Estimating Translation Probabilities from the Web for Structured Queries on CLIR

Xabier Saralegi and Maddalen Lopez de Lacalle

Elhuyar Foundation, R & D,
20170 Usurbil, Spain
{xabiers,maddalen}@elhuyar.com

Abstract. We present two methods for estimating replacement probabilities without using parallel corpora. The first method proposed exploits the possible translation probabilities latent in Machine Readable Dictionaries (MRD). The second method is more robust, and exploits context similarity-based techniques in order to estimate word translation probabilities using the Internet as a bilingual comparable corpus. The experiments show a statistically significant improvement over non weighted structured queries in terms of MAP by using the replacement probabilities obtained with the proposed methods. The context similarity-based method is the one that yields the most significant improvement.

Keywords: Cross-lingual Information Retrieval, Structured Query Translation, Web as Corpus.

1 Introduction

Several techniques have been proposed for dealing with translation ambiguity for the query translation task on CLIR, such as structured query-based translation (also known as Pirkola's method) [1], word co-occurrence statistics [2] and statistical translation models [3]. Structured queries are adequate for less resourced languages, rare pairs of languages or certain domains where parallel corpora are scarce or even nonexistent. The idea behind this method is to treat all the translation candidates of a source word as a single word (*syn* operator) when calculating TF and DF statistics. This produces an implicit translation selection during retrieval time. There are many works dealing with structured queries, and some variants are proposed. [4] for example, proposes that weights or replacement probabilities be included in the translation candidates (*wsyn* operator). One drawback with this approach is that it needs parallel corpora in order to estimate the replacement probabilities.

Following this line of work, we propose a simple method based on the implicit translation probabilities of a dictionary, and also a more robust one which uses translation knowledge mined from the web. We have analyzed different ways of accessing web data: **Web As Corpus** tools, **News** search engines, and **Blog** search engines. Our aim is to examine how the characteristics of each access strategy influence the representation of the constructed contexts, and also, how far these strategies are adequate

C. Gurrin et al. (Eds.): ECIR 2010, LNCS 5993, pp. 586–589, 2010.

for estimating translation probabilities by means of the cross-lingual context similarity paradigm. All experiments have been carried out taking Spanish as source language and English as target.

2 Obtaining Translation Probabilities from a Dictionary

The first method proposed for estimating translation probabilities relies on the hypothesis that, in a bilingual MRD (*D*), the position (*pos*) of the translation (*w*) among all the corresponding translation candidates *f* for a source word (*v*) is inversely proportional to its translation probability ($p(w|v)$). If we assume that it is an exponential decay relation, we can model the translation probability through this formula:

$$p(w|v)=1/\sum_{(v,f)\in D}\left(\frac{1}{pos(D,v,f)}\right)\cdot pos(D,v,w) \tag{1}$$

The principal problems of these assumptions are, firstly, that translations are not ordered in all MRD (partially or at all) by frequency of use, and secondly, that the proposed relation above does not fit all translation equivalents. So, we propose a method that is useful for ordering the translations of an MRD as well as for estimating more accurate translation probabilities, as presented in the following section.

3 Translation Probabilities by Context Similarity

The idea is to obtain translation probabilities by using the web as a bilingual comparable corpus. This strategy is based on estimating the translation probability of the translation candidates taken from the MRD in accordance with the context similarity of the translation pairs [5]. The hypothesis is that the more similar the contexts are, the more probable the translation will be. The computation of the context similarity requires a large amount of data (contexts of words), which has to be representative and from comparable sources. The Internet is a good source of large amounts of texts, and that is why, we propose that different search-engines be analyzed to obtain these contexts. These search engines have different features, such as domain, coverage and ranking, which affect both the degree of comparability and the representativeness of the contexts, as follows:

WebCorp: This Web Concordancer is based on main search APIs. Therefore, navigational queries and popular ones are promoted. These criteria can reduce the representativeness of the contexts retrieved. Since we take a maximum number of snippets for each query, the selected contexts depend on the ranking algorithm. It guarantees good recall, but perhaps poor precision. Thus, the comparability degree between contexts in different languages can be affected negatively.

Google News Archive: The content is only journalistic. It seems appropriate if we want to deal with journalism documents but not with other registers or more specialized domains. In short, it offers good precision, enough recall and a good degree of comparability.

Google Blog search: The language used is more popular, and although the register is similar to that of journalism, the domain is more extensive. This could offer good recall but not very comparable contexts.

The method to estimate the translation probabilities between a source word (*v*) and its translations f ($(v,f)\in D$) starts by downloading, separately, the snippets of both words as returned by the search engines mentioned above. Then, we set up context vectors for the source $\vec{v}$ and the translation word $\vec{w}$ by taking keywordness (using log-likelihood scores) of the content words (nouns, adjectives, verbs and adverbs selected by using *Treetagger*) belonging to all their snippets. The next step is to translate the Spanish context vector $\vec{v}$ into English $\overrightarrow{tr(v)}$. This is done by taking the first translation from a Spanish-English MRD (*D*) (34,167 entries). Cross-lingual context similarity is calculated according to cosine measure which is transformed into translations probabilities:

$$p(w|v)=\frac{\cos\left(\overrightarrow{tr(v)},\vec{w}\right)}{\sum_{(v,f)\in D}\left(\cos\left(\overrightarrow{tr(v)},\vec{f}\right)\right)} \tag{2}$$

We analyze the differences between the translation rankings obtained with the different search engines and those in the original dictionary. We computed Pearson's correlation for the translation rankings obtained for the polysemous content words in all 300-350 Spanish CLEF topics. The correlation scores (cf. Table 1) show that the different characteristics of each search engine produce translation rankings which are quite different from those in the dictionary (**Dic.**) and also from each other.

Table 1. Mean of Pearson's correlation coefficients for translation rankings compared to each other

	WebCorp	**News**	**Blog**
Dic.	0.42	0.31	0.40
WebCorp		0.44	0.54
News			0.49

4 Evaluation and Conclusions

We evaluated 50 queries (title+description) taken from 300-350 CLEF topics against collections from CLEF 2001 composed by LA Times 94 and Glasgow Herald 95 news. Previously, nouns, adjectives, verbs and adverbs were selected manually both in Spanish and English topics. *Indri* was used as the retrieval model and the queries were translated using several methods: taking the first translation of the MRD (**First**); taking all the translations and grouping them by the *syn* operator (**All or Pirkola**); and weighting the translations by using the *wsyn* operator and the methods described in sections 2 (**Dic**.) and 3 (**Webcorp**, **News** and **Blog**). The results are shown in Table 2.

In the first column we show the MAP results obtained with each method, with the English monolingual results first. In the second column we show the percentage of the

Table 2. MAP for 300-350 topics

Method	MAP	% Monolingual	% Improv. over All
Monolingual (en)	0.3651		
First	0.2462	67.43	
All	0.2892	79.21	
Dic.	0.2951	80.83	2.04
WebCorp	0.2943	80.55	1.76
News	0.2993	82.63	3.49
Blog	0.2960	81.07	2.35

cross lingual MAP with respect to the monolingual result. We can see that using all translations with their replacement probability estimated according to the dictionary order produces better results than using only the first translation or using all translations, with a significant improvement (according to the Paired Randomization Test with α=0.05) over the **All** method. So, exploiting the translation knowledge latent in the position of the translations improves the MAP when provided by the dictionary. Otherwise, the web-based estimation techniques also improve significantly over the **First** and **All** strategies (α=0.05). However, there is no significant improvement over the **Dic.** method. It seems that context similarity calculated from **Blog** or **News** sources is more suited to estimating translation probabilities since they significantly outperform **WebCorp** in terms of MAP. Therefore, comparability between sources of both languages, domain precision and informational snippets seem to be important factors in order to obtain useful context for context-similarity, although deeper analyses must be carried out to determine the importance of each more precisely. Finally, we conclude that translation knowledge obtained from the Internet, offers an adequate means, and by means of cross-lingual context similarity, it is useful for estimating replacement probabilities. Moreover, it could be an alternative when parallel corpora or MRDs with translations sorted according frequency of use are not available.

References

1. Pirkola, A.: The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. In: SIGIR 1998, pp. 55–63 (1998)
2. Ballesteros, L., Croft, W.B.: Resolving Ambiguity for Cross-Language Retrieval. In: SIGIR 1998, pp. 64–71 (1998)
3. Hiemstra, D., De Jong, F.: Statistical Language Models and Information Retrieval: natural language processing really meets retrieval. University of Twente (2001)
4. Darwish, K., Oard, D.W.: Probabilistic structured Query Methods. In: SIGIR 2003 (2003)
5. Fung, P., Yuen Yee, L.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In: COLING-ACL (1998)