web (actually about the same volume of texts as in the Hansard), we can do as well as with a well controlled parallel corpus.

4.2. Evaluation of the submitted runs

From the official evaluation, we extracted those for AP and SDA collections. We use this set of judgements as our reference. The following table gives the average precision for each run.

E2EF		F2EF	
Hansard	Web	Hansard	Web
0.3027	0.2744	0.3002	0.3012

Fig 2. Merged CLIR tests with TREC8 queries

This table shows that the Web model performs slightly worse than the Hansard model in the E2EF case. In F2EF, the performances are equivalent. At this point, several questions may be raised: Why the optimal numbers set for TREC6 and TREC7 do not work well for TREC8? Is this difference due to the different numbers of translation words used in different runs? To the difference between the sets of queries? Or to the merging method used?

These questions can only be answered when we have thoroughly analyzed the translation and retrieval results with different models. This will be reported later.

Notice that the submitted runs do not use a combination of a probabilistic model and a bilingual dictionary. Our previous tests all confirmed that such a combination improve the performances. Our goal in TREC8 is solely to compare the two probabilistic models. Therefore, the possible improving techniques (such as the combination as well as quasi-relevance feedback) that may be used for both models are not used. In so doing, we hope to be able to have a more clear comparison between the models.

In order to evaluate the performances of the translation models, without considering the problem of result merging, we compare the result of simple cross-language results (from English query to French documents, or vice versa) with those of the monolingual runs. The following table shows this comparison.

French mono: 0.3946		English mono: 0.3090	
E2F (% mono)		F2E (% mono)	
Hansard	Web	Hansard	Web
0.3253 (84%)	0.3109 (79%)	0.2842 (92%)	0.2784 (90%)

Fig. 3. Single CLIR runs

Let us look at some of the problems in query translation.

Wrong translations of the key concepts:

Query 59 - exportation of dangerous medicines

We observe a drastic drop in the case of web model (from 0.4062 of monolingual run to 0.0448 only). The reason is the wrong translation of "medicines" as "médecine" (medical area). We obtained the same performance as the monolingual run using the Hansard model.

Query 60 - Rare Birds Stolen

The Hansard model translated "bird" by itself, and attributed the strongest probability to it. This is the main reason of drastic drop of effectiveness: 0.0568 compared to 0.3392 in monolingual run (we obtained 0.1684 with the Web model). By both models, several terms such as "navire" (ship) have been given very strong probabilities. These terms are translations of some less important terms in the English query (e.g. "shipped") in the description field.

Query 71 - saving the dolphin

In the Web model, the word "dolphin" is translated by itself. This topic also contains the word "net" (fishing net). This raised a lot of problem to both translation models. It is translated as "net" (an adjective in French, or Internet).

Wrong proper name

The French word "dauphin" has been translated as "dauphin" by both translation models. In the case of the Hansard model, this is because "Dauphin" is the name of a place. In the Web model, "Dauphin" is also taken in this way. We found many occurrences in the English Web documents talking about "Dauphin Lake Basin", or phone number in "Dauphin".

Unknown words:

In the Query 64, the words "fertilizer" and "fertilizing" are unknown to the Hansard model. Whereas only "fertilizing" is unknown to the Web model. As a consequence, the CLIR run with the Web model (0.2759) is comparable to that of the monolingual run (0.2519), whereas that with the Hansard model is much worse (0.0260).

"ONU" (UN) is an important concept in French query 61 (on "German UN force"). However, its translation "United Nation" is only attributed with low probabilities (especially by the Web model). A possible reason is the very low frequency of occurrences of "ONU" in the training corpus.

For the Web model, the word "Galiciens" is an unknown French word. For the Hansard model, the situation is even worse: "Catalan", "Galice" and "ETA" are also unknown. The performances obtained with the translations are only 1/2 and 1/3 of the monolingual run.

Related words included

In several cases, we observed the interesting phenomenon that related words are also included in the translation. These related words may be even absent in the original queries. For example, the word "movie" does not appear in the English query about "European film industry". In the translations (by moth models) from French to English, it is included and attributed with a strong probability. As a consequence, the translated queries lead to higher effectiveness (around 0.26) than the original English query (0.1544). The same phenomenon is observed for Query 65 (on synthetic fertilizing): From the French query, some related English words (that are absent in the original English query) have been included in the translations (e.g. environment). In this case, the translated queries also lead to higher performance than the monolingual English run.

5. Final remarks

Some of the above mentioned problems may be solved to certain extent by using the translation models in conjunction with a bilingual dictionary. For example, the unknown words problem and the wrong proper name problem. Such a combination has proven to be effective [Nie99].

The other problems (especially the wrong translation problem) seem difficult to solve. However, it is to be noted that the same problem also occurs for query translation with any tool (MT or bilingual dictionary). These problems explain why CLIR effectiveness is usually lower than the monolingual runs, even with the best translation tools of the world. On the other hand, if we compare the probabilistic translation models with other translations means (in particular, with MT systems), their performances are very close [Nie99]. This suggests that probabilistic models are translation tools that are as valuable as MT systems for the CLIR purposes.

Our tests in TREC8 showed that using Web documents to train a probabilistic model is a reasonable approach. The final performance is only slightly lower than using a controlled parallel corpus. The great advantage of this approach is that it may be easily extended to several other

language pairs with little additional cost. We are extending this approach to several other language pairs.

References

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

[Gale93] W. A. Gale, K.W. Church, A program for aligning sentences in bilingual corpora, *Computational Linguistics*, 19 :1, 75-102 (1993).

[Franz98] M. Franz, J.S. McCarley, S. Roukos, Ad hoc and multilingual information retrieval at IBM, *The Seventh Text Retrieval Conference (TREC-7)*, NIST SP 500-242, pp. 157-168 (1998)

[Nie98] J.Y. Nie, TREC-7 CLIR using a probabilistic translation model, *The Seventh Text Retrieval Conference (TREC-7)*, NIST SP 500-242, pp. 547-553 (1998).

[Nie99] J.Y. Nie, P. Isabelle, M. Simard, R. Durand, Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web, *ACM-SIGIR conference*, Berkeley, CA, pp. 74-81(1999).

[Simard92] M. Simard, G. Foster, P. Isabelle, Using Cognates to Align Sentences in Parallel Corpora, *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal (1992).