Automatic Generation of Bilingual Dictionaries using Intermediary Languages and Comparable Corpora

Pablo Gamallo Otero[1] and José Ramom Pichel Campos[2]

[1] Departamento de Língua Espanhola,
Universidade de Santiago de Compostela, Galiza, Spain
pablo.gamallo@usc.es
[2] Departamento de Tecnologia Linguística da Imaxin|Software
Santiago de Compostela, Galiza
jramompichel@imaxin.com

Abstract. This paper outlines a strategy to build new bilingual dictionaries from existing resources. The method is based on two main tasks: first, a new set of bilingual correspondences is generated from two available bilingual dictionaries. Second, the generated correspondences are validated by making use of a bilingual lexicon automatically extracted from non-parallel, and comparable corpora. The quality of the entries of the derived dictionary is very high, similar to that of hand-crafted dictionaries. We report a case study where a new, non noisy, English-Galician dictionary with about $12,000$ correct bilingual correspondences was automatically generated.

1 Introduction

In this paper we describe a method to derive a new bilingual lexicon from two existing ones using comparable corpora to validate candidate correspondences. The method is entirely unsupervised and consists of two tasks. First, given two existing bilingual lexicons for two languages pairs (A, B) and (B, C), we can obtain a new pair (A, C) by simple transitivity. Second, the generated bilingual correspondences are validated using translation equivalents automatically extracted from comparable corpora. In particular, we will derive a new $(English, Galician)$ lexicon from two existing dictionaries, $(English, Spanish)$ and $(Spanish, Galician)$, by making use of English-Galician comparable corpora.

The strategy described in the paper is especially well suited to create new language resources for minority languages (e.g., Galician) from languages such as English or Spanish, which have a lot more resources. Our method does not require the minority language being provided with many and large linguistic resources: only a bilingual dictionary and some raw text is required. This is enough to automatically build a new non-noisy, bilingual lexicon.

This strategy is also useful to create new bilingual dictionaries for multilingual machine translation systems, such as Opentrad-Apertium[3]. The number of bilingual dictionaries required by a multilingual translator increases as a quadratic function of the number of languages the system aims to translate [15]. So, the process of automatically deriving new bilingual resources can drastically reduce the amount of work.

The paper is organized as follows: the following section (2) introduces some related work. Then, Section 3 describes the different steps of our method. Next, in Section 4, we report a case study where a new, non-noisy, English-Galician dictionary with about $12,000$ bilingual correspondences was automatically generated. Finally, some conclusions are put forward in Section 5.

2 Related work

There exist some approaches to derive bilingual lexicons from existing ones [11, 1, 16, 10, 15]. Our work is directly inspired by [10], who sketch a very similar methodology to that proposed here. They use two bilingual lexicons sharing the same language (the pivot) and derive a new bilingual dictionary by using the pivot language as intermediate. The new lexicon is derived by transitivity. For instance, given the language pairs $(English, Spanish)$ and $(Galician, Spanish)$, as Spanish as language pivot, their method build a new bilingual pair without the pivot language: $(English, Galician)$. The crucial aspect of this strategy is the validation of correspondences. The validity of the retained correspondences was checked using a parallel corpus, i.e., only the correspondences found in the parallel corpus are kept.

The specificity of our method is the fact that we used comparable corpora, instead of parallel texts, to validate the correspondences retained by transitivity. So, our main contribution is to propose a strategy to validate new bilingual lexicons by making use of translation equivalents extracted from non-parallel, comparable corpora. This kind of corpus is easier available than parallel texts, especially for minority languages.

Unlike most approaches to extract word translations from non-parallel corpora [6, 7, 12, 4, 14, 13], which are based on baseline windowing techniques, our method relies on syntactically analyzed text. In [9], it is showed that the use of syntactic dependencies instead of window-based strategies significantly improves the accuracy of the extraction.

3 The method

Our strategy consists of two main tasks: both to generate candidate bilingual correspondences by transitivity and to validate them by using translation equivalents extracted from comparable corpora.

[3] `http://www.opentrad.com/`

3.1 Generation by transitivity

The first task is inspired by that described in [10]. Given two bilingual dictionaries represented as two relations (A, B) and (B, C), we generate a derived dictionary (A, C) as follows:

- First, we create the relation (A, C') taking two existing dictionaries (A, B) and (B, C), where B is the pivot language. For each bilingual correspondence (a_i, b_i) belonging to the relation (A, B), we create a set of new correspondences $\{(a_i, c_1), (a_i, c_2), \ldots, (a_i, c_n)\}$, where $c_1, \ldots, c_n$ are those words and terms associated with b_i within (B, C). The derived dictionary (A, C') is the set of all new bilingual correspondences.
- Then, we remove the redundant bilingual pairs from (A, C'). The result is the relation (A, C).
- Finally, we split (A, C) into two complementary subsets: $(A, C)_{amb}$, which consists of those correspondences containing at least one ambiguous word, and $(A, C)_{unamb}$, containing only unambiguous words. Note that the former is a many-to-many relationship whereas the latter is one-to-one.

As in [10], the derived dictionary with only unambigous words, $(A, C)_{unamb}$, can be considered as a non-noisy lexical resource. In Lexicography, words with only one translation equivalent behave as not ambiguous terms. Therefore, all the unique correspondences derived from unambigous words (one-to-one) are of good quality and must be validated. By contrast, $(A, C)_{amb}$ is a noisy lexicon. The translation by transitivity of ambiguous words can overgenerate odd bilingual correspondences. For instance, in one of our $(English, Spanish)$ dictionaries, the verb *subside* is translated in Spanish as *bajar*, which is translated, in turn, by the $(Spanish, Galician)$ dictionary as *baixar* and *apear*. Therefore, the derived $(English, Galician)$ dictionary must contain the correspondences $(subside, baixar)$ and $(subside, apear)$. While the former translation is correct, the latter is clearly odd. The galician verb *apear* does not mean *subside* in any context; it means *take down*, which is one of the senses of the spanish word *bajar*.

In the next task, all correspondences of $(A, C)_{amb}$ will be checked using translation equivalents between language A and C extracted from comparable corpora.

3.2 Validation with comparable corpora

The second process is the main contribution of our work. It consists in filtering out those ambiguous correspondences that are not in a lexicon of translation equivalents automatically generated from a non-parallel corpus syntactically annotated with dependencies. The lexicon of translation equivalents, called $(A, C)_{corpus}$ is organized as follows. Each term of language A, a_i, is assigned a ranked list of terms of language C, $c_1, c_2, \ldots, c_n$, which are the top-N best translation candidates of a_i. Conversely, each term of language C, c_i, is assigned a ranked list of terms of language A, $a_1, a_2, \ldots, a_n$, which are the top-N best translation candidates of c_i. So, the relation $(A, C)_{corpus}$ consists of correspondences between words and their candidate translations inferred from the corpus. To validate $(A, C)_{amb}$, we make the intersection between $(A, C)_{amb}$ and $(A, C)_{corpus}$. The resulting relation is a set of correct bilingual correspondences containing ambiguous words. Finally the new non-noisy, derived lexicon,$(A, C)_{not-noisy}$, is the union of this validated relation with the bilingual lexicon of unambiguous words:

$$(A, C)_{not-noisy} = (A, C)_{amb} \cap (A, C)_{corpus} \cup (A, C)_{unamb}$$

In the following subsection, it is described how $(A, C)_{corpus}$ is learned.

3.3 An approach to extract translation equivalents from comparable corpora

Our method to extract translation equivalents from syntactically annotated comparable corpora was described in detail in previous work [9, 8]. Here, we only sketch the main properties of the approach. The starting point is the following: word w_1 is a candidate translation of w_2 if the lexical-syntactic contexts in which w_1 occurs are translations of the lexical-syntactic contexts in which w_2 occurs. Words (or multiword terms) are previously lemmatized. This strategy relies on a list of bilingual lexical-syntactic contexts (called *seed contexts*) provided by an external bilingual dictionary, (A, C), and a list of generic syntactic dependencies: subject, direct object, adjective modification, prepositional complement, etc. So, w_1 is a candidate translation of w_2 if they tend to co-occur with the same seed contexts. For instance, let's suppose that the dictionary (A, C) contains the correspondence *(subside, baixar)*. As they are two specific verbs, we can build a bilingual correspondence between two lexical-syntactic contexts introduced by their corresponding verbs:

$$(< Subject; subside, NOUN >, < Subject; baixar, NOUN >)$$

where $< Subject; subside, NOUN >$ is used to identify those English nouns appearing in the subject position of *subside*, while $< Subject; baixar, NOUN >$ allows to select those Galician nouns playing the role of subject of *baixar*. This bilingual correspondence is used as a "seed context" in the process of selecting translation equivalents. This way, if English nouns such as *fever* or *swelling* appear as subject of *subside*, the Galician nouns occurring in the subject position of *baixar* (e.g., *febre* or *inchazón*) are candidate to be their translations.

The extraction method consists of the following subtasks[4] :

Multilingual parsing The two corpora are analyzed using a multilingual dependency based parser, DepPattern[5].

[4] This method was implemented in a prototype system available at `http://gramatica.usc.es/~gamallo/prototypes/BilingualExtraction.tar.gz`.

[5] Available, under GPL license, at `http://gramatica.usc.es/pln/tools/deppattern.html`.

Seed contexts A list of seed lexical-syntactic contexts is created from the noisy bilingual dictionary, (A, C), and a small set of generic syntactic rules. Note that the bilingual dictionary used as source is that derived by transitivity in the previous task. It contains both ambiguous and unambiguous correspondences, even if the former ones can contain several errors.

Hash table The word dependencies identified in the corpora and the list of seed contexts are organized in a word-context matrix (stored in memory as a hash table of non-zero values). Each item of the table represents a word (or multiword term), a seed context, and the word-context frequency observed in the corpus.

Similarity Then, we compute dice similarity [5] of each bilingual pair of words. For each word of the source language, we select its top-N ($N = 10$) most similar ones in the target language. They are their candidate translations.

At the end of the process, we obtain the relationship $(A, C)_{corpus}$, which will be used to validate $(A, C)_{amb}$ by identifying correct ambiguous correspondences. As it was stated above, the selection of correct correspondences is the result of intersecting $(A, C)_{amb}$ with $(A, C)_{corpus}$.

4 A case study: the elaboration of an English-Galician dictionary

To verify whether the method is useful, we apply it to perform a particular task, namely to derive a new English-Galician dictionary from two existing ones. This case study has two limitations: given that Galician is a language with few electronic resources, the Galician part of our comparable corpus is considerably smaller than the English one. On the other hand, since the extraction method only works at the moment on nouns, verbs, and adjectives, the dictionary elaboration is restricted to these three grammatical categories.

4.1 The existing dictionaries and generation by transitivity

The $(English, Galician)$ dictionary was derived from both $(English, Spanish)$ and $(Spanish, Galician)$ existing dictionaries, where Spanish is the pivot language. In particular, the bilingual dictionaries we used are part of the lexical resources integrated in an open source machine translation system: OpenTrad-Apertium [2]. In fact, one of the short-mid term objectives of our experiments is to update the bilingual resources of OpenTrad in order to improve the results of the machine translation system, which is used by *La Voz de Galicia*, the sixth most widely read Spanish newspaper.

The $(English, Spanish)$ dictionary contains 8,432 bilingual correspondences, while the $(Spanish, Galician)$ reaches 27,640. Both dictionaries are freely available[6]. Given that the former dictionary is too small, we also made use of a

[6] http://sourceforge.net/projects/apertium/files/

Collins dictionary[7], which we call ($English\_C, Spanish\_C$), and contains 48, 637 entries. This resource is not freely available. Note that we only count bilingual correspondences between verbs, nouns, and adjectives. All of these dictionaries were manually created by lexicographers.

Table 1. Dictionaries derived by transitivity

derived dictionaries	number of entries	ambiguous entries	not ambiguous entries	source dictionaries
($English, Galician$)	7, 687	3, 890	3, 797	($Galician, Spanish$) ($Spanish, English$)
($English\_C, Galician$)	23, 094	17, 601	5, 494	($Galician, Spanish$) ($Spanish\_C, English\_C$)

Using the strategy described above in Section 3.1, we generated two new noisy bilingual dictionaries: ($English, Galician$) and ($English\_C, Galician$) (see Table 1). The first raw of the table shows the different elements of ($English, Galician$), which was derived from the two OpenTrad-Apertium dictionaries (sources). It contains 7, 687 correspondences that was splitted into two subsets:

- ambiguous correspondences: $(English, Galician)_{amb}$
- not ambiguous ones: $(English, Galician)_{not-amb}$

They contain 3, 890 and 3, 797 entries, respectively (third and fourth columns of the table). The same was made to obtain ($English\_C, Galician$), which was derived from ($Spanish\_C, English\_C$) (Collins) and ($Galician, Spanish$) (OpenTrad-Apertium). Here, the size of the resulting lexicon is larger because of the higher number of entries provided by the Collins dictionary.

4.2 Comparable corpora and validation

To validate the English-Galician correspondences with ambiguous words, we used the strategy described in sections 3.2 and 3.3. First, we built different non-parallel, (and somehow) comparable corpora. Then, the automatic extraction of translation equivalents were performed on those corpora.

Building three comparable corpora The Galician part was crawled from two online daily newspapers, Vieiros and Galicia-Hoxe, which are the only general purpose newspaper written in Galician language. The crawler retrieved all news published by these newspaper since they are available in the net. We built a corpus with 35 million word tokens.

The English part was divided in three different corpora:

[7] http://www.collinslanguage.com/

- 35M words selected from British National Corpus (BNC)[8],
- 35M words containing breaking news from Reuters Agency[9]
- 1M words containing news crawled from New York Times (NYT)

Given that we could not find more Galician Newspapers, to obtain a corpus size comparable to that of the English part, we decided to build 3 non-parallel corpus as follows:

BNC-based This corpus is constituted by all Galician news (35M words) and the 35M words selected from BNC.

Reuters-based It constituted by all Galician news and the 35M words from Reuters

NYT-based It contains 1M words selected from the Galician corpus and 1M words crawled from NYT.

So, BNC-based and Reuters-based corpora contains the same Galician corpus while NYT-based is constituted by a small partition of that corpus. We followed this strategy because of the few electronic resources in Galician language. Let's note that the BNC-based corpus is less comparable than the others since the English part does not only contain news articles. It consists of many types of documents, including oral speech.

Extraction The extraction method was sketched in Section 3.3. First, all texts were parsed with DepPattern to extract all word dependencies (we focused on dependencies containing verbs, nouns, or adjectives). DepPattern takes as input the output of the PoS tagger Freeling[3]. Then, a list of seed lexical syntactic contexts was generated from the largest English-Galician lexicon: $(English\_C, Galician)$. Even if it is likely to contain some odd bilingual correspondences, we consider that it is sound enough to be used for stochastic-based extraction. Then, on the basis of word dependencies and a list of contexts, three context-word bilingual matrices were created (one for each corpus). Finally, word similarity was computed on each matrix. For each English word, the 10 most similar Galician words were retained to define 10 candidate bilingual correspondences. Since similarity is an asymmetric relationship, the same was done from Galician to English. At the end of the process, we built three corpus-based bilingual lexicons: $(English\_C, Galician)_{bnc}$, $(English\_C, Galician)_{reuters}$, and $(English\_C, Galician)_{nyt}$. Table depicts the number of correspondences of each dictionary.

Table 2 shows the results obtained. Corpus-based dictionaries are much bigger than those directly derived by transitivity, and so they contain much more noisy correspondences. The goal is to generate for each word, at least, a good bilingual correspondence which will be used to validate dubious pairs derived by transitivity. Notice also than the Reuters-based dictionary is significantly larger

[8] http://www.natcorp.ox.ac.uk/

[9] http://trec.nist.gov/data/reuters/reuters.html

Table 2. Corpus-based dictionaries

dictionaries	number of entries
$(English\_C, Galician)_{bnc}$	$400,440$
$(English\_C, Galician)_{reuters}$	$531,710$
$(Spanish\_C, English)_{nyt}$	$132,490$

than the BNC-based, even if the corpus size over which the extraction was performed is the same. This is probably due to the fact that the BNC-based corpus is less comparable (it is just a "non-parallel" corpus).

Validation To check the validity of the dubious correspondences within the ambiguity-based lexicons (i.e., containing ambiguous words), we make their intersection with the corpus-based lexicons. Table 3 shows the outputs of all possible intersections between the three corpus-based dictionaries (columns) and the two lexicons with ambiguous words (rows). The third row is the union of the two ambiguity-based dictionaries, while the last column is the union of the three corpus-based lexicons. Each absolute number is assigned a percentage: the ration between the correspondences validated (i.e., resulting of the intersection) divided by the total number of correspondences found in the dictionary with ambiguous words.

Table 3. Corpus-based validation

	bnc	*reuters*	*nyt*	**Union**
$(English, Galician)_{amb}$	$1,123$ (29%)	$1,350$ (35%)	396 (10%)	$1,573$ (40%)
$(English\_C, Galician)_{amb}$	$2,404$ (14%)	$2,940$ (17%)	619 (4%)	$3,584$ (20%)
Union	$2,837$ (15%)	$3,475$ (18%)	759 (4%)	**$4,248$ (22%)**

For instance, The intersection of $(English, Galician)_{amb}$ with the smallest corpus-based lexicon, $(English\_C, Galician)_{reuters}$, gives rise to $1,350$ correspondences, which represent 35% of $(English, Galician)_{amb}$. Notice that successive unions of dictionaries improve the results by making the output dictionary larger. The largest lexicon was obtained by intersecting the union of the corpus-based lexicons with the union of the two ambiguity-based dictionaries: $4,248$ correct entries. It represents 22% of entries found in the union of the two ambiguity-based dictionaries ($19,425$ entries). These results are not very far from those obtained by [10] using parallel corpora. These authors reported an experiment to derive by transitivity an English-German dictionary, whose ambiguity-based correspondences were validated using parallel corpora. The result of this checking process allowed them to validate $6,282$ correspondences, which represent 26% of all candidate correspondences with ambiguous words.

Even if we use non-parallel corpora, our results are very close to that score, which is very promising.

The quality of the validated correspondences is very good. No error was found.

4.3 The final not-noisy lexicon

Table 4. non-noisy dictionary

	number of entries
OpenTrad + Collins	25,790
Validated correspondences	4,248
Not ambiguous correspondences	7,816
Total not-noisy dictionary	**12,064** (**47%**)

At the end of the process, we made the union of the validated correspondences with the lexicons containing unambiguous words (i.e., one-to-one correspondences). Table 4 summarizes the number of entries obtained in each step of the process. The last row shows the total number of non-noisy correspondences, 12,064, our method was able to automatically generate. This represents 47% of the total correspondences, 25,790, resulting of the union of $(English, Galician)$ with $(English\_C, Galician)$.[10]

To summarize, the output dictionary is the result of the following set-theoretic operations:

$\boldsymbol{(English, Galician)_{not-noisy}} =$

$((English, Galician)_{amb} \cup (English\_C, Galician)_{amb})$
$\cap$
$((English_C, Galician\_C)_{bnc} \cup (English\_C, Galician\_C)_{reuters} \cup$
$(English\_C, Galician\_C)_{nyt})$
$\cup$
$((English, Galician)_{not-amb} \cup (English\_C, Galician)_{not-amb})$

Let's note that the final lexicon, even if it only contains 47% of all candidate correspondences generated by transitivity, is much larger than the smallest hand-crafted dictionary, $(English, Spanish)$, which is one of the existing dictionaries used as source to derive the new one. We generated more than 12,000 correct correspondences against 7,687 entries in the smallest existing lexicon. The quality of the derived entries is similar to those found in dictionaries built by hand by lexicographers.

[10] The final dictionary can be downloaded at `http://gramatica.usc.es/~gamallo/dicosFromComparable.htm`

5 Conclusions and Future Work

The lexicographic method proposed in this paper is entirely automatic. It does not require any manual revision to generate a new bilingual dictionary since the quality of the derived correspondences is very high, similar to that achieved by a human lexicographer. The main contribution of the method is the use of lexicon extracted from syntactically annotated comparable corpora to validate correspondences derived by transitivity. Moreover, the experiments showed that the information provided by other source dictionaries and more corpus allowed us to easily make derived dictionaries much larger without losing quality.

The main drawback of the method is to be language dependent since it requires a syntactic parser to annotate the corpus. However, in order to cope with as many language as possible, we make use of a robust multilingual parser, DepPattern, designed and implemented by our research group.

In future work, we'll integrate the resulting dictionaries into a machine translation system, namely OpenTrad-Apertium, with the aim of adapting the system to new pairs of languages.

Acknowledgments

This work has been supported by the Galician Government, within the projects with reference PGIDIT07PXIB204015PR and 2008/101.

References

1. Kisuh Ahn and Matthew Frampotn. Automataic generation of translation dictionaries using intermediary languages. In *Cross-Language Knowledge Induction Workshop of EACL06*, pages 41–44, Trento, Italy, 2006.
2. Carme Armentano-Oller, Rafael C. Carrasco, Antonio M. Corb-Bellot, Mikel L. Forcada, Mireia Ginest-Rosell, Sergio Ortiz-Rojas, Juan Antonio Prez-Ortiz, Gema Ramrez-Snchez, Felipe Snchez-Martnez, and Miriam A. Scalco. Open-source portuguese-spanish machine translation. In *Lecture Notes in Computer Science, 3960*, pages 50–59, 2006.
3. X. Carreras, I. Chao, L. Padró, and M. Padró. An open-source suite of language analyzers. In *4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004.
4. Y-C. Chiao and P. Zweigenbaum. Looking for candidate translational equivalents in specialized, comparable corpora. In *19th COLING'02*, 2002.
5. James R. Curran and Marc Moens. Improvements in automatic thesaurus extraction. In *ACL Workshop on Unsupervised Lexical Acquisition*, pages 59–66, Philadelphia, 2002.
6. Pascale Fung and Kathleen McKeown. Finding terminology translation from nonparallel corpora. In *5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong, 1997.
7. 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.

8. Pablo Gamallo. Learning bilingual lexicons from comparable english and spanish corpora. In *Machine Translation SUMMIT XI*, Copenhagen, Denmark, 2007.
9. Pablo Gamallo and J-R. Pichel. Learning spanish-galician translation equivalents using a comparable corpus and a bilingual dictionary. *LNCS*, 4919:413–423, 2008.
10. Luka Nerima and Eric Wehrli. Generating bilingual dictionaries by transitivity. In *LREC-2008*, pages 2584–2587, 2008.
11. Kyonghee Paik, Satoshi Shirai, and Hiromi Nakaiwa. Automatic construction of a transfer dictionary considering directionality. In *COLING-2004 Multilingual Linguistic Resources Workshop*, pages 25–32, Geneva, 2004.
12. Reinhard Rapp. Automatic identification of word translations from unrelated english and german corpora. In *ACL'99*, pages 519–526, 1999.
13. 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.
14. 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.
15. Eric Wehrli, Luca Nerma, and Yves Scherrer. Deep linguistic multilingual translation and bilingual dictionaries. In *Foruth Workshop on Statistical Machine Translation*, pages 90–94, Athens, Greece, 2009.
16. Yujie Zhang, Qing Ma, and Hitoshi Isahara. Building japanese-chinese translation dictionary based on EDR japanese-english bilingual dictionary. In *MT Summit XI*, pages 551–557, Copenhagen, 2007.