Using Query-Relevant Documents Pairs for Cross-Lingual Information Retrieval*

David Pinto[1,2], Alfons Juan[1], and Paolo Rosso[1]

[1] Department of Information Systems and Computation,
Polytechnic University of Valencia, Spain
[2] Faculty of Computer Science,
B. Autonomous University of Puebla, Mexico
{dpinto, ajuan, prosso}@dsic.upv.es

Abstract. The world wide web is a natural setting for cross-lingual information retrieval. The European Union is a typical example of a multilingual scenario, where multiple users have to deal with information published in at least 20 languages. Given queries in some source language and a target corpus in another language, the typical approximation consists in translating either the query or the target dataset to the other language. Other approaches use parallel corpora to obtain a statistical dictionary of words among the different languages. In this work, we propose to use a training corpus made up by a set of Query-Relevant Document Pairs (QRDP) in a probabilistic cross-lingual information retrieval approach which is based on the IBM alignment model 1 for statistical machine translation. Our approach has two main advantages over those that use direct translation and parallel corpora: we will not obtain a translation of the query, but a set of associated words which share their meaning in some way and, therefore, the obtained dictionary is, in a broad sense, more semantic than a translation one. Besides, since the queries are supervised, we are working in a more restricted domain than that when using a general parallel corpus (it is well known that in this context results are better than those which are performed in a general context). In order to determine the quality of our experiments, we compared the results with those obtained by a direct translation of the queries with a query translation system, observing promising results.

1 Introduction

The fast growth of the Internet and the increasing multilinguality of the web poses an additional challenge for language technology. Therefore, the development of novel techniques for managing of data, especially when we deal with information in multiple languages, is needed. There are sufficient examples in which users may be interested in information which is in a language other than their own native language. A common language scenario is where a user has some comprehension ability for a given language but s/he is not sufficiently proficient to confidently specify a search request in that language. Thus, a search engine that may deal with this cross-lingual problem should be of a high benefit.

* This work has been partially supported by the MCyT TIN2006-15265-C06-04 research project, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant.

V. Matousek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 630–637, 2007.

In Cross-Language Information Retrieval (CLIR), the usual approach is to first translate the query into the target language and then retrieve documents in this language by using a conventional, monolingual information retrieval system. The translation system might be of any type (rule-based, statistical, hybrid, etc.). For instance, in [1] and [2], a statistical machine translation system is used, but it had to be previously trained from parallel texts. See [3], [4], and [5] for a survey on CLIR.

Since our perspective, the above two-step approach is too sensitive to translation errors produced during the first step. In fact, even if we have a very accurate retrieval system, translation errors prevent correct retrieval of relevant documents. To overcome this drawback, we propose to use a set of queries with their respective set of relevant documents as an input training set for a direct probabilistic cross-lingual information retrieval system which integrates both steps into a single one. This is done on the basis of the IBM alignment model 1 (IBM-1) for statistical machine translation [6]. Probabilistic approaches which use parallel corpora in order to translate the input queries by means of a statistical dictionary in CLIR have been used from many years ago (see [2]). However, our aim is *not* to translate queries but to obtain a set of associated words for a given query. Therefore, a parallel corpus does not have sense for our purpose, since we need to find a possible set of relevant documents for each query given. To our knowledge, this novel approach has not been presented earlier in literature.

We carried out some experiments by using a subset of the EuroGOV corpus [7] which was first used in the bilingual English to Spanish subtask of WebCLEF 2005 [8]. A document indexing reduction was also proposed in order to improve precision of our approach and to diminish its storing space. The corpus reduction was based on the use of a technique for selecting mid-frequency terms, named the Transition Point (TP), which was used in other research works with the same purpose [9,10]. We evaluated four different percentages of TP observing that it is possible to improve precision by reducing the number of terms for a given corpus.

Section 2 and 3 describe the query-relevant document pairs model in detail. Section 4 introduces the corpus used in the experiments, and explains the way we implemented the reduction process. The results obtained after the evaluation are illustrated in Section 5 and discussed in Section 6.

2 The QRDP Probabilistic Model

Lex x be a query text in a certain *input (source)* language, and let $y_1, y_2, \cdots, y_W$ be a collection of W web pages in a different *output (target)* language. Let $\mathcal{X}$ and $\mathcal{Y}$ be their associated input and output vocabularies, respectively. Given a number $k < W$, we have to find the k most relevant web pages with respect to the input query x. To do this, we have followed a probabilistic approach in which the k most relevant web pages are computed as those most probable given x, i.e.,

$$\{y_1^*(x), \cdots, y_k^*(x)\} = \underset{\substack{S \subset \{y_1, \cdots, y_W\} \\ |S|=k}}{\operatorname{argmax}} \min_{y \in S} p(y \mid x) \tag{1}$$

In the particular case of k=1, Equation (1) is simplified to

$$y_1^*(x) = \underset{y=y_1,\cdots,y_W}{\operatorname{argmax}} \; p(y \mid x) \tag{2}$$

In this work, $p(y \mid x)$ is modelled by using the well-known IBM alignment model 1 (IBM-1) for statistical machine translation [6,11]. This model assumes that each word in the web page is *connected to exactly one word* in the query. Also, it is assumed that the query has an initial "null" word to which words in the web page with no direct connexion are linked. Formally, a hidden variable $a = a_1 a_2 \cdots a_{|y|}$ is introduced to reveal, for each position i in the web page, the query word position $a_i \in \{0, 1, \ldots, |x|\}$ to which it is connected. Thus,

$$p(y \mid x) = \sum_{a \in \mathcal{A}(x,y)} p(y, a \mid x) \tag{3}$$

where $\mathcal{A}(x, y)$ denotes the set of all possible alignments between x and y. The *alignment-completed* probability $p(y, a \mid x)$ can be decomposed in terms of individual, web page position-dependent probabilities as:

$$p(y, a \mid x) = \prod_{i=1}^{|y|} p(y_i, a_i \mid a_1^{i-1}, y_1^{i-1}, x) \tag{4}$$

$$= \prod_{i=1}^{|y|} p(a_i \mid a_1^{i-1}, y_1^{i-1}, x) p(y_i \mid a_1^{i}, y_1^{i-1}, x) \tag{5}$$

In the case of the IBM-1 model, it is assumed that a_i is uniformly distributed

$$p(a_i \mid a_1^{i-1}, y_1^{i-1}, x) = \frac{1}{|x| + 1} \tag{6}$$

and that y_i only depends on the query word to which it is connected

$$p(y_i \mid a_1^{i}, y_1^{i-1}, x) = p(y_i \mid x_{a_i}) \tag{7}$$

By sustitution of (6) and (7) in (5); and thereafter (5) in (3), we may write the IBM-1 model as follows by some straighforward manipulations:

$$p(y \mid x) = \sum_{a \in \mathcal{A}(x,y)} \prod_{i=1}^{|y|} \frac{1}{(|x| + 1)} p(y_i \mid x_{a_i}) \tag{8}$$

$$= \frac{1}{(|x| + 1)^{|y|}} \prod_{i=1}^{|y|} \sum_{j=0}^{|x|} p(y_i \mid x_j) \tag{9}$$

Note that this model is governed only by a *statistical dictionary* $\boldsymbol{\Theta}$=$\{p(w|v)$, for all $v \in \mathcal{X}$ and $w \in \mathcal{Y}\}$. The model assumes that the order of the words in the query is not important. Therefore, each position in a document is equally likely to be connected to each position in the query. Although this assumption is unrealistic in machine translation, we consider the IBM-1 model is particularly well-suited for our approach.

3 Maximum Likehood Estimation

It is not difficult to derive an Expectation-Maximisation (EM) algorithm to perform maximum likelihood estimation of the statistical dictionary with respect to a collection of training samples $(X, Y) = \{(x_1, y_1), \ldots, (x_N, y_N)\}$. The *(incomplete)* log-likelihood function is:

$$L(\boldsymbol{\Theta}) = \sum_{n=1}^{N} \log \sum_{a_n} p(y_n, a_n | x_n) \tag{10}$$

with

$$p(y_n, a_n | x_n) = \frac{1}{(|x_n| + 1)^{|y_n|}} \prod_{i=1}^{|y_n|} \prod_{j=0}^{|x_n|} p(y_{ni} \mid x_{nj})^{a_{nij}} \tag{11}$$

where, for convenience, the alignment variable, $a_{ni} \in \{0, 1, \ldots, |x_n|\}$, has been rewritten as an indicator vector in (11), $a_{ni} = (a_{ni0}, \ldots, a_{ni|x_n|})$, with 1 in the query position to which it is connected, and zeros elsewhere.

The so-called *complete* version of the log-likelihood function (10) assumes that the hidden (missing) alignments $a_1, \ldots, a_N$ are also known:

$$\mathcal{L}(\boldsymbol{\Theta}) = \sum_{n=1}^{N} \log p(y_n, a_n | x_n) \tag{12}$$

The EM algorithm maximises (10) iteratively, through the application of two basic steps in each iteration: the E(xpectation) step and the M(aximisation) step. At iteration k, the E step computes the expected value of (12) given the observed (incomplete) data, (X, Y), and a current estimation of the parameters, $\boldsymbol{\Theta}^{(k)}$. This reduces to the computation of the expected value of a_{nij}:

$$a_{nij}^{(k)} = \frac{p(y_{ni} \mid x_{nj})^{(k)}}{\sum_{j'} p(y_{ni} \mid x_{nj'})^{(k)}} \tag{13}$$

Then, the M step finds a new estimate of $\boldsymbol{\Theta}$, $\boldsymbol{\Theta}^{(k+1)}$, by maximising (12), using (13) instead of the missing a_{nji}. This results in:

$$P(v|w)^{(k+1)} = \frac{\sum_n \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} a_{nij}^{(k)} \, \delta(y_{ni}, w) \, \delta(x_{nj}, v)}{\sum_{w'} \sum_n \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} a_{nij}^{(k)} \, \delta(y_{ni}, w') \, \delta(x_{nj}, v)} \tag{14}$$

$$= \frac{\sum_n \frac{p(w \mid v)^{(k)}}{\sum_{j'} p(w \mid x_{nj'})^{(k)}} \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} \delta(y_{ni}, w) \, \delta(x_{nj}, v)}{\sum_{w'} \left[\sum_n \frac{p(w' \mid v)^{(k)}}{\sum_{j'} p(w' \mid x_{nj'})^{(k)}} \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} \delta(y_{ni}, w') \, \delta(x_{nj}, v) \right]} \tag{15}$$

for all $v \in \mathcal{X}$ and $w \in \mathcal{Y}$; where $\delta(a, b)$ is the Kronecker delta function; i.e., $\delta(a, b) = 1$ if $a = b$; 0 otherwise.

An initial estimate for $\boldsymbol{\Theta}$, $\boldsymbol{\Theta}^{(0)}$, is required for the EM algorithm to start. This can be done by assuming that the translation probabilities are uniformly distributed; i.e.,

$$p(w \mid v)^{(0)} = \frac{1}{|\mathcal{Y}|} \tag{16}$$

for all $v \in \mathcal{X}$ and $w \in \mathcal{Y}$.

4 The EuroGOV Corpus

We have used a subset of the EuroGOV corpus for the evaluation of the QRDP model. This subset was made up by a set of Spanish Internet pages, originally obtained from European government-related sites and particularly used in the WebCLEF track of the Cross-Language Evaluation Forum[1] (CLEF) [8]. A better reference to this corpus can be seen in [7].

We refined the evaluation corpus, with those documents automatically identified as in the "Spanish" language, by using the TexCat language identification program [2]. For the evaluation of this corpus, a set of 134 supervised queries in the "English" language was used. The pre-processing step was applied to both, the web pages and the queries, and consisted of the elimination of punctuation symbols, Spanish and English stopwords, numbers, html tags, script codes and cascading style sheets codes.

For convenience, we built a training corpus comprising pairs of query and target web page. We observed that a possible improvement in time indexing and search engine precision may be obtained by reducing the size of this corpus. Therefore, we applied a term selection technique, named transition point, in order to obtain only the mid-frequency terms which will represent every document (see [9] and [10] for further details).

For this purpose, a term frequency value of the web page vocabulary is selected as the transition point, and then a neighbourhood of TP is used as threshold for determining those terms which will be selected. After using four different thresholds (10%, 20%, 40%, and 60%), we obtained five corpora for the evaluation. Table 1 shows the size of every test corpus used, as well as the percentage of reduction obtained for each of them. As can be seen, the TP technique obtained a high percentage of reduction (between 75 and 89%), which also implied a time reduction for constructing the statistical dictionary.

Table 1. Test corpora

Corpus	Size (≈Kb)	Reduction (%)
Full	117	0
TP60	29	75.37
TP40	20	82.55
TP20	19	83.25
TP10	13	89.25

[1] `http://www.clef-campaign.org/`

[2] `http://www.let.rug.nl/~vannoord/TextCat/`

5 Evaluation of the Results

In the experiments, we used the leave-one-out procedure which is a standard procedure in predicting the generalisation power of a classifier, both from a theoretical and empirical perspective [12].

Table 2 shows the results for every run executed by applying only 10 iterations in the EM algorithm. The first column indicates the name of the run carried out for each corpus. The last column shows the Mean Reciprocal Rank (MRR) obtained for each run. Additionally, the Average Success At (ASA) different number of documents retrieved is shown. As can be seen, an improvement by using an evaluation corpus was obtained employing the TP technique with a neighbourhood of 40%, which is exactly the same percentage used in other research works (see [10] and [13]). We consider that this improvement is derived from the elimination of noisy words, which helps to rank better the web pages.

Table 2. Evaluation results

	ASA					
Run	**1**	**5**	**10**	**20**	**50**	**MRR**
FULL	0.0000	0.0299	0.0970	0.2687	0.3955	0.0361
TP10	0.0149	0.0522	0.0672	0.0970	0.4030	0.0393
TP20	0.0149	0.0299	0.0448	0.0746	0.4030	0.0323
TP40	0.0149	0.0448	0.1045	0.1940	0.3881	**0.0470**
TP60	0.0000	0.0448	0.1269	0.2164	0.4030	0.0383

Three teams participated at the bilingual "English to Spanish" subtask at WebCLEF in 2005. Every team submitted at least one run [14,10,15]. A comparison among the results obtained by each team and our best results can be seen in Table 3. In this case, we are presenting the results obtained with the TP40 corpus and by applying 100 iterations in the EM algorithm. Each of these teams translated each query from English to Spanish and thereafter they used a traditional monolingual information retrieval system for carrying out the searching process. Particularly, the UNED team reported two results (UNED_FULL and UNED_BODY) which are related with the information of each web page used; their first aproximation makes use of information stored in html fields or tags identified during the preprocessing, like *title*, *metadata*, *heading*, *body*, *outgoing links*. Their second aproximation (UNED_BODY) only considered the information in the *body* field. We also considered only the information inside the *body* html tag and, therefore, the UNED_BODY run can be used for comparison. On the other hand, the ALICANTE's team has used a combination of three translation systems for obtaining the best translation of a query. Thereafter, they used a passage retrieval-based system as a search engine, indexing in the documents all the information except html tags.

We may observe that by using the same information from a web page, we have slightly outperformed the results obtained by other approaches, even when we have trained our model with only 3 target web pages in average per query, and executing 100 iterations on the Expectation-Maximization model.

Table 3. Comparison results over 134 topics

	ASA					
Run name	**1**	**5**	**10**	**20**	**50**	**MRR**
OurApproach	0.0672	0.1045	0.1418	0.2164	0.4403	**0.0963**
UNED_FULL	0.0821	0.1045	0.1194	0.1343	0.2090	0.0930
BUAP/UPV40	0.0597	0.0970	0.1119	0.1418	0.2164	0.0844
UNED_BODY	0.0224	0.0672	0.1045	0.1716	0.2612	0.0477
BUAP/UPVFull	0.0224	0.0672	0.1119	0.1418	0.1866	0.0465
ALICANTE	0.0299	0.0522	0.0597	0.0746	0.0970	0.0395

6 Conclusions

We have described a query-relevant document pairs based model for cross-language information retrieval. The QRDP model uses a statistical dictionary of associated words directly to rank documents according to their relevance with respect to the query. We consider that inaccuracies of query translation have a negative effect on document retrieval and, therefore, using the probabilistic values of association should help to overcome this problem.

The application of statistical machine translation for CLIR may be often seen in literature, but what we proposed in this paper is to study the derivation of the translation (association) dictionary from query-relevant document pairs. The probabilistic model assumes that the order of the words in the query is not important. Therefore, each position in a document is equally likely to be connected to each position in the query. Although this assumption is unrealistic in machine translation, we consider the IBM-1 model to be particularly well-suited for our approach.

We have used a term selection technique in order to reduce the size of the training corpus with good findings. For instance, by using a 82.5% of reduction, the results can improve those of using the complete corpus.

Last but not least, we would emphasize that the QRDP probabilistic model is language independent and, therefore, it can be employed to model cross-language query-document pairs in any language.

References

1. Franz, M., McCarley, J.S., Roukos, S.: Ad-hoc and multilingual information retrieval at ibm. In: Proceedings of the TREC-7 Conference, pp. 157–168 (1998)
2. Kraaij, W., Nie, J.Y., Simard, M.: Embedding web-based statistical translation models in cross-language information retrieval. Computational Linguistics 29(3), 381–419 (2003)
3. Fuhr, N.: Probabilistic models in information retrieval. The Computer Journal 35(3), 243–255 (1992)
4. Rijsbergen, C.J.V.: Information Retrieval, 2nd edn. Dept. of Computer Science, University of Glasgow (1979)
5. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, New York, Addison-Wesley (1999)

6. Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Computational Linguistics 16(2), 79–85 (1990)
7. Sigurbjörnsson, B., Kamps, J., de Rijke, M.: Eurogov: Engineering a multilingual web corpus. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 825–836. Springer, Heidelberg (2006)
8. Sigurbjörnsson, B., Kamps, J., de Rijke, M.: Overview of webclef 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 810–824. Springer, Heidelberg (2006)
9. Pinto, D., Jiménez-Salazar, H., Rosso, P.: Clustering abstracts of scientific texts using the transition point technique. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 536–546. Springer, Heidelberg (2006)
10. Pinto, D., Jiménez-Salazar, H., Rosso, P.: Buap-upv tpirs: A system for document indexing reduction on webclef. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 873–879. Springer, Heidelberg (2006)
11. Civera, J., Juan, A.: Mixtures of ibm model 2. In: Proceedings of the EAMT Conference, pp. 159–167 (2006)
12. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1995)
13. Rojas-López, F., Jiménez-Salazar, H., Pinto, D.: A competitive term selection method for information retrieval. In: Gelbukh, A. (ed.) CICLing 2007. LNCS, vol. 4394, pp. 468–475. Springer, Heidelberg (2007)
14. Artile, J., Peinado, V., Peñas, A., Verdejo, F.: Uned at webclef 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 888–891. Springer, Heidelberg (2006)
15. Martínez, T., Noguera, E., noz, R.M., Llopis, F.: University of alicante at the clef2005 webclef track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 865–868. Springer, Heidelberg (2006)