Multilingual Information Retrieval based on Parallel Texts from the Web

Jian-Yun Nie, Michel Simard, 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, simardm, foster}@iro.umontreal.ca

Abstract. In this paper, we describe our approach in CLEF Cross-Language IR (CLIR) tasks. In our experiments, we used statistical translation models for query translation. Some of the models are trained on parallel web pages that are automatically mined from the Web. Others are trained from bilingual dictionaries and lexical databases. These models are combined in query translation. Our goal in this series of experiments is to test if the parallel web pages can be used effectively to translate queries in multilingual IR. In particular, we compare models trained on Web documents with models that also combine other resources such as dictionaries. Our results show that the models trained on the parallel web pages can achieve reasonable CLIR performance. However, combining models effectively is a difficult task, and single models still yield better results.

1. Introduction

In Cross-Language Information Retrieval (CLIR), the usual approach is to translate queries to the target language of the documents. One of the ways to perform query translation is to use a large set of parallel texts to train a statistical translation model. This approach has been successfully applied in previous CLIR experiments [4]. However, a possible obstacle is the lack of parallel texts for many language pairs. In order to overcome this obstacle, we conducted a research project to try to find parallel web pages automatically. In the past two years, we were able to build models for French-English and Chinese-English translations. Our results showed comparable performance to MT systems.

This year, we successfully mined several sets of parallel Web pages for the following language pairs: English-Italian, English-German, in addition to the English-French corpus we found previously. Our goal in this year's CLEF experiments is to see if the parallel Web documents can also apply to multilingual IR.

In our previous experiments, we observed that a certain combination of the translation models with a dictionary could improve IR effectiveness. However, the combination remained ad hoc: a dictionary translation is attributed a certain "default

probability" and combined with translation words provided by a statistical translation model. In CLEF experiments, we tested a new combination method. First, a dictionary is transformed into a statistical translation model. This is done by considering a word/term and its translation words/terms as two parallel texts. Then different statistical translation models are combined linearly. The parameters of the combination are set so as to maximize the translation probability of held-out data.

In this paper, we will first describe the mining system we used to gather parallel texts from the Web. Then a brief description of the training process of the statistical model will be provided. The CLEF experimental results will be reported. We provide some analysis of the translation process before the concluding remarks.

2. Mining parallel texts from the Web

Statistical models have often been used in computational linguistics for building MT systems or constructing translation assistance tools. The problem we often have is the unavailability of parallel texts for many language pairs. The Hansard corpus is one of the few existing corpora for English and French. For other languages, such a corpus is less (or not at all) available. In order to solve this problem, we conducted a text-mining project on the Web in order to find parallel texts automatically. The first experiments with the mined documents have been described in [5]. The experiments were done with a subset (5000) of the mined documents. However, we obtained a reasonably high CLIR performance. This experiment showed the feasibility of the approach based on parallel web pages. Later on, we trained another translation model with all the Web documents found, and the CLIR effectiveness obtained is close to that with a good MT system (Systran).

The mining process proceeds in three steps:

1. selection of candidate Web sites
2. finding all the documents from the candidate sites
3. pairing the texts using simple heuristic 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 text such as "English version", "english", and so on; and similarly for English documents. The idea is, if a French document contains such an anchor text, the link to which the anchor is associated usually points to the parallel text in English (fig. 1).

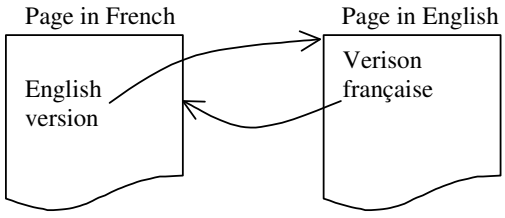

Fig. 1. Detection of candidate web sites

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. In addition, as search engines only index a subset of all the web pages at each site, a host crawler is used to explore each candidate site more completely. This crawler follows the links in each web page. If a link points to another web page on the same site, then this page is added to the collection of web pages. In this way, many more web pages have been found.

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

1. 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. Some of the parallel URLs are shown below:

Table 1. Examples of parallel URLs

French page	English page
www.booksatoz.com/french/Museumf.htm	www.booksatoz.com/Museum.htm
www.c3ed.uvsq.fr/esee/french/state.htm	www.c3ed.uvsq.fr/esee/english/state.htm
www.gov.nb.ca/dot/adm/adminf.htm	www.gov.nb.ca/dot/adm/admin.htm
www.psac.com/comp/upwe/upwe-f.htm	www.psac.com/comp/upwe/upwe-e.htm
www.psac.com/comm/news/9602007f.htm	www.psac.com/comm/news/9602007e.htm

   Therefore, by examining the URLs of the documents, we can quickly determine which files may be a pair.

2. We then use other criteria such as the length of the file to further confirm or reject a pair.

3. The above criteria do not require downloading the files. 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 French-English parallel corpus was constructed last year at the RALI laboratory. This year, we cooperated with Twenty-One (W. Kraaij) to construct English-Italian and English-German parallel corpora, using the same mining system - PTMiner [2]. The following table shows the number of text pairs as well as volume of the corpora for different language pairs.

Table 2. Training corpora

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

The corpora found from the Web will be called WAC corpora (Web Aligned Corpora). The models trained with these corpora will be called the WAC models.

3. Principle of building a probabilistic translation model

Given a set of parallel texts in two languages, it is first aligned into parallel sentences. The criteria used in sentence alignment are the position of the sentence in the text (parallel sentences have similar positions in two parallel texts), the length of the sentence (they are also similar in length), and so on [3]. In [6], it is proposed that cognates may be used as an additional criterion. Cognates refer to the words (e.g. proper names) or symbols (e.g. numbers) that are identical (or very similar in form) in two languages. If two sentences contain such cognates, it provides additional evidence that they are parallel. It has been shown that the approach using cognates performs better than the one without cognates. Before the training of models, each corpus is aligned into parallel sentences using cognate-based alignment algorithm.

Once a set of parallel sentences is obtained, word translation relations are estimated. First, it is assumed that every word in a sentence may be the translation of every word in its parallel sentence. Therefore, the more often a pair of words appears in parallel sentences, the better its chances of being a valid translation. In this way, we obtain the initial probabilities of word translation.

At the second step, the probabilities are submitted to a process of Expectation Maximization (EM) in order to maximize the probabilities with respect to the given parallel sentences. The algorithm of EM is described in [1]. The final result is a probability function $P(f|e)$ which gives the probability that f is the translation of e. Using this function, we can determine a set of probable word translations in the target language for each source word, or for a complete query in the source language.

4. The training of multiple models and their combination

For English and French, we also have other resources: the Hansard corpus (a set of parallel French and English texts from the Canadian parliament debates), a large terminology database (Termium) and a small bilingual dictionary (Ergane). A translation model is trained from the Hansard data, in the same way as for the Web documents (WAC).

In both the terminology database and the bilingual dictionary, we have English words/terms, and their French translations (words/terms). In some way, we can also think of these two resources as two sets of special parallel "sentences". Therefore, the translation probability between words can also be estimated with the same statistical training process. In this way, two additional translation models are estimated from them. In total, we obtain 4 different translation models between English and French from four different resources (in each direction). The question now is how we can combine them in a reasonable way.

We choose a linear combination of the models. Each model is assigned a coefficient denoting our confidence in it. The coefficients are tuned from a set of

"held-out" data - a set of parallel sentences (about 100K words), by using the EM algorithm to find values which maximize the probability of this data according to the combined model. This set is selected from different resources (distinct from those used for model training) so that it gives a good balance of different kinds of texts.

Finally, the following coefficients are assigned to each model:

Table 3. Parameters for linear combination of models

Model	Parameter
Ergane	0.041
Hansard	0.301
Termium	0.413
WAC	0.245

We observe that the combination seems to favor models with larger vocabularies. Termium is attributed the highest coefficient because it contains about 1 million words/terms in each language. The Hansard corpus and the WAC corpus contain about the same volume of texts. So their coefficients are comparable. The Ergane dictionary is a small dictionary that only contains 9000 words in each language. Its coefficient is very low The main reason for this is that the EM algorithm penalizes models which assign zero probabilities to target-text words, and models with small vocabularies will assign zero probabilities more often than those with large vocabularies. Therefore a larger model will usually be preferred over a smaller model, even though the translations it contains may not be as accurate. Although the coefficients we used are the best for the held-out data in the sense of maximizing its likelihood, they may not be suitable to our data in CLEF, and the maximum-likelihood approach may not be ideal in this context

5. Experiments

We used a modified version of SMART system [9] 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 50 words for each query. The value of 50 seemed to be a reasonable number on TREC6 and TREC7 data.

5.1. Monolingual IR

Monolingual IR results have been submitted for the following languages: French, Italian and German. This series of experiments uses the SMART *ltn* weighting scheme for queries as well. In addition, a pseudo-relevance feedback is applied, which uses the 100 most important terms among the top 30 documents retrieved to revise the original queries. The parameters used for this process is: α=0.75, and β=0.25. The results obtained are shown below:

Table 4. Monolingual IR effectiveness

	French	Italian	German
≥ medium	18	18	12
< medium	16	16	25
Av.p. With feedback	0.4026	0.4334	0.2301
Av.p. Without feedback	0.3970	0.4374	0.2221

The comparisons with medium runs are only done for the submitted runs with pseudo-relevance feedback. As we can see, a great difference can be observed in effectiveness in the above runs. Several factors have contributed to this.

1. The use of stoplist
   In the case of French, a set of stopwords is set up carefully by French speaking people. In the case of Italian and German, we used two stoplists found from the Web [7]. In addition, a small set of additional stopwords was added manually for Italian.

2. The use of a lemmatizer or a stemmer
   For French, we used a lemmatizer developed in the RALI laboratory that first uses a statistical tagger, then transforms a word to its citation form according to its part-of-speech category. For Italian and German, two simple stemmers obtained from the Web [8] are used. There is no particular processing for compound words in German. This may be an important factor that affected the effectiveness of German IR.

Overall, the French and Italian monolingual runs seem to be comparable to the medium performance of the participants; but the German run is well below the medium performance. We think the main reason is due to the lack of special processing on German (e.g. compound words).

5.2. Tests on bilingual IR

The bilingual task consists in finding documents in a language different from that of the queries. We tested the following bilingual IR: E-F (i.e. English queries for French documents), E-I and E-G. For this series of tests, we first used the translation models to obtain a set of 50 weighted translation words for each query. Unknown words are not translated. They are added into the translation words with a default probability of 0.05. The same pseudo-relevance feedback process as that in monolingual IR is used.

Between English and Italian, English and German, we only have the Web parallel documents to train our translation models. For French and English, we have multiple translation resources: the Web documents, the Hansard corpus, and two bilingual dictionaries. So we also compare the model with only the Web documents (the WAC model) and the model with all the resources combined (the Mixed model). The following table summarizes the results we obtained for bilingual IR. Only 33 queries have relevant documents, and are considered in these evaluations.

Table 5. Bilingual IR with different models

	F-E		I-E (WAC)	G-E (WAC)
	WAC	Mixed		
≥ medium	20	16	21	13
< medium	13	17	13	21
Av.p. With feedback	0.2197	0.1722	0.2032	0.1437
Av.p. Without feedback	0.2410	0.1728	0.2102	0.1456

The runs we submitted are those with pseudo-relevance feedback. These runs are compared with Medium runs in the above table. For F-E and I-E cases, the WAC models perform better than the medium. The Mixed model of F-E gives a medium performance. The comparison between the two translation models for French to English is particularly interesting. We expected that the Mixed model could perform better because it is trained with more data from difference sources. Surprisingly, its effectiveness is worse than the WAC model alone. We see two possible explanations for this:

- The combination of different resources is tailored for a set of held-out data that does not come from the CLEF document set. So there may be a bias in the combination.
- During the combination, we observed that the combination results tend to favor dictionary translations. A high priority is attributed to dictionary translations. This may also be attributed to the biased tuning of combination.

In Table 2, we showed that the I-E training corpus is smaller than both F-E and G-E corpora. However, the model trained with it seems to be better suited to our CLIR task than the G-E model. This may be due to two possible reasons.

1. The quality of the translation model is determined by not only the size of the training corpus, but also the correspondence of the training data to the application corpus.
2. The quality of the model is dependent on the languages and on the processing on them.

In our case, the processing on German is the weakest. In particular, we did not consider compound words in German. This may have had a great impact on the trained model. It is also in translating German queries that we encountered the most unknown words, as we can see in Table 6. Quite a number of them are compound words such as "welthandelsorganisation", "elektroschwachtheorie" and "golfkriegssyndrom".

Table 6. Number of unknown words encountered by WAC models

Model	F-E	I-E	G-E
Unknown words	67	30	128

Another observation of Table 5 is that the pseudo-relevance feedback we used led to a general decrease in effectiveness, especially in the case of the WAC model for F-

E. This may be due to the fact that the initial retrieval effectiveness is too low or the setting of the feedback parameters is not suitable.

In the case of F-E CLIR, we tested several models separately. The following table shows the effectiveness between French and English using different translation models. It also compares the effectiveness with and without pseudo-relevance feedback.

Table 7. Comparison of bilingual IR with different individual models

Model	WAC	Hansard	Termium	Mixed
Av.p. Without feedback	0.2410	0.2869	0.2182	0.1728
Av.p. With feedback	0.2197	0.2914	0.2359	0.1722

We can see that the mixed model performed worse than any of the individual models. This indicates clearly that the combination of the models is not suitable for the CLEF data. Again, the pseudo-relevance feedback did not have a uniform impact on effectiveness. In the case of the mixed model, the impact is almost null. In the Hansard and Termium models, the impacts are positive, whereas in the WAC model, it is negative.

This table clearly shows that the effectiveness in the official runs could be improved greatly by 1) a better relevance feedback process (or by removing this process), and 2) a better combination of models.

5.3. Multilingual runs

In our case, the multilingual runs are only possible from English to all the languages (English, French, Italian and German). In these experiments, we followed these three steps:

1. Translate English queries to French, Italian and German, respectively;
2. Retrieve document from different document sets;
3. Merge the retrieval results.

The translation of English queries to German and Italian was done by the WAC translation model (trained from the Web documents). For English to French, we also have the alternative of using the Mixed model. The translation words are submitted to the *mtc* transformation of SMART. This scheme is chosen because it leads to comparable similarity values between results from different data sets, therefore, makes the result merging easier. The merging is done according to the similarity scores. The top 1000 retrieved are selected as the final results and submitted for evaluation.

The following table describes the results of different runs. In the WAC column, all the models used to translate English queries are WAC models. In the Mixed case, only the English to French translation uses the Mixed model, whereas the other translations still use the WAC models.

Table 8. Multilingual IR effectiveness

	WAC	Mixed
≥ medium	14	12
< medium	26	28
Av.p. With feedback	0.1531	0.1293
Av.p. Without feedback	0.1548	0.1544

As we can see, these performances are all below the medium performance. One of the main reasons may be that the German monolingual retrieval does not use any linguistic preprocessing, and has a very poor effectiveness. This may greatly affect the multilingual runs. Another possible reason is the over-simplified merging method we used. In fact, in order to render the English monolingual runs compatible (in terms of similarity values) with other bilingual runs, we had to choose the *mtc* weighting scheme as for the other cases. In our tests, we observe that this weighting scheme is not as good as *ltc*. Therefore, the ease of result merge has been obtained at the detriment of English effectiveness.

We observe again the negative impact of the Mixed model in this task. When the WAC model for English-French is replaced by the Mixed model, the effectiveness decreases. This shows once again that the coefficients we set for different models are not suitable for the CLEF data.

6. Analysis of CLEF results

In analyzing the translation results, we observed several problems in query translation.

6.1. Translation of ambiguous words

The translation models we constructed are IBM Model 1. These models do not consider the context during translation. It is a word-by-word translation; i.e. each word is translated in isolation. Therefore, they cannot solve word ambiguity in translation. For example, the word "drug" may be translated as "médicament" or "drogue" in French. These two senses are included in the translations of all the models, as we can see in Table 8. The same phenomenon is produced for the word "union" (in a query on European Union), which is translated to both "union" and "syndicat".

Table 9. Related translation words of "drug"

Model	Translation	Prob.
Hansard	médicament	0.1027
	drogue	0.0464
	stupéfiant	0.0042
WAC	drogue	0.0862
	médicament	0.0692
	drug	0.0042
Termium	drogue	0.0889
	médicament	0.0534
	drug	0.0101
	médicamenteux	0.0049
	stupéfiant	0.0046
Mixed	drogue	0.0746
	médicament	0.0715
	stupéfiant	0.0062
	médicamenteux	0.0020
	remède	0.0018

Table 10. Related translation words of "union"

Model	Translation	Prob.
Hansard	syndicat	0.0781
	communauté	0.0358
	union	0.0323
	collectivité	0.0125
	syndical	0.0111
	syndiqué	0.0042
	cee	0.0036
	unir	0.0032
WAC	syndicat	0.0666
	union	0.0508
	syndical	0.0341
	communauté	0.0158
	ue	0.0153
	collectivité	0.0131
Termium	union	0.0961
	syndicat	0.0327
	communautaire	0.0146
	assemblage	0.0049
	assemblée	0.0133
	syndical	0.0123
	communauté	0.0094
	collectivité	0.0067
	community	0.0044
Mixed	union	0.0673
	syndicat	0.0546
	communauté	0.0185
	syndical	0.0167
	communautaire	0.0131
	collectivité	0.0098
	assemblage	0.0060
	assemblée	0.0055
	ue	0.0037

6.2. Translation of compound terms

The example of "European Union" also shows the necessity to translate compound terms as a unit, instead of translating it word by word. By translating a compound term together, the word "union" in "European Union" is much less ambiguous than when it is translated in isolation. To do this, two approaches are possible. 1). One can detect compound terms in the parallel training texts before using the texts for model training. These compound terms will be considered as a "word" in the IBM model 1. If this "word" appears in a query, then it is translated as a unit (possibly by a

compound term in the target language). 2). On can also use a higher model than IBM 1. In fact, in IBM 1, words in a sentence are considered independently. In order to capture relationship between words in a compound term, or to capture some contextual information, it would be useful to use at least a language model together with the translation model. That is:

$$(f_1, f_2, \ldots f_l) = \text{argmax } P((f_1, f_2, \ldots f_l)|E) = \text{argmax } \prod_{i=(1,l)} P(f_i|E)*P(f_1, f_2, \ldots f_l). \quad (1)$$

where $P(f_1, f_2, \ldots f_l)$ is a language model that estimates the probability of $f_1, f_2, \ldots f_l$ appearing together in the target language, and $P(f_i|E)$ is the translation model.

In so doing, the best translation words would be those that not only have high translational probability, but also have a high probability to appear together in the target language. An alternative is to use IBM model 2 or 3 instead of model 1.

6.3. The effect of mixing models

From the above translation examples, we cannot observe any advantage of combining different models together. In fact, the mixed model only takes the translation words from different models, and re-calculates their probability according to a linear combination. This does not affect the ambiguity problem. Ambiguous words remain as ambiguous as before model combination.

The parameters used for linear combination of models are estimated on a small set of held-out data that are not necessarily adapted to the IR document collection used in these experiments. A better way to train the parameters is to use a similar IR test collection (e.g. the collections used for TREC 5, 6 and 7 CLIR tracks). The combination with better tuned parameters could allow us to achieve higher effectiveness than single translation models. We can also think about a different combination method than linear combination, or a method of estimating combining coefficients that corrects for vocabulary-size bias.

6.4. Coverage of the models

The effectiveness of each translation model may be strongly affected by their coverage. A model that produces many unknown words will not be able to translate many key concepts correctly (except for proper names). Table 6 showed the number of unknown words in translating queries to English from different languages. The G-E and I-E cases are comparable in both the size of training corpora and types of pre-processing on Italian and German. Nevertheless, there was a large difference between G-E run and I-E run. A strong factor that may have affected these performances is

unknown words. Below we show the case of a query on "electroweak theory"[1]. All the words marked * are unknown words.

Table 11. Word coverage in translating the query on "electroweak theory"

Model	Known words	Top translation words	
Hansard	*electroweak	théorie	0.0505
	*subnuclear *weinberg-	nucléaire	0.0497
	salam-glashow	découverte	0.0326
	*subatomic *quark	confirmer	0.0323
	*photon	proposer	0.0231
		domaine	0.0223
		modèle	0.0184
		physique	0.0182
WAC	*electroweak	théorie	0.0683
	*subnuclear *weinberg-	physique	0.0378
	salam-glashow	nucléaire	0.0353
	*subatomic *photon	découverte	0.0281
		domaine	0.0260
		modèle	0.0243
		proposer	0.0238
		confirmer	0.0238
Termiu	*weinberg-salam-	théorie	0.0864
m	glashow	électrofaible	0.0605
		nucléaire	0.0548
		physique	0.0457
		particule	0.0368
		infra-atomique	0.0303
		quark	0.0303
		modèle	0.0281
Mixed	*weinberg-salam-	théorie	0.0740
	glashow	nucléaire	0.0505
		physique	0.0336
		découverte	0.0298
		électrofaible	0.0250
		particule	0.0239
		modèle	0.0237
		confirmer	0.0229

As we can see, most unknown words are key concepts of the query. The Hansard model seems to have the worst coverage for this query. Most of the key concepts are

[1] Although this query does not contribute in official effectiveness measurements (because there is no relevant document in English and French collections), it does show the potential problem that low coverage may give rise to.

unknown. The model based on parallel web pages is slightly better. The Termium lexical database is the best for this query. It recognizes all the concepts, except the proper names. For this particular query, the effectiveness would have been greatly affected by the coverage of the models if its effectiveness were measured.

For other queries, there are only a few unknown words. In the English to French case, the Hansard model encountered 17 unknown words in total, the WAC model 13 and the Termium model 14. This appears surprising for the WAC model, which is a resource constructed without manual control. This shows that an automatically constructed parallel corpus can have a very good coverage.

7. Final remarks

In this CLEF, we successfully used parallel Web pages to train several translation models for language pairs other than English and French. Our experiments on mining the web for parallel texts further confirm that the automatic mining approach is feasible for many language pairs.

For monolingual IR, we used some basic IR methods, including simple stemmers and publicly available stoplists. The effectiveness for French and Italian monolingual IR is similar to the medium performance. The German monolingual run is well below the medium. We think the main reason is that we did not carry out any particular processing on German morphology, which is an important problem for German IR.

For bilingual IR between English and French, and between English and Italian, the effectiveness seems to be reasonable. It is better than the medium effectiveness. Between English and German, however, the effectiveness is well below the medium effectiveness. The reason may be the same as for the German monolingual run.

For multilingual runs, the performance is below the medium. We believe the reason is once again the low effectiveness for German. In addition, result merging may also have affected the global effectiveness.

Between English and French, we also tried to combine different resources in our translation models. We used a linear combination of the models trained with different data, including two dictionaries, a manually constructed parallel corpus, an automatically constructed parallel corpus and a lexical database. The coefficients of the combination were determined using a small set of held-out data. However, to our surprise, the mixed model performed worse than the model trained with the Web documents only. In fact, its effectiveness is lower than any of the individual translation models. This clearly indicates that the combination is not well suited to the CLEF data. In other words, the held-out data do not correspond to the document collection used in these IR experiments.

These experiments reveal several problems in using statistical translation models for CLIR. 1). The IBM model 1 has difficulty translating ambiguous words correctly. In order to deal with this problem, we will need to take into account a language model or use a more elaborate translation model in the future. 2). Compound terms should be translated as a whole, instead of being decomposed into single words. 3). Models should be combined in a better way. These are some of the problems we will study in our future research.

References

1. Brown, P. F., Pietra, S. A. D., Pietra, V. D. J., Mercer, R. L.: The mathematics of machine translation: Parameter estimation. Computational Linguistics, vol. 19 (1993) pp. 263-312
2. Chen, J.: Parallel Text Mining for Cross-Language Information Retrieval using a Statistical Translation Model, M.Sc. Thesis, DIRO, University of Montreal (2000)
3. W. A. Gale, K.W. Church, A program for aligning sentences in bilingual corpora, Computational Linguistics, 19:1 (1993) 75-102
4. Franz, M., McCarley, J.S., Roukos, S.: Ad hoc and multilingual information retrieval at IBM, The Seventh Text Retrieval Conference (TREC-7), NIST SP 500-242 (1998) 157-168
5. 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 (1999) 74-81
6. 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)
7. http://www.cs.ualberta.ca/~oracle8/oradoc/DOC/cartridg.804/a58165/appa.htm
8. http://www.muscat.com
9. ftp://ftp.cs.cornell.edu/pub/smart/