XRCE participation to CLEF2002

Jean-Michel Renders, Hervé Dejean, Eric Gaussier
XRCE
Firstname.Lastname@xrce.xerox.com
6, chemin de Maupertuis, 38240 Meylan – France

Abstract

In this paper, we describe the methods we used for the Cross-Lingual Evaluation Forum CLEF 2002, and more specifically for the GIRT Task. The methods are based on (1) the extraction of two bilingual lexicons, one from parallel corpora and the other one from comparable corpora, (2) the optimal combination of these bilingual lexicons in Cross-Language Information Retrieval and (3) the combination with monolingual IR on parallel corpora. While our original submission to CLEF2002 was restricted to short queries (using only the title field), we present here the results extended to complete queries.

1. Introduction

The GIRT Task of CLEF2002 consists of retrieving documents in a German corpus dedicated to Social Science, starting from English queries. For this cross-lingual task, resources such as the ELRA bilingual dictionary and the GIRT bilingual thesaurus are available. The corpus contains scientific articles, whose titles are translated in English. Several articles (about 6%) have their body translated as well.

When dealing with English queries on this corpus, the most obvious approach is the monolingual one, by considering only the translated parts of the German documents. Of course, this can be not satisfying, because the translated titles are a poor representation of the original content (actually, there are more than 20% articles whose titles are not even translated).

A more elaborated approach is to use the bilingual resources to translate the queries. However, it is well known that using these resources as such can raise covering problems (entries are missing ; translations are not domain-specific, etc.). It is therefore necessary to extend these resources by extracting specialized bilingual lexicons for the corpus. This can be done by alignment form parallel or comparable corpora.

The titles and the translated bodies constitute a parallel corpus. Starting from classical techniques of alignment from parallel corpora, a bilingual lexicon can be extracted, which gives $P_{par}(t|s)$, the probability of selecting target word t as translation of source word s. On the other hand, we could imagine to enrich the extracted lexicon by extracting another one from a comparable corpus. The lexicon built from the comparable corpus, besides being of more general use, can provide more reliable translation candidates when terms are of very low frequency in the parallel corpus. We describe later in the paper a technique to extract a bilingual lexicon from a comparable corpus, which results in providing $P_{comp}(t|s)$, the probability of selecting target word t as translation of source word s following this model.

The main idea of our approach is to optimize the combination of all these approaches:

- the monolingual approach
- the use of a lexicon extracted from the parallel corpus
- the use of a lexicon extracted from the comparable corpus.

Indeed, they all have different strengths and weaknesses and combining them should provide better performance than each approach alone.

This paper is organized as follows: sections 2 and 3 briefly describe the methods of bilingual lexicon extraction from parallel and comparable corpora that we used in CLEF-GIRT2002. Section 4 is dedicated to optimizing the combination of the methods, when solving Cross-Lingual Information Retrieval tasks. The corresponding experimental results are presented in Section 5. These results involve not only what we submitted to CLEF2002 (at that time, we restricted ourselves to the “short query problem”), but also the extension of the work to the more general problem (dealing with the complete topics, i.e. all fields of the queries).

2. Bilingual lexicon Extraction : Alignment from parallel corpora

Bilingual lexicon extraction from parallel corpora has received much attention since the seminal works of [5, 2, 7] on sentence alignment. Recent research has demonstrated that statistical alignment models can be highly successful at extracting word correspondences from parallel corpora (see [3, 6, 8] among others). We follow the approach proposed by [6] and represent co-occurrences between words across translations by a matrix, the rows of which represent the source language words[1], the columns the target language words, and the elements of the matrix the expected alignment frequencies for the words appearing in the corresponding row and column.

The estimation of the expected alignment frequency is based on the Iterative Proportional Fitting Procedure (IPFP) [1], which consists, given an initial estimate of all cell counts, in the following two computations at the k^{th} iteration, for each element of the matrix n_{ij}:

$$n_{ij}^{(k,1)} = n_{ij}^{(k-1,2)} \times \frac{m_{i.}}{n_{i.}^{(k-1,2)}} \quad (1.a)$$

$$n_{ij}^{(k,2)} = n_{ij}^{(k,1)} \times \frac{m_{.j}}{n_{.j}^{(k,1)}} \quad (1.b)$$

where $n_{i.}$ and $n_{.j}$ are the current row and column marginals, whereas $m_{i.}$ and $m_{.j}$ are the observed row and column term frequencies. Empty words are added in both languages in order to deal with words with no equivalent in the other language.

At each iteration, the algorithm considers each pair of aligned sentences, and updates local expected counts of the source and target words they contain through the above equations. The local expected counts are then summed up into global expected counts for the whole corpus, that will serve as initial values for local expected counts at the next iteration. Once the global expected counts are stable, they are normalized so as to yield probabilistic translation lexicons, where each source word is associated with a target word through a score. In the remainder of the paper, we will use $P_{par}(t|s)$, to denote the probability of selecting target word t as translation of source word s.

3. Bilingual lexicon Extraction : Alignment from comparable corpora

Bilingual lexicon extraction from non-parallel but comparable corpora has been studied by a number of researchers ([10, 9, 13, 12, 4] among others). Their work relies on the assumption that if two words are mutual translations, then their more frequent collocates (taken here in a very broad sense) are likely to be mutual translations as well. Based on this assumption, a standard approach consists in building context vectors, for each source and target word, which aim at capturing the most significant collocates. The target context vectors are then translated using a general bilingual dictionary, and compared with the source context vectors. This approach is reminiscent of the way similarities between terms are built in information retrieval, through the use of the cosine measure between term vectors extracted from the term-document matrix [11].

Our implementation is somehow different in the sense that we still use bilingual resources (here the ELRA dictionary and the GIRT thesaurus), but no vector translation is done. Instead, we compute the similarity between each word of the source corpus with each class in the bilingual resource. The same computation is done with the target side. Then the similarity between source words and target words is given by the following equation:

$$sim(s,t) = \sum_{C} p(C \mid s)\, p(C \mid t) \quad (2)$$

The different steps followed are:

- for each word w occurring in the corpus, build a context vector by considering all the words occurring in a window encompassing several sentences that is run through the corpus. Each word i in the context vector of w is then weighted with a measure of its association with w. We chose the log-likelihood ratio test to measure this association that we will denote v_{wi}.

[1] We use “source” and “target” to refer to elements of different languages, which does not imply, in our case, any privileged direction.

- for each class (entry) of the general lexicon and of the GIRT thesaurus, build a context vector using the context vectors and the weights previously defined. A class can have a context vector if and only if it has terms occurring in the corpus. If a class contains a multi-word term, the corresponding context vector is the intersection of the context vector of each word. If a class has several (equivalent) terms occurring in the corpus, the vector of this class corresponds to the union of the vectors of each term.
- the similarity of each source word *s*, for each class *C*, is computed on the basis of the cosine measure:

$$sim(s,C) = \frac{\sum_i v_{si} v_{Ci}}{\sqrt{\sum_i v_{si}^2 \sum_i v_{Ci}^2}} \quad (3)$$

  where the context word *i* ranges over the set of source words.
- The same is done on the target side, resulting in $sim(t,C)$
- The similarities are then normalized to yield a probabilistic translation lexicon:

$$P_{comp}(t|s) = \sum_C P(t \mid C)P(C \mid s) = \sum_C \frac{sim(t,C).sim(s,C)}{\sum_i sim(i,C) \sum_{Kj} sim(s,K)} \quad (4)$$

For practical reasons, the previous expansion (sum over all *C* classes) can be limited to the *n* closest classes as determined by $P(t|C)$ or $P(C|s)$. Experimentally, n=100 gives results very similar to the complete expansion.

For the GIRT task, we constituted our comparable corpus as follows. We have selected in the British National Corpus a subset of documents whose size is roughly equivalent to the size of the German GIRT corpus. This subset is chosen is such a way that it contains all words of the English queries (GIRT 2000 and GIRT 2001). Indeed, the translations of these words, as defined in the corresponding German queries, appear in the German corpus, which fairly ensures a comparable corpus.

4. Optimization

Three approaches are combined here: the monolingual approach, the bilingual approach based on the parallel corpus, and the one based on the comparable corpus.

The monolingual Information Retrieval method basically returns relevant German documents, just by considering the similarity of the English queries with the (translated) English titles of the documents constituting the parallel corpus. In our case, German documents are here indexed by the English words (in a lemmatized form) of their translated titles. Adopting the Vector Space model, let us note:

- q_e: the vector of (normalized) words of the English queries (NB. A stoplist is applied, which removes non relevant words)
- d_e: the vector representing a GIRT document, indexed by the normalized English terms of the titles.

Then, after applying standard weighting schemes on both query and document vectors, a first monolingual score can be computed and associated with each document:

$$s_{\text{english}} = \tilde{q}_e^t . \tilde{d}_e \quad (5)$$

where the "~" denotes weighted vector (see the experimental part to know which weighting schemes were applied).

When adopting the bilingual lexicon extracted from the parallel corpus and using it as a probabilistic translation lexicon, the new vector representing the query in the target language is given by $P_{par}(t|s).q_e$. Similarly, when adopting the bilingual lexicon extracted from the comparable corpus, the new vector representing the query in the target language is given by $P_{comp}(t|s).\ q_e$.

The first stage of the combination is to combine the lexicons extracted from parallel and comparable corpora. This can be done by a simple convex linear combination: the new vector representing the query in the target language is then given by:

$$q_g = [\alpha(s)P_{par}(t|s) + (1-\alpha(s))\ P_{comp}(t|s)].\ q_{e.} \quad (6)$$

Note that we introduced a dependence of the combining factor (α) on the particular word (s). To make the approach feasible, we divided the set of normalized English words into 3x3 subsets. The first dimension of the division is the frequency of the word in the (English) parallel corpus, discretized by 3 values (HIGH, MEDIUM, LOW). The second dimension is the proximity of the word to the thesaurus, defined by $\max_C sim(s, C)$ (the similarity is given by equation 3) and discretized as well. The dependence on s is now reduced to the dependence on these features (HIGH/MEDIUM/LOW frequency ; HIGH/MEDIUM/LOW proximity). The choice of optimal values for $\alpha(s)$ is our first set of degrees of freedom.

In practice, the vector q_g can be limited to its k largest components (in order to take a "finite" and small enough number of translation candidates). The value of the k threshold should be optimized as well (there is some trade-off between recall and precision when varying k).

After applying the probabilistic translation lexicons, the translated queries q_g and the vectors representing the documents indexed by the segmented and normalized German terms d_g are respectively weighted by a traditional SMART-like weighting scheme, in order to provide our second score:

$$s_{\text{german}} = \tilde{q}_g^t . \tilde{d}_g \quad (7)$$

Then both scores can be combined. This is the second stage of the combination:

$$s_{\text{final}} = \beta s_{\text{english}} + (1-\beta) s_{\text{german}} \quad (8)$$

Once again, the value of β must be optimized.

To briefly summarize the method, we have to find the optimum values of $\alpha(s)$, the k threshold, the β coefficient, as well as the weighting schemes for German documents, English titles, English queries and translated (German) queries. We have decided to optimize it with respect to the average precision (non interpolated) for the 50 queries of GIRT 2000 and GIRT 2001.

5. Results of experiences

As described above, there are a lot of parameters to be optimized. Therefore, we have adopted the following heuristic (non necessarily optimal) strategy. We began to search for the optimal weighting scheme for both queries and documents (β=0 or 1; α=1 ; k=100). Then, we optimized the threshold (k retained candidate translations). We then searched for a constant α, which resulted in the optimal value of the criterion (average non-interpolated precision), and then tried to find independently the values of α for each of the nine classes of features. Finally, we optimized β, keeping all other parameters at their optimal value found so far. Of course, this search strategy assumes some independence between the parameters. The experiments we have conducted so far seem to confirm this assumption.

The experiments are separated into two groups. The first one deals with "short" queries (queries limited to the "title" field). This group of experiments constituted our official submission to GIRT in June 2002. The second group considers "long" queries (all fields are used, even if some parts of the "narrative" field are automatically filtered).

The performance measure that we used here is the non-interpolated average precision. This measure is given:

- for the training setting (which consists of 48 queries of GIRT 2000 and 2001, and on a set of 16000 documents for which the relevance judgments were known)
- for the GIRT 2002 setting (23 English queries and the complete set of about 80.000 documents constituting the GIRT corpus).

The main results are given in the following table.

Query Type	β	weighting scheme query	weighting scheme document	*k*	α	average precision on training	average precision on GIRT 2002
Short	0	nin	lic	200	1	0.308	0.114
Short	0	nin	lic	200	0	0.105	0.081
Short	0	nin	lic	200	0.06	0.322	0.126
Long	0	nin	lic	25	1	0.327	0.166
Long	0	nin	lic	25	0.15	0.330	0.194
Long	0	nin	lic	25	α*	0.332	0.206
Long	1	lin	lic	--	--	0.212	0.098
Long	0.01	lin (e) / nin (g)	lic	25	α*	0.360	0.212

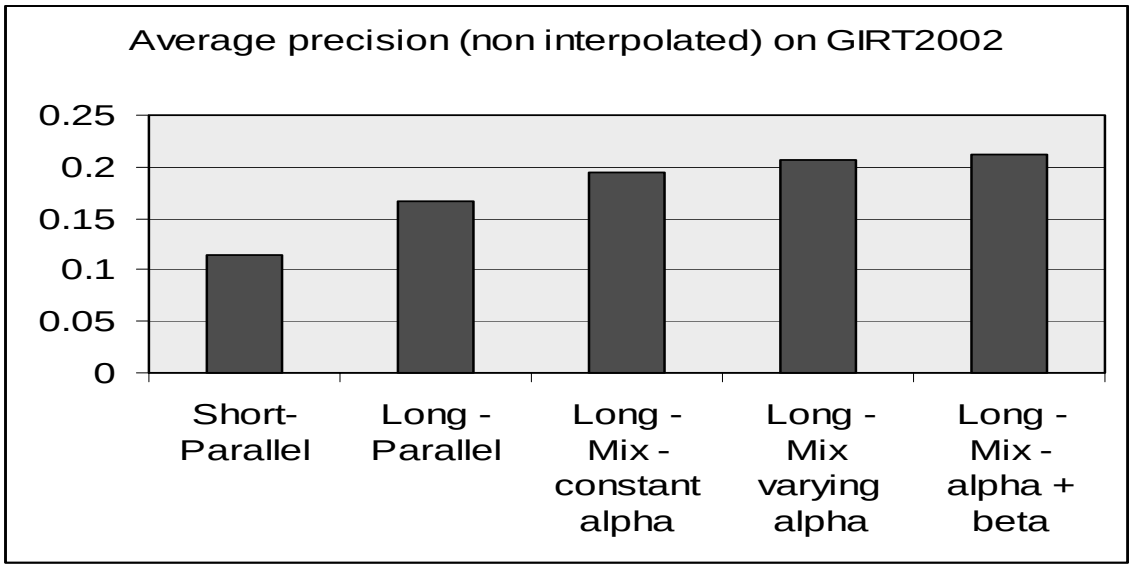
The weighting scheme "nin" only consists in applying the IDF (inverse document frequency) weighting, the "lin" scheme performs a logarithmic transformation of the term frequencies before the IDF weighting, while the "lic" scheme also performs a final normalization (cosine transformation). The optimal values of α are:

	HIGH frequency	MEDIUM frequency	LOW frequency
HIGH similarity	0.10	0.13	0.18
MEDIUM similarity	0.13	0.15	0.21
LOW similarity	0.15	0.20	0.21

The relatively low values of α are due to the difference between the distribution of $P_{par}(t|s)$ and $P_{comp}(t|s)$. The distribution obtained with the comparable corpus is flatter, with lower values for the most likely translation candidates when compared with the best candidates obtained with the parallel corpus. So, in order to take into account the influence of the lexicon extracted from the comparable corpus, α must be kept small.

The optimal values of α can be explained intuitively at least for the evolution with the "similarity" feature: the lexicon originated from the comparable corpus is less reliable for words with lower similarity with respect to the thesaurus. On the other hand, the dependency of α along the "frequency" dimension seems to be less obvious and needs more investigations.

The following chart summarizes the progression of the average precision (non-interpolated) on GIRT2002, when adopting the different stages of the combination.



6. Conclusions

Starting from three single approaches to cope with Cross-Lingual Information Retrieval tasks, we have shown how to combine them in an optimal way. Experimental results show that such combination provide performance in the GIRT2002 task which is better than the performance obtained by each approach taken in isolation. Significant improvement was achieved mainly for the queries of this year (GIRT), where it appears that taking into account the complete query and enriching lexicons by exploiting comparable corpora can bring important benefits.

Bibliography

[1] Y. Bishop and S. Fienberg and P. Holland. *Discrete Multivariate Analysis*, MIT Press, 1975.

[2] P. Brown and J.C. Lai and R.L. Mercer. Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics*, pp. 169-176, 1996.

[3] P. Brown and S. Della Pietra and V. Della Pietra and R.L. Mercer. The mathematics of Statistical Machine Learning Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263-311, 1993.

[4] P. Fung, A Statistical view on Bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In Jean Veronis, editor, *Parallel Text Processing*, 2000.

[5] W. A. Gale and K. W. Church. A Program for Aligning Sentences in Bilingual Corpora. In *Meeting of the Association for Computational Linguistics*, pp. 177-184, 1991.

[6] D. Hiemstra. Using Statistical Methods to create a Bilingual Dictionary. Master's Thesis, Universiteit Twente, 1996.

[7] M. Kay and M. Röscheisen. Test-translation alignment. *Computational Linguistics*, 19(1):121-142, 1993.

[8] I. Dan Melamed, A Word-to-Word Model of Translational Equivalence. In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 490-497, 1997.

[9] C. Peters and E. Picchi. Capturing the Comparable: A System for Querying Comparable Text Corpora. In *JADT'95 - 3rd International Conference on Statistical Analysis of Textual Data*, pp. 255-262, 1995.

[10] R. Rapp. Identifying word translations in nonparallel texts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1995.

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

[12] I. Shahzad and K. Ohtake and S. Masuyama and K. Yamamoto. Identifying Translations of Compound Nouns Using Non-aligned Corpora. In *Proceedings of the Workshop MAL'99*, pp. 108-113, 1999.

[13] K. Tanaka and Hideya Iwasaki. Extraction of Lexical Translations from Non-Aligned Corpora. In *International Conference on Computational Linguistics*, COLING'96, 1996.