Extraction of Bilingual Cognates from Wikipedia *

Pablo Gamallo Otero[1] and Marcos Garcia[1]

Centro de Investigação em Tecnologias da Informação (CITIUS)
Universidade de Santiago de Compostela, Galiza, Spain
pablo.gamallo@usc.es , marcos.garcia.gonzalez@usc.es

Abstract. In this article, we propose a method to extract translation equivalents with similar spelling from comparable corpora. The method was applied on Wikipedia to extract a large amount of Portuguese-Spanish bilingual terminological pairs that were not found in existing dictionaries. The resulting bilingual lexicons consists of more than 27,000 new pairs of lemmas and multiwords, with about 92% accuracy.

1 Introduction

A comparable corpus consists of documents in two or more languages, which are not translation of each other and deal with similar topics. The use of comparable corpora to acquire bilingual lexicons has been growing in the last years [4, 5, 14, 3, 17, 16, 9, 20, 19, 15]. However, the number of these studies is not so large in comparison to those using a strategy based on aligned, parallel texts. A small but representative sample of extraction methods based on parallel texts is the following: [6, 12, 1, 18, 11].

The main advantage of comparable corpora is that they are easily available using the Web as a huge resource of multilingual texts. By contrast their main drawback is the low performance of the extraction systems based on them. According to [13], bilingual lexicon extraction from comparable corpora is a too difficult and ambitious objective, and much more complex than extraction from parallel, and aligned corpora. In fact, we can conceive a continuum of comparability between two poles: completely unrelated corpora (non comparable) and fully related parallel texts. The degree of comparability is directly related to the quality of the extracted lexicons. The more comparable is the corpus, the more precise the extracted lexicons are.

In this paper, we propose a method to learn bilingual lexicons from comparable corpora that try to overcome the low precision reached by most of the methods relying on comparable corpora. Two aspects will be taken into account to improve precision:

* This work has been supported by Ministerio de Ciencia e Innovación, within the project OntoPedia, ref: FFI2010-14986.

The degree of comparability of the corpus: we will discover articles in the Wikipedia with very high degree of comparability (pseudo-parallel texts).

The extraction of bilingual cognates: the extraction will be focused on bilingual pairs of words with similar spelling (*cognates*), which are in fact true translation equivalents and not false friends nor false cognates. *Bilingual cognates* are considered here as those words in two languages with similar spelling and similar meaning.

We assume that it is possible to build high quality bilingual lexicons if the extraction is performed on very comparable corpora considering only bilingual cognates. To minimize the low coverage of the lexicons acquired by this method, it is convenient to use it on family related languages sharing many cognates. So, our experiments were performed on Portuguese and Spanish, two very close Latin languages.

Moreover, among the different web sources of comparable corpora, Wikipedia is likely the largest repository of similar texts in many languages. We only require the appropriate computational tools to make them comparable. The proposed method is based on the Wikipedia structure, even if it can be easily generalized to be adapted to other sources of comparable corpora. In this paper, we will use the method to enlarge an existing Portuguese-Spanish bilingual dictionary with new bilingual cognates, most of them representing domain-specific terminology found in Wikipedia.

This article is organized as follows. In the next two sections, 2 and 3, we describe the extraction method. Then, in section 4 some experiments are performed in order to learn a large Portuguese-Spanish bilingual lexicon. Finally, some conclusions are presented in 5.

2 Method Overview

Our extraction method relies on the multilingual structure of Wikipedia. It consists of the following four steps:

Corpus alignment First, we identify the Wikipedia articles in two languages whose titles are translations of each other.

Degree of comparability Then, to calculate a degree of comparability between two aligned articles, we apply a similarity measure and select the most comparable pairs of bilingual articles.

Candidates for translation equivalents From each very comparable pair of articles, we calculate the Dice similarity between lemmas and select the most similar ones, which are considered as being candidates for translation equivalents.

Selecting cognates Finally, using the Edit distance, we check whether the candidates are *cognates* and select the most similar ones as true translation equivalents.

This whole method runs in time linear in the size of its input, and thus scales readily as the corpus grows. In the following section, we will describe in detail the four steps of our method.

3 Method Description

The method is based on the following assumption:

> The use of distributional similarity to extract bilingual cognates from very comparable corpora should generate high quality bilingual correlations

Following this assumption, we develop a strategy adapted to the Wikipedia structure. The output is a bilingual dictionary containing many bilingual pairs of domain-specific terms.

3.1 Alignment of Bilingual Wikipedia Articles

The input of our strategy is CorpusPedia[1], a friendly and easy-to-use XML structure, generated from Wikipedia dump files. In CorpusPedia, all the internal links found in the text are put together in a vocabulary list identified by the tag *links*. In addition, the tag *translations* codifies a list of interlanguage links (i.e., links to the same articles in other languages) found in each article. Both internal and interlanguage links are very useful features to build comparable corpora.

For this purpose, the first task is to extract all pairs of bilingual articles related by interlanguage links. For instance, given the Portuguese article entitled *"Arqueologia"*, we search for its Spanish counterpart within the list of Spanish *translations* associated to the Portuguese article. If the Spanish translation, *"Arqueología"*, is in the list, we select the article entitled *"Arqueología"* within the Spanish Wikipedia, and build a small comparable corpus with the pair of articles. This algorithm is applied article by article and results in a large set of small, comparable, and aligned pairs of bilingual texts.

3.2 Wikipedia-Based Comparability Measure

The next step is to measure the degree of comparability for each pair of bilingual texts, such as it was described in [8]. The measure of comparability between two Wikipedia articles is defined as follows.

For a comparable corpus $\mathcal{C}$ built from the Wikipedia, and constituted for instance by a Portuguese article $\mathcal{C}_p$ and a Spanish article $\mathcal{C}_s$, a comparability coefficient can be defined on the basis of finding, for each Portuguese term t_p in the vocabulary $\mathcal{C}_p^v$ of $\mathcal{C}_p$, its interlanguage link (i.e., translation) in the vocabulary $\mathcal{C}_s^v$ of $\mathcal{C}_s$. The vocabulary of a Wikipedia article is the set of "internal links" found in that article. Those internal links are the key words and terms representing the content of the article. So, the two articles, $\mathcal{C}_p$ and $\mathcal{C}_s$, tend to have a high degree of comparability if we find many internal links in $\mathcal{C}_p^v$ that can be translated (by

[1] The software to build CorpusPedia, as well as CorpusPedia files for English, French, Spanish, Portuguese, and Galician, are freely available at `http://gramatica.usc.es/pln/`

means of interlanguage links) into many internal links in $\mathcal{C}_s^v$. Let $Trans_{bin}(t_p, \mathcal{C}_s^v)$ be a binary function which returns 1 if the translation of the Portuguese term t_p is found in the Spanish vocabulary $\mathcal{C}_s^v$. The binary Dice coefficient, $Dice_{comp}$, between the two articles $\mathcal{C}$ is then defined as:

$$Dice_{comp}(\mathcal{C}_p, \mathcal{C}_s) = \frac{2 \sum_{t_p \in C_p^v} Trans_{bin}(t_p, \mathcal{C}_s^v)}{|\mathcal{C}_p^v| + |\mathcal{C}_s^v|} \quad (1)$$

It follows that two texts in two languages have a high degree of comparability if the main terms found in one text are translations of the main terms found in the other text. For instance, the pair of articles *"Arqueologia/Arqueología"* has a low comparability degree because the two texts only share two bilingual pairs of terms (e.g., *"sociedade/sociedad"*, *"antropologia/antropología"*) out of 95 (i.e., $Dice_{comp} = 0.04$). By contrast, the degree of comparability of the Portuguese and Spanish entries for the scientist *Brian Goodwin* reaches a $Dice_{comp}$ coefficient of 0.77, since the two articles share 13 out of 43 terms. Given that some authors of Wikipedia articles translate (part of) their contributions from the English version, the two articles on Brian Goodwin are likely to be translations from the English entry.

This comparability measure is applied to each pair of bilingual articles, and those pairs whose degree of comparability is higher than a specific threshold (empirically set to >= 0.3) are finally selected. It results in a set of very comparable pairs of bilingual texts. The whole set of very similar bilingual texts can be perceived as a pseudo-parallel corpus, since we observed that many pairs are constituted by paragraphs that are either (pseudo) translations of each other or translations sharing a common source.

3.3 Identifying Translation Equivalent Candidates

In this step, we identify all pairs of translation candidates within each pair of bilingual texts selected in the previous step. For this purpose, we apply a distributional-based strategy defined in [9]. The starting point of this strategy is as follows: lemma l_1 is a candidate translation of l_2 if the lemmas with which l_1 co-occurs are translations of the lemmas with which l_2 co-occurs. This strategy relies on a bilingual list of lemmas (called *seed words*) provided by an external bilingual dictionary. So, l_1 is a candidate translation of l_2 if they tend to co-occur with the same seed words. In our strategy, co-occurrences are defined by means of syntactic contexts, which are identified with a robust dependency parser, DepPattern [7]. Only three PoS categories of lemmas were taken into account: common nouns, adjectives, and verbs. In this experiment, proper names were not considered. They will be processed using a simpler strategy we will describe later.

Similarity between lemmas l_1 and l_2 is computed using the following version of the *Dice* coefficient:

$$Dice_{distrib}(l_1, l_2) = \frac{2 \sum_i min(f(l_1, s_i), f(l_2, s_i))}{f(l_1) + f(l_2)} \quad (2)$$

where $f(l_1, s_i)$ represents the number of times the lemma l_1 co-occurs with seed s_i, and $f(l_1)$ the total frequency of l_1 in the corpus. As a result, each lemma of the source language is assigned a list of candidate translations. Potential candidates are restricted to be of the same category: nouns are compared with nouns, verbs with verbs, and adjectives with adjectives.

It is worth mentioning that this strategy takes into account multiwords. Before computing similarity, the most representative multiword terms are identified with GaleXtra[2], a multilingual term extractor that uses both patterns of PoS tags and association measures to select term candidates. So, Dice similarity is computed on pairs of bilingual lemmas containing single lemmas and/or multiword terms.

Since our objective is to enlarge the existing bilingual dictionary, only bilingual pairs of lemmas that are not in that source dictionary are considered as candidate translation equivalents.

3.4 Identifying Bilingual Cognates

The final step is to identify *cognates* out of the set of translation equivalent candidates. For this purpose, we use a spelling similarity measure based on Edit distance. The spelling based similarity, noted $Dice_{eds}$, between two lemmas, l_1 and l_2, is defined by means of the following equation:

$$Dice_{eds}(l_1, l_2) = 1 - \frac{2\ eds(l_1, l_2)}{length(l_1) + length(l_2)} \quad (3)$$

where $eds(l_1, l_2)$ is the Edit distance of lemmas l_1 and l_2, and $length(l_i)$ represents the number of characters of l_i. It means that the $Dice_{eds}$ similarity between the spelling of two lemmas is a function of their Edit distance and their string lengths. This similarity measure allows us to select those pairs of translation candidates that share similar spelling, i.e., that can be perceived as being bilingual cognates. We assume that the final selected bilingual cognates are very probably correct translations.

4 Experiments

We performed an experiment aimed at learning a large set of new bilingual correlations from the Portuguese and Spanish versions of Wikipedia.

[2] `htpp://http://gramatica.usc.es/~gamallo/gale-extra/index2.1.htm`

4.1 Existing dictionaries

Our method requires a list of seed words taken from existing bilingual resources. We used two different dictionaries:

OpenTrad The general purpose bilingual dictionary Portuguese-Spanish integrated in the open source machine translation system, OpenTrad-Apertium [2]. The dictionary is freely available[3]. For our experiment, we selected all bilingual pairs containing nouns, verbs, and adjectives.

Wikipedia We created a new Portuguese-Spanish dictionary using the interlanguage links of Wikipedia. Since Wikipedia is an encyclopedia dealing with named entities and terms, this new dictionary only contains proper names and domain-specific terminology.

Table 1 shows the size of the two existing dictionaries and the total union of them. We count the different number of bilingual correspondences (not the number of entries). A bilingual correspondence is, for instance, the pair *"sociedade/sociedad"*. The total size obtained by the union of both resources is 263,362 different bilingual correspondences.

	nouns	**adject.**	**verbs**	**total**
OpenTrad	4,210	1,428	4,226	9,854
Wiki	253,367	-	-	253,367
Union	257,708	1,428	4,226	**263,362**

Table 1. Existing lexical resources

Note that the two dictionaries are complementary: we only found 131 entries in common.

4.2 Extraction

	nouns	**adject.**	**verbs**	**total**
single lemmas	9,374	5,725	2,215	17,314
multiword terms	9,585	-	944	10,529
all lemmas	18,959	5,725	3,159	**27,843**

Table 2. Results of the extraction

After applying our method on the whole Wikipedia in Portuguese and Spanish, we extracted 27,843 new bilingual correspondences. None of them were in the

[3] `http://sourceforge.net/projects/apertium/files/`

two input dictionaries. Preliminary tests led us to set the thresholds of $Dice_{comp}$, $Dice_{distrib}$, and $Dice_{eds}$ to 0.3, 0.6, and 0.6, respectively. Table 2 depicts the final results. In the first row, we show the extractions of single lemmas, while the second row is just focused on multiwords. The total extractions considering both multiword terms and single lemmas are shown in the third row.

Notice that the size of the new bilingual dictionary, 27843 lemmas, is much larger than that of the general purpose dictionary of OpenTrad, which only contain 9,854 bilingual correspondences.

4.3 Evaluation

	accuracy
nouns	91%
verbs	89%
adjectives	95%
total	**92%**

Table 3. Accuracy of the extracted bilingual pairs.

To evaluate the quality of the extracted dictionary, a test set of 450 bilingual pairs were randomly selected. The test set was created with the aim of obtaining these three balanced subsets:

- 150 bilingual pairs of nouns
- 150 bilingual pairs of verbs
- 150 bilingual pairs of adjectives

Results are depicted in Table 3. Accuracy is the number of correct pairs divided by the number of evaluated pairs. The best performance was achieved by the extraction of adjectives, since it reaches 95% accuracy. By contrast, verb extraction only achieves 89%. The total accuracy is still high: about 92%. This performance is much better than state-of-the-art work on extraction from comparable corpora, whose best scores were about 70% accuracy [14]. The good quality of the generated translation equivalents allows us to reduce the time spent in manual correction. It follows that our method permits to minimize the effort to build a new bilingual dictionary of two related languages.

4.4 Error Analysis

We found 39 errors out of 450 evaluated extractions. Most of them (58%) were due to foreign words, namely English words appearing in the input text as part of titles or citations. For instance, the translation pair "*about/about*" was incorrectly learnt from two Portuguese and Spanish texts containing such a word within a non translated English expression. It would not be difficult to avoid

this kind of problem if we use a language identifier to find parts of the input text written in other languages.

The second type of most common errors (8%) were caused by prefixes appearing in one of the two correlated words, for instance:

americanismo / **anti**-americanismo
anti-fascista / fascista
hispanorabe / **neo**-hispano-árabe

Note that it would be possible to filter out those cases by making use of a list of productive prefixes.

In Table 4, we show some types of errors found in the evaluation. As the two most common errors (foreign words and prefixes), which represent 66% of the total number of errors, can be easily filtered out, the total accuracy of our system could achieve 97%.

	frequency
foreign words	58%
prefixes	8%
typos	8%
multiwords	5%
PoS-tagging	3%

Table 4. Types of errors.

4.5 Further Experiments: the Case of Proper Names

As proper names are less ambiguous than common nouns, verbs, and adjectives, we consider it is not necessary to grasp high quality bilingual correspondences by using distributional-contextual similarity. So, we defined a simpler strategy to extract translation equivalents of proper names:

Given two bilingual pairs of articles in Wikipedia, for instance the Portuguese and Spanish articles entitled *"Arqueologia/Arqueología"*, all proper names found in those articles are identified, then a list of bilingual pairs is created, and the $Dice_{eds}$ similarity is computed between them. If two pairs of bilingual proper names with similar spelling has a $Dice_{eds}$ similarity higher than 0.9, they are selected as being a translation equivalent candidate. Notice that we do not compute the degree of comparability between the two Wikipedia articles nor the distributional $Dice_{distrib}$ similarity between the two proper names. In this case, we assume that two proper names with very similar spelling in two languages, and appearing in two texts with similar or identical titles, are strong candidates to be true translation equivalents.

This strategy led us to extract $818,797$ new bilingual pairs of proper names that were not found in the two existing dictionaries. This kind of bilingual pairs

can be integrated into rule-based machine translation systems, as OpenTrad. As this system is not provided with a Named Entity Recognition module, it proposes bilingual translations of proper names on the basis of large lists of bilingual correspondences of proper names.

5 Conclusions

We have proposed a method to extract new bilingual terminology from Wikipedia-based comparable corpora, achieving more than 90% accuracy. The proposed strategy was adapted to the internal structure of Wikipedia, but it could be applied to other types of comparable text corpora with minor changes.

One of the main drawbacks of our strategy is due to the inherent limitations of Edit distance to measure the spelling based similarity between lemmas that have undergone systematic changes in the past. For instance, action nouns in Portuguese may contain the suffix *-ção* while the equivalent suffix in Spanish is *-ción*. If we consider that these two suffix represent the same abstract concept, the distance between two words such as *organização* and *organización*, whose dissimilarity is just due to the suffix, must be close to 0, i.e., they should be taken as identical cognates. However, the Edit distance between these two words does not consider the strong relationship between the two suffixes. To address these cases, [10] proposed a new spelling similarity based on the generalization of substitution patterns. A different strategy, based on linguistic knowledge, could be the use of a list of equivalent bilingual pairs of prefixes and suffixes. In future work, we will make use of both strategies to enlarge the coverage of the extracted dictionaries.

References

1. Lars Ahrenberg, Mikael Andersson, and Magnus Merkel. A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 29–35, Montreal, 1998.
2. Carme Armentano-Oller, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, and Miriam A. Scalco. Open-source Portuguese-Spanish machine translation. In *Lecture Notes in Computer Science, 3960*, pages 50–59, 2006.
3. Y-C. Chiao and P. Zweigenbaum. Looking for candidate translational equivalents in specialized, comparable corpora. In *19th COLING'02*, 2002.
4. Pascale Fung and Kathleen McKeown. Finding terminology translation from non-parallel corpora. In *5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong, 1997.
5. Pascale Fung and Lo Yuen Yee. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Coling'98*, pages 414–420, Montreal, Canada, 1998.
6. Willian Gale and Kenneth Church. Identifying Word Correspondences in Parallel Texts. In *Workshop DARPA SNL*, 1991.

7. Pablo Gamallo and Isaac González. A grammatical formalism based on patterns of part-of-speech tags. *International Journal of Corpus Linguistics*, 16(1):45–71, 2011.
8. Pablo Gamallo and Isaac González. Measuring comparability of multilingual corpora extracted from wikipedia. In *Workshop on Iberian Cross-Language NLP tasks (ICL-2011)*, Huelva, Spain, 2011.
9. Pablo Gamallo and José Ramom Pichel. Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary. *LNCS*, 4919:413–423, 2008.
10. L. Gomes and G.P. Lopes. Measuring spelling similarity for cognate identification. In *EPIA 2011, LNAI 7026*, pages 624–633, 2011.
11. Oi Yee Kwong, Benjamin K. Tsou, and Tom B. Lai. Alignment and extraction of bilingual legal terminology from context profiles. *Terminology*, 10(1):81–99, 2004.
12. Dan Melamed. A Portable Algorithm for Mapping Bitext Correspondences. In *35th Conference of the Association of Computational Linguistics (ACL'97)*, pages 305–312, Madrid, Spain, 1997.
13. Hiroshi Nakagawa. Disambiguation of single noun translations extracted from bilingual comparable corpora. *Terminology*, 7(1):63–83, 2001.
14. Reinhard Rapp. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *ACL'99*, pages 519–526, 1999.
15. Raphael Rubino and Georges Linarés. A multi-view approach for term translation spotting. In *CICLing 2011*, pages 29–40, LNCS 6609, 2011.
16. X. Saralegui, I. San Vicente, and A. Gurrutxaga. Automatic generation of bilingual lexicons from comparable corpora in a popular science domain. In *LREC 2008 Workshop on Building and Using Comparable Corpora*, 2008.
17. Li Shao and Hwee Tou Ng. Mining New Word Translations from Comparable Corpora. In *20th International Conference on Computational Linguistics (COLING 2004)*, pages 618–624, Geneva, Switzerland, 2004.
18. Jorg Tiedemann. Extraction of Translation Equivalents from Parallel Corpora. In *11th Nordic Conference of Computational Linguistics*, Copenhagen, Denmark, 1998.
19. Kun Yu and Junichi Tsujii. Bilingual dictionary extraction from wikipedia. In *Machine Translation Summit XII*, Ottawa, Canada, 2009.
20. Kun Yu and Junichi Tsujii. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *NAACL HLT 2009*, pages 121–124, Boulder, Colorado, 2009.