Mining Term Translations from Domain Restricted Comparable Corpora

Extracción de Traducciones de Términos a partir de Corpus Comparables pertenecientes a áreas específicas

Xabier Saralegi
Elhuyar R&D
Zelai Haundi kalea, 3
20170 Usurbil.
xabiers@elhuyar.com

Iñaki San Vicente
Elhuyar R&D
Zelai Haundi kalea, 3
20170 Usurbil.
inaki@elhuyar.com

Maddalen López de Lacalle
Elhuyar R&D
Zelai Haundi kalea, 3
20170 Usurbil.
maddalen@elhuyar.com

Abstract: Several approaches have been proposed in the literature for extracting word translations from comparable corpora, almost all of them based on the idea of context similarity. This work addresses the aforementioned issue for the Basque-Spanish pair in a popular science domain. The main tasks our experiments focus on include: designing a method to combine some of the existing approaches; adapting this method to a popular science domain for the Basque-Spanish pair; and analyzing the performance of different approaches both for translating the contexts of the words and computing the similarity between contexts. We finally evaluate the different prototypes by calculating the precision for different cutoffs. The yielded results show the validity of the designed hybrid method, as well as the improvement obtained by using the probabilistic models we propose for computing the similarity between contexts.
Keywords: Bilingual Terminology Extraction, Comparable Corpora, Machine Translation.

Resumen: En la literatura se han propuesto diferentes estrategias para la tarea de extracción automática de traducciones a partir de corpus comparables, estando basadas la mayoría de ellas en la idea de similitud entre contextos. Este trabajo aborda la citada tarea para el par de lenguas Euskera-Castellano y el género científico-divulgativo. Los principales puntos en los que se centra este trabajo son los siguientes: diseñar un método que combine las existentes aproximaciones; adaptar este método al par de lenguas Euskera-Castellano y al género científico-divulgativo; y por último analizar el comportamiento de distintas técnicas tanto para el proceso de traducción de contextos como el cálculo de similitud entre ellos. Finalmente, evaluaremos los diferentes prototipos implementados de acuerdo a la precisión obtenida para distintos cutoffs. Los resultados obtenidos muestran que el método híbrido diseñado resulta adecuado y una mejora para el cálculo de similitudes entre contextos mediante los modelos probabilísticos propuestos.
Palabras clave: Extracción de Terminología Bilingüe, Corpus Comparables, Traducción Automática.

1 Introduction

In the literature, several strategies have been proposed for extracting lexical equivalences from corpora. Most of them are designed to be used with parallel corpora. Although these kinds of corpora give the best results, they are a scarce resource, especially when we want to deal with certain language pairs and certain domains and genres. To overcome this limitation the first algorithms (Rapp, 1995), (Fung, 1995) were developed for automatic extraction of translation pairs from comparable corpora. These kinds of corpora can be easily built from the Internet.

The techniques proposed for the extraction task are mainly based on the idea that translation equivalents tend to co-occur within similar contexts. An alternative is to detect translation equivalents by means of string similarity (cognates). Nevertheless, none of these techniques achieve the precision and recall obtained with the parallel corpora techniques.

This work focuses on the Basque-Spanish pair and popular-science domain. We channeled our efforts towards designing a hybrid approach by combining the methods proposed in the literature, adapting it to the scenario, and analyzing the performance of different strategies for the two main steps of the extraction approach based on context similarity: translation of the context of the source word to the target language, and calculation of the similarity between contexts. On the one hand we have compared a number of methods for resolving the two main problems in this first phase, which are translation selection and treatment of Out of Vocabulary (OOV) words. On the other, we have tested different models for representing contexts and different ranking algorithms to calculate the similarity between contexts.

Finally it must be said that this work is the continuation of the research started in (Saralegi, San Vicente and Gurrutxaga, 2008), focusing on the Basque-English pair.

2 Comparable Corpora

Comparable multilingual corpora are defined as collections of documents sharing certain characteristics and written in more than one language. In bilingual lexicon extraction some of these characteristics depend on the lexicon type we aim to extract. Thus, achieving a high degree of comparability with regard to these characteristics is very important, since context similarity techniques will be more effective. The more similar the corpora are, the higher the comparability between the collocated words of the equivalent translations (Morin et al., 2007). Therefore, it is essential to ensure that some characteristics are equal in both parts of the corpora built for terminology extraction purposes.

3 Identification of Equivalents

3.1 Context Similarity

The main method is based on the idea that the same concept tends to appear with the same context words in both languages, that is, it maintains many collocates. The methods based on context similarity consist of two steps: modeling of the contexts, and calculation of the degree of similarity using a seed bilingual lexicon (Rapp, 1999), (Fung, 1998).

The majority of the methods for modeling are based on the "bag-of-words" paradigm. Thus, the contexts are represented by weighted collections of words. In fact, the context similarity calculation tasks can be seen as a Cross Language Information Retrieval (CLIR) problem. Therefore, all the paradigms proposed in the CLIR literature can be useful in this context. There are several techniques for determining which words make up the context of a word: distance-based window, syntactic based-window, etc.

Different models have been proposed to represent the context of words. The most widely used combines the Vector Space Model and Association Measures (AM) for establishing the weight of the context words with regard to a word: Log-likelihood ratio (LLR), Mutual Information, Dice coefficient, Jaccard measure, frequency, tf-idf, etc. After representing word contexts in both languages, the proposed algorithms compute for each source word a ranking of translation candidates according to the similarity between its context vector and the context vectors of all the target words. The similarity score is computed by means of measures such as Cosine, Jaccard or Dice.

Nevertheless, the number of works that exploit the recent advances obtained in the CLIR community is limited, in particular works involving translation selection techniques and probabilistic models. (Shao et al., 2004) can be an example of the use of probabilistic models. It represents the contexts by using language models. Other probabilistic retrieval-models proposed for IR tasks, which can also be of use in context similarity calculation, are Okapi (Robertson, Walker and Beaulieu, 1998) or Divergence From Randomness (DFR) (Amati and Van Rijsbergen, 2002). Okapi (BM25) represents the state of the art in IR and is often used as baseline. The DFR paradigm is, like

Okapi, a generalization of the Harter's 2-Poisson (Harter, 1974) indexing-model which offers different models. The Terrier[1] toolkit offers many of these DFR models as well as others, such as tf-idf, Okapi and language models.

3.2 Context words translation

To be able to compute the similarity, the context vectors are put in the same space by translating one of them. The methods proposed in the literature for the translation in CLIR tasks can be divided into two main groups (Hull, 1997): corpus-based methods and dictionary-based methods. Corpus-based methods use parallel and sometimes comparable corpora for mining query translations. Unfortunately, parallel corpora constitute a scarce resource and the results obtained using comparable corpora are still poor. On the other hand, dictionary-based methods use a bilingual dictionary to lookup the translations of the components of the query. However, the dictionary poses two main problems: it fails to solve the ambiguous translations and it has a coverage problem (OOV).

3.3 Translation selection

Many algorithms have been proposed for dealing with the translation disambiguation resulting from query translation guided by bilingual dictionaries. The simplest method is to select the first translation given by the dictionary as the best since the translations are often sorted by use frequency. However this approach fails to take into account the domain of the query, so the disambiguation can be very rigid. Other more flexible approaches (Pirkola, 1998), which perform better, take all the translations and group them as a unique word when the TF and DF values of the document words are calculated by the ranking method. The syn operator offered by the Indri and Inquery query languages allows this type of grouping (Pirkola, 1998). Other more complex approaches (Ballesteros and Croft, 1998) (Liu, Jin, and Chai, 2005) (Chen, Bian, and Lin, 1999) (Gao and Nie, 2006), which also use statistical information of monolingual word concurrences, are those based on the degree of cohesion or association between the translation candidates. They try to obtain the combination of translation candidates that maximize the coherence between them. A corpus in the target language is used to compute association scores.

3.4 Cognates

Another technique proposed in the literature, also useful for the treatment of OOV, is the identification of translations by means of cognates (Al-Onaizan and Knight, 2002). This method could be appropriate in a science domain where the presence of cognates is high. In fact, using a Basque-Spanish technical dictionary we were able to calculate automatically that around 26% of the translation pairs were cognates. Dice coefficient and LCSR (Longest Common Subsequence Ratio) measures are proposed for computing string similarity.

4 Experiments

4.1 Term Extraction from Comparable Corpora

4.1.1 Preprocess

We needed to identify the words we considered to be meaningful for our process, that is, content-words. POS tags were used for this task. Treetagger[2] is the tagger we chose to tag the Spanish corpus, and Eustagger[3] in the case of the Basque corpus. Only nouns, adjectives and verbs are regarded as content words. In our experiments, adverbs were found to produce noise. Proper nouns also produced noise due to a cultural bias effect. Both were removed.

4.1.2 Contexts Construction

We established a distance-based window to delimit the contexts of the words. The window size was determined empirically: 10 words for Basque (plus and minus 5 around a given word) and 14 for Spanish (plus and minus 7). Furthermore, our experiments showed that using punctuation marks to delimit the window improved the results. So, we also included this technique in our system.

We calculated the weight of the words within the context by means of absolute frequency, LLR, Dice coefficient or Jaccard measure, and then the contexts were modeled in a vector space. The best results were achieved by using the LLR.

[1] http://ir.dcs.gla.ac.uk/terrier/

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

[3] A POS tagger for Basque developed by the IXA group of the University of the Basque Country.

We also modeled the contexts of words by using different probabilistic models offered by the Terrier Toolkit. Specifically, we carried out tests with two models, Okapi and PL2, which is an instantiation of the DFR framework appropriate for tasks that require high precision. We indexed the context words of a word like a document. That is, the words that make up a context of a word throughout the collection are included in the same document that is indexed.

4.1.3 Context Translation

To compute the translation of a Basque word, we translated its contexts in order to make them comparable with Spanish contexts. A bilingual Machine Readable Dictionary (MRD) was used for this purpose.

4.1.4 OOV words

The recall of the MRD determines the representativity of the context vector. In our experiments with a general dictionary, the average translation recall by vector was 55%. The higher the recall the greater the possibilities of finding the right translation for a word, because context vectors held more detailed information about the word in question.

To increase the recall of our translated vectors, we try to find equivalents not included in the dictionary by means of cognates. For all the OOV words, we looked for cognates among all the context words in the target language. The identification of these cognates is made by calculating the LCSR between the Basque and Spanish context words. Before applying the LCSR, we processed some typographic rules to normalize equal phonology n-grams (e.g., c→k acta=akta) or regular transformation ones (e.g., -ción→-zio, acción=akzio) in both equivalent candidates. The candidates that exceeded the empirically determined threshold of 0.8 were taken as translations.

4.1.5 Translation selection methods

One of the problems of using bilingual dictionaries to translate contexts is that many translation candidates from the dictionary are obtained. This fact causes many problems for the subsequent calculation of the similarity between contexts. Incorrect translations distort the modeling of the context, and hence disfigure the semantic lexical representation and the distribution of the context words. Therefore, techniques for choosing the correct translations can help in this task. In this work we propose two techniques used in CLIR systems that do not need the use of parallel corpora:

First translation: The first translation of an entry is usually the most probable. Although this fact can vary depending on the domain, taking the first translation is a general translation selection method.

Concurrences-based translation: The best translations are selected by using a concurrences-based method. The basic idea is that the degree of association between the correct translations is higher than between other translations. The algorithm tries to obtain the combination of translation candidates that maximizes that degree of association. The algorithm we use to obtain that combination is a greedy one because of the np-hard nature of the process. Some independence assumptions between translation candidates are adopted. Specifically we have used the algorithm proposed by (Gao et al., 2001):

(1) Given a Basque (source language) query $\mathbf{e}=\{e_1,e_2,...,e_n\}$, for each query term e, we define a set of m distinct Spanish translations according to a bilingual dictionary

D: $\mathbf{D}(e_i)=\{c_{i,1},c_{i,2},...,c_{i,m}\}$

(2) For each set $\mathbf{D}(e_i)$:

(a) For each translation $c_{i,j} \in D(e_i)$, define the similarity score between the translation $c_{i,j}$ and a set $\mathbf{D}(e_k)$ ($k \neq i$) as the sum of the similarities between $c_{i,j}$ and each translation in the set $\mathbf{D}(e_k)$ according to Eq. (1)

$$am(c_{i,j}, D(e_k)) = \sum_{c_{k,l} \in D(e_k)} am(c_{i,j}, c_{k,l}) \quad (1)$$

(b) Compute the cohesion score for $c_{i,j}$ as

$$cohesion(c_{i,j} \mid e, D) = \log \sum_{D(e_k)} am(c_{i,j}, D(e_k)) \quad (2)$$

(c) Select the translation $c \in D(e_i)$ with the highest cohesion score

$$c = argmax_{c_{e,j} \in D(e_i)} cohesion(c_{e,j} \mid e, D) \quad (3)$$

We use a Spanish corpus of 10M words obtained from Madri+d to calculate the concurrences for the target collection. We adopted the Mutual Information measure to

calculate the degree of association at document level between translation candidate pairs.

4.1.6 Context Similarity Calculation

To obtain a ranked list of the translation candidates for a Basque word, we calculated the similarity between its translated context vector and the context vectors of the Spanish words by using two different ranking methods. Cosine distance for the case of weighting by LLR, and the aforementioned rank-models for the case of probabilistic models.

Furthermore, to prevent noise candidates in both strategies, after obtaining the rankings, we pruned those that had a different grammatical category from that of the word to be translated.

4.1.7 Equivalent Similarity Calculation

In addition to context similarity, string similarity between source words and equivalent candidates is also used to rank candidates. LCSR is calculated between each source word and its first 100 translation candidates in the ranking obtained after context similarity calculation. LCSR is applied in the same way as in context vector translation.

When used in combination with context similarity, LCSR data is used as the last ranking criterion. The candidates that exceeded an empirically established threshold (0.8) are ranked first, while the position in the ranking of the remaining candidates remains unchanged. A drawback to this method is that cognate translations are promoted over translations based on context vector similarity.

5 Evaluation

5.1 Building Test Corpora

We built one test corpus. The sources of the documents were the science news websites Zientzia.net[4] (Basque), and Madri+d[5] (Spanish).

Zientzia.net and Madri+d are quite similar with respect to the distribution of topics and register, so we chose them to build the test corpus. A correlation between topic and date was expected and for that reason we downloaded all news items between 2000 and 2008, only. Moreover, other types of documents like dossiers, etc. were rejected in order to maintain the same register throughout the corpus. Finally, the HTML documents were cleaned and converted into text using Kimatu (Saralegi and Leturia, 2007). The size of this corpus was 1.092 million tokens for Basque and 1.107 for Spanish. We mapped the different domains in order to compare the distribution of documents among the different domains (table 1). The distribution of the documents among the domains was quite similar, so we expected an acceptable degree of comparability between the two corpora.

Domain	Madri+d	Zientzia.net
Biology, food, Agriculture & fishing	36.59%	24.31%
Health	9.73%	16.26%
Earth sciences	6.12%	10.44%
Physics, Chemistry & Math	6.65%	7.18%
Technology & Industry	29.45%	24.15%
Energy & Environment	11.45%	7.35%

Table 1: Domain distribution of documents for the test corpus.

corpus	#word		#doc	
	eu	es	eu	es
Test corpus	1,092K	1,107K	2521	1242

Table 2: Characteristics of test corpora

5.2 Tests

For the automatic evaluation of our system, we needed a list of Basque-Spanish equivalent terms occurring in each part of the corpora and which were not included in the dictionary used for the translation of content words in the construction of context vectors. To build the list, firstly we took all the Basque content words obtained in the preprocess step for the two corpora, which had been built. Secondly, those words were searched in the Basque-Spanish Elhuyar dictionary[6], and for all the Basque words not included in that dictionary, we randomly selected 200 pairs of words that yielded a minimum frequency (10) and which appeared in one of two terminology Basque-Spanish dictionaries (Elhuyar Science and

[4] http://www.zientzia.net

[5] http://www.madridmasd.org

[6] An abridged version of the Elhuyar Spanish/Basque dictionary including 20,000 entries.

Technology Dictionary[7] and Euskalterm terminology bank[8]).

This enabled us to estimate the precision automatically. We computed, for each source word, the precision of the ranked translation candidates at different cutoff points. We took as correct translation only the one included in the test list as the Spanish translation of the source Basque word. In order to analyze the impact the frequency has on the results, we divided this set into two subsets. The first one includes words of high frequency (>50), and the other one, medium-low frequency words (within the 10-30 frequency range).

We analyzed different variables: the modeling of the contexts, translation methods, and the way to combine the different approaches:

- Modeling of contexts and similarity computation: LLR and cosine, and probabilistic models: Okapi (b=0.75) and PL2 (c=1).
- Translation methods: Cognate detection for treatment of OOV words in the context translation step, first translation selection, and concurrences based selection methods.
- Ranking of translation candidates: context similarity, cognates detection.

5.3 Results

The following tables show the results for the test corpora.

	Mean precision				
	Top 1	Top 5	Top 10	Top 15	Top 20
LLR+cos	0.27	**0.52**	**0.62**	0.65	0.65
Okapi	0.34	0.47	0.60	0.65	0.69
PL2	**0.37**	0.50	0.61	**0.68**	**0.73**

Table 3: Precision results for high frequency test words. Context similarity (cosine+LLR, Okapi, PL2) combined with first translation selection.

[7] Encyclopaedic dictionary of science and technology including 15,000 entries in Basque with equivalences in Spanish, French and English.

[8] Terminological dictionary including 100,000 terms in Basque with equivalences in Spanish, French, English and Latin.

	Mean precision				
	Top 1	Top 5	Top 10	Top 15	Top 20
LLR+cos	**0.07**	0.15	0.17	0.18	0.23
Okapi	0.05	0.12	0.17	0.21	0.23
PL2	0.06	**0.16**	**0.21**	**0.23**	**0.24**

Table 4: Precision results for high frequency test words. Context similarity (cosine+LLR, Okapi, PL2) combined with first translation selection.

	Mean precision				
	Top 1	Top 5	Top 10	Top 15	Top 20
PL2 + First	0.37	0.50	0.61	0.68	0.73
PL2 + Coo	0.37	0.50	0.64	0.68	0.72
PL2 + First + Cog	0.30	0.54	0.59	0.72	0.74
PL2 + Coo + Cog	0.32	0.55	0.67	0.71	0.74
PL2 + Coo + Cog + Cog-re	**0.38**	**0.61**	**0.72**	**0.75**	**0.78**

Table 5: Precision results for high frequency test words. Context similarity (PL2) combined with first translation (First), concurrences based selection (Coo), cognates detection for vector translation (Cog) and re-ranking (Cog-re).

	Mean precision				
	Top 1	Top 5	Top 10	Top 15	Top 20
PL2 + First	0.06	0.16	0.21	0.23	0.24
PL2 + Coo	0.07	0.13	0.19	0.22	0.22
PL2 + First + Cog	0.05	0.16	0.23	0.25	0.26
PL2 + Coo + Cog	0.06	0.18	0.19	0.22	0.25
PL2 + Coo + Cog + Cog-re	**0.28**	**0.40**	**0.39**	**0.46**	**0.45**

Table 6: Precision results for low frequency test words. Context similarity (PL2) combined with first translation (First), concurrences based selection (Co), cognates detection for vector translation (Cog) and re-ranking (Cog-re).

We have observed that combining the identification of cognates in the list of equivalents with context similarity (as proposed in section 4.1.7) improves the precision of the final ranking. The high presence of these kinds of translations explains this improvement.

Otherwise, the results obtained for the low frequency words are poorer than the ones obtained for the high frequency words, as we expected.

The detection of cognates in the translation of the context vectors slightly outperforms translation based exclusively on dictionaries.

The probabilistic models Okapi and PL2 perform much better than the LLR cosine combination for calculating the context similarity. Between Okapi and PL2 the latter is more appropriate.

As for the translation selection methods, there is little difference, but the first translation selection yields better results. This can be due to the short length of the contexts, or to the nature of the context. The contexts used as queries contain fewer specific words than topic queries. This fact could make more difficult the translation selection process.

6 Conclusions

We have performed the first experiments aimed at terminology extraction from comparable corpora by integrating different existing techniques and adapting them for a new language pair.

The combination of the cognate detection in the final ranking as well as in the translation process of the context vectors seems suitable for corpora of the science domain, in which the presence of cognates is high, as we saw for the Basque-English pair.

On the other hand, the concurrences-based algorithm has not improved the quality of the translations achieved with the first translation selection method. This means that selection method adapted to the context sentences is worse than the general selection method. Nonetheless, further experiments will be carried out in order to explore these results in greater depth and to fine-tune the concurrences-based algorithm.

Finally, the representation of contexts and calculation of similarity is improved by using more advanced probabilistic models like Okapi and PL2.

References

Amati G. and C.J. Van Rijsbergen. 2002. "Probabilistic models of information retrieval based on measuring divergence from randomness" In *the Transactions on Information Systems journal,* vol. 20, issue 4*,* pp.357-389.

Al-Onaizan, Y. and K. Knight. 2002. "Machine transliteration of names in Arabic text." In *Proceedings of the ACL-02 workshop on Computational approaches to Semitic languages*, pp.1-13.

Ballesteros, L. and W. B. Croft. 1998. "Resolving ambiguity for cross-language retrieval. In *Proceedings of SIGIR, pp. 64-71.*

Chen, Hsin-Hsi, Guo-Wei Bian, and Wen-Cheng Lin. 1999. "Resolving translation ambiguity and target polysemy in cross-language information retrieval." In *Proceedings of ACL*, pp.215-222.

Déjean, H., E. Gaussier and F. Sadat. 2002. "An Approach Based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction." In *Proceedings of COLING*.

Fung, P. 1995. "Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus." In *Proceedings of the Third Workshop on Very Large Corpora*, pp.173-183.

Fung, P. and L. Y. Yee. 1998. "An IR Approach for Translating New Words from Nonparallel Comparable Texts." In *Proceedings of COLING-ACL* 1998, pp.414-420.

Gao, J. and J. Nie. 2006. "A study of statistical models for query translation: finding a good unit of translation." In *Proceedings of SIGIR*, pp.194-201.

Gao, J., J. Nie, E. Xun, J. Zhang, M. Zhou and C. Huang. 2001. "Improving query translation for cross-language information retrieval using statistical models." In *Proceedings of SIGIR*, pp.96-104.

Kilgarriff, A. and T. Rose. 1998. "Measures for corpus similarity and homogeneity." In *Proceedings of the 3rd EMNLP conference,* pp.46-52.

Hull, D. A. 1997. “Using structured queries for disambiguation in cross-language information retrieval.” In *Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, pp.73-81.

Liu, Y., R. Jin, and J. Y. Chai. 2005. “A maximum coherence model for dictionary-based cross-language information retrieval.” *In Proceedings of SIGIR*, pp.536-543.

Morin, E., B. Daille, K. Takeuchi and K. Kageura. 2007. “Bilingual Terminology Mining - Using Brain, not brawn comparable corpora.” In *Proceedings of ACL*, pp.664-671.

Pirkola, A. 1998. “The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval.” In *Proceedings of SIGIR*, pp.55–63.

Rapp, R. 1995. “Identifying word translations in non-parallel texts.” *In Proceedings of ACL*, pp.320-322.

Rapp, R. 1999. “Automatic identification of word translations from unrelated English and German corpora.” In *Proceedings of ACL*, pp.519-526.

Robertson, S. E., S. Walker and M. Beaulieu. 1998. “Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive”. In *Proceedings of TREC*, pp.199-210.

Saralegi, X. and I. Leturia. 2007. “Kimatu, a tool for cleaning non-content text parts from html docs.” In *Building and exploring web corpora, Proceedings of the 3rd Web as Corpus workshop,* pp. 163-167.

Saralegi, X., I. San Vicente and A. Gurrutxaga. 2008. “Automatic extraction of bilingual terms from comparable corpora in a popular science domain”. In *Proceedings of the Building and Using Comparable Corpora workshop (LREC08).*

Shao, L. and H.T. Ng. 2004. “Mining New Word Translations from Comparable Corpora.” In *Proceedings of COLING,* pp. 618-624.

Rayson, P. and R. Garside. 2000. “Comparing corpora using frequency profiling.” In *Proceedings of the workshop on Comparing Corpora (38th ACL)*, pp.1–6.