Using Natural Alignment to Extract Translation Equivalents*

Pablo Gamallo Otero

Departamento de Língua Espanhola, Faculdade de Filologia
Universidade de Santiago de Compostela, Galiza, Spain
`pablogam@usc.es`

Abstract. Most methods to extract bilingual lexicons from parallel corpora learn word correspondences using relative small aligned segments, called sentences. Then, they need to get a corpus aligned at the sentence level. Such an alignment can require further manual corrections if the parallel corpus contains insertions, deletions, or fuzzy sentence boundaries. This paper shows that it is possible to extract bilingual lexicons without aligning parallel texts at the sentence level. We describe a method to learn word translations from a very roughly aligned corpus, namely a corpus with quite long segments separated by "natural boundaries". The results obtained using this method are very close to those obtained using sentence alignment. Some experiments were performed on English-Portuguese and English-Spanish parallel texts.

1 Introduction

To automatically extract a bilingual lexicon from a parallel corpus, we need first to align it, i.e., the various pieces or segments of text must be put into correspondence. This makes the lexicon extraction easier and more reliable. Many statistical algorithms for extraction of bilingual lexicon entries use statistical information derived from a sentence-aligned clean parallel corpus [5, 9, 1, 15, 12, 16, 8] The main drawback of this algorithm is that they are not very reliable when corpora have unclear sentence boundaries or the bilingual texts contain deletions, insertions and merged paragraphs. As [4] reported, many bilingual corpora contain noisy translations. In such a case, the aligned texts need to be manually corrected. It means that most approaches to bilingual dictionaries construction rely, to a certain extent, on supervised strategies.

The objective of this paper is to describe a specific method to compile bilingual lexicon entries from parallel corpora. Unlike most approaches cited above, we do not align texts at the sentence level. Bilingual texts are aligned using basic boundaries such as those separating chapters or articles. We call this strategy

---

* This work has been supported by Ministerio de Educación y Ciencia of Spain, within the project CESAR+, ref: BFF2003-02866.

"natural alignment". As the detection of these long and *natural* segments is a trivial task that does not need further corrections, the extraction method we will describe is entirely unsupervised.

To achieve the same accuracy as comparable approaches to bilingual lexicon extraction, the correlation metric we propose in this paper will be well-suited to account for longer aligned segments than sentences. The correlation metrics used by most related work measure, given two words, if they co-occur in the list of paired sentences. These metrics rely on a boolean test: they count the times two words do and do not co-appear in the correlated sentences. Such a test is not suitable for long segments, where several tokens of the same word can occur. Our metric, on the contrary, takes into account the fact that a word can occur several times in a single (but long) segment.

Using this metric in several experiments, we verified that the accuracy of the extraction method remains stable when the parallel corpus contains a reduced number (no more than $2,000-3,000$) of aligned segments. The fact of increasing the number of segments does not improve accuracy in a significant way.

The main advantages of the method proposed in this paper are the following: First, lexicon extraction can be done from noisy bilingual corpus without making any correction (it is an entirely unsupervised method). Second, since we need very few (but long) aligned segments, the system turns out to be computationally more efficient than the standard strategies relying on many aligned sentences. And third, accuracy is not lower than in comparable approaches. In sum, a bilingual lexicon acquired using our method is, at least, as complete and accurate as a lexicon extracted from a sentence-aligned corpus. So, it can be used as a set of reliable anchor points to incrementally search for smaller aligned segments, until fine-grained word or multi-word level is achieved.

The remaining paper is organised as follows. In Section 2, we discuss some related work. Then, Section 3 introduces the notion of natural alignment. Section 4 defines the similarity coefficient proposed, and finally, in Section 5, some relevant experiments are performed and evaluated. The experiments were performed on both English-Spanish and English-Portuguese corpora.

2 Related Work

Most work on bilingual lexicon acquisition primarily relies on text aligned at the sentence level. The first approaches for sentence alignment used length-based algorithms [2,3]. These algorithms assume that both the source and target sentences in the parallel corpus will be of constantly related length. Whereas [2] defined sentence length taking into account the number of words within a sentence, [3] defined length by counting the number of characters. Sentence delimiters are full stops. These strategies perform well only if the parallel texts have clear sentence boundaries. But full stops do not always mark sentence boundaries. Problems also arise when the target text contains insertions or deletions.

Other approaches to sentence-alignment use text based algorithms. They try to exploit more types of *correspondence points*, also called *anchor points*.

The most common anchor points used by these approaches are the sequences of characters with the same form in both texts (e.g., homographs such as proper names and numbers) [12], or with similar forms (e.g., cognates as "president" and "presidente" in English and Spanish) [11, 13, 10]. Other anchor points can be taken from bilingual dictionaries or even from previously known translation equivalents. Anchor points are more robust to insertion and deletion than length-based algorithms. However, the main problem of text-based approaches is the large number of noisy points that make alignment less precise. To find reliable correspondence points, it is necessary to define heuristics to filter out those points that are not correct anchors.

Instead of elaborating complex strategies to filter out noisy correspondence points, we consider natural boundaries as reliable delimiters for parallel text alignment.

3 Natural Alignment

We align the source and target texts by detecting natural boundaries such as chapters of novels, directives of laws, articles of journals, articles of constitutions, etc. As natural boundaries preserve the basic structure of the document, they are more reliable anchor points than punctuation marks such as full stops or paragraph delimiters. For instance, the English translation of "El Quijote", found in the free catalogue of the Gutenberg project[1], contains 1,639 paragraphs whereas the source text is up 4,000. This makes automatic sentence alignment difficult even for well-known software aligners such as *Vanilla*[2]. By contrast, the two parallel texts - i.e., the source novel and its English translation - share the same number of chapters. So, we do not need to make further manual corrections if the alignment is made at the level of chapters.

The main drawback of this type of alignment is obviously the fact that it often selects long segments. Two advantages, however, deserve to be mentioned. First, deletions and additions found in some parts of the source texts do not prevent the correct alignment of the whole parallel corpus. And second, this alignment does not need manual correction. To overcome the problems derived from an alignment containing long segments, we posit a particular version of the Dice correlation metric.

4 Correlation Metric

To extract bilingual dictionaries from aligned texts, we need to measure the co-occurrence of words in the different aligned segments.

Following [14], we consider the Dice coefficient as an appropriate metric to measure translation correlations between word types. Given a source word w_1

---

[1] http://www.gutenberg.org/

[2] http://nl.ijs.si/telri/Vanilla/

and a candidate translation w_2, our particular version of the Dice coefficient is defined as follows:

$$Dice(w_1, w_2) = \frac{2F(w_1, w_2)}{F(w_1) + F(w_2)}$$

where

$$F(w_1, w_2) = \sum_i min(f(w_1, s_i), f(w_2, s_i))$$

and

$$F(w_n) = \sum_i f(w_n, s_i)$$

Note that $f(w_1, s_i)$ represents the frequency of the word type w_1 occurring in segment s_i. Unlike most approaches to bilingual lexicon extraction, we consider that the frequency of a word in a particular segment carries a very significant information. As the segments we use to align the corpus are longer as those used for sentence alignment, then, the same word can occur several times in the same segment. So, we assume a word of the target language, w_2, is likely to be a translation of a source expression, w_1, if they tend to have a similar frequency in each segment s_i. This is the main difference with regard to standard approaches. In most approaches, as segments are supposed to be small, the metric only takes into account boolean searches. More precisely, they check if the source word and its candidate translation do or do not appear in the same aligned segments. However, as in our approach many different words can occur at least once in all segments, we need a more informative feature, namely the number of times a word occur in each segment.

`have`	51	11	16	7	...
`haber`	46	8	13	4	...
`estar`	1	2	3	1	...

Table 1. Word vectors with frequency information.

Let's take an example. Table 1 shows the first 4 positions of three word vectors: "have", "haber", and "estar". They were defined using the *EuroParl* parallel corpus (English-Spanish).

Concerning the verb "have", the first aligned segment contains 51 verb occurrences, the second segment 11, and so on. As natural segments are quite large, we need to take into account the number of occurrences in each segment in order to learn that "haber", and not "estar", is the likely translation of "have". Our particular definition of $F(w_1, w_2)$ allows us to grasp such a behaviour. By contrast, if the word vectors were defined using only a boolean test ("a word do and do not appear in a segment"), it would be impossible to distinguish between "haber" and "estar", since both Spanish words would be 100% similar to "have". See Table 2.

`have`	1 1 1 1 . . .
`haber`	1 1 1 1 . . .
`estar`	1 1 1 1 . . .

Table 2. Word vectors with boolean information.

5 Experiments and Evaluation

Our objective is to test whether natural segments are useful to extract correct translation equivalents. For this purpose, we need to observe the behaviour of both the length and the number of aligned segments in the extraction process. Two different experiments were performed.

5.1 First Experiment

Table 3 shows the results of our first experiment. We used two different English-Spanish parallel corpora: the European Parliament proceedings, *EuroParl* [7], and the European Constitution. We selected a bitext of 1 million word tokens from EuroParl and then performed two different alignments. On the one hand, we used as natural segments the speaker interventions in the European Parliament. As a result, we obtained 3,000 aligned segments, where the average length of each segment is 2 KBytes. On the other hand, we aligned the same corpus using shorter segments (paragraphs). The number of paragraphs does not differ in the two languages because they were previously filtered by Koehn [7]. As paragraphs are 4 times shorter than the speaker interventions, this alignment gave rise to 12,000 segments. Concerning the European Constitution, we also aligned the corpus in two different ways. First, the alignment was performed using as natural segments the articles themselves. As a result, we obtained 881 segments. Second, we made an alignment at the sentence level by using the *Vanilla* aligner. We obtained 6,357 sentences. Each sentence is about 8 times shorter than an article.

For each aligned corpus, we generated a bilingual lexicon of single words tagged as nouns. The English corpus was tagged using TreeTagger[3], whereas we used FreeLing[4] to tag the Spanish text. The strategy to select the best translation equivalent was the following. For an English word, we selected the Spanish word with the highest Dice coefficient if and only if this coefficient is not lower

---

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[4] http://garraf.epsevg.upc.es/freeling/

than an empirically set threshold (0.3). Notice that, in this work, as we are not interested in evaluating the extraction strategy itself but only the input data (many and short against few and long segments), we used a very simple and intuitive extraction algorithm. In [6], we defined a more complex and efficient method to extract translation equivalents which takes into account word polysemia.

To evaluate the results, *precision* is characterised as the ability of the system to assign the correct translation equivalent to a word type in the dictionary. Let's note that the evaluation is made at the word type level. So, this *precision* does not consider the fact that a word can have several senses in different contexts. A translation proposed by the system is considered as correct if only if it is the correct translation in at least one particular context. *Recall* is the number of equivalents proposed by the system divided by the total number of word types found in the text. *F-score* is the harmonic mean between recall and precision. Table 3 shows that the fact of increasing the number of segments (by making their length shorter) has two effects: on the one hand, it improves precision, but, on the other, it makes recall lower. Notice that the f-score values obtained with the longest segments (speakers in EuroParl and articles in the Constitution) are higher than those obtained with small segments (paragraphs and sentences). These results seem to show that the mean between precision and recall (the f-score) is more balanced when the lexicon extraction is performed using fewer segments with a larger size, as in natural alignment.

Partitions	Segments	Precision	Recall	F-Score
EuroParl (speakers)	3,000	.77	.30	.43
EuroParl (paragraphs)	12,000	.85	.23	.32
Constit. (articles)	881	.69	.33	.45
Constit. (sentences)	6,357	.76	.29	.42

Table 3. Evaluation of two parallel corpus aligned in two different ways.

5.2 Second Experiment

To check this idea, we made a more reliable experiment. We selected an excerpt from the English-Portuguese parallel corpus built within the Translation Project[5]. In order to tag the Portuguese corpus, we parametrised TreeTagger for Portuguese using the training corpus built by the NILC group ("Núcleo Interinstitucional de Linguística Computacional). As lexical resource, we made use of

---

[5] This project aims at translating the Linux manuals to several languages. See more details in http://www.iro.umontreal.ca/translation/HTML/index.html

the "Léxico Multifuncional Computorizado do Português Contemporáneo[6]. The parametrised file for Portuguese is freely available[7].

The English-Portuguese parallel corpus we selected consists of 165,000 English word tokens and 28,000 manually aligned sentences. This corpus was splitted in 5 incremental partitions, where partition n is a subset of partition $n + 1$. The smallest partition represents only the 7% of the whole corpus; it consists of 2,000 segments. The second one is larger (18%) and consists of 5,000 segments. The third partition (36%) contains 10,000 segments. The forth (71%) has 20,000 segments, and finally, the firth partition is the whole corpus, that is, 28,000 aligned sentences. Table 4 depicts the precision, recall, and f-score achieved on these 5 partitions.

Partitions	Segments	Precision	Recall	F-Score
1 (7%)	2,000	.84	.17	.28
2 (18%)	5,000	.87	.21	.34
3 (36%)	10,000	.90	.23	.37
4 (71%)	20,000	.91	.19	.31
5 (100%)	28,000	.89	.19	.31

Table 4. Evaluation of the 5 partitions extracted from the parallel corpus built by the Translation Project. Each aligned segment consists of 1 sentence.

Later, this corpus was also aligned in 4,000 segments, where each segment is the result of putting 7 sentences together. In this experiment, the specific goal was to build an aligned corpus with few and long segments, as those usually generated by natural alignment. This aligned corpus was also splitted in 5 incremental partitions. The smallest one consists of 500 segments (12% of the corpus). The second contains 1,000 (25%). The third and the forth consists of 2,000 and 3,000, respectively (50% and 75%). Finally, the largest partition embraces all the 4,000 segments. Table 5 shows the evaluation scores obtained on these partitions.

Partitions	Segments	Precision	Recall	F-Score
1 (12%)	500	.83	.20	.32
2 (25%)	1,000	.83	.22	.35
3 (50%)	2,000	.89	.24	.38
4 (75%)	3,000	.89	.19	.31
5 (100%)	4,000	.88	.20	.33

Table 5. Evaluation of the 5 partitions extracted from the parallel corpus built by the Translation Project. Each aligned segment consists of 7 aggregated sentences.

---

[6] The trained corpus is available at http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html. The lexicon can be downloaded at http://www.clul.ul.pt/sectores/projecto_lmcpc.html.

[7] http://gramatica.usc.es/∼gamallo/tagger.htm.

As in the experiments performed on both Europarl and the European Constitution, the alignment based on long segments entails a lower precision (.88 against .89 in the whole corpus), whereas recall is slightly better: .20 against .19. It results in a higher f-score: .33 against .31. These results confirm that the use of longer aligned segments provides a more balanced relationship between precision and recall than that produced by the alignment at the sentence level.

On the other hand, these experiments also show that precision becomes stable when the corpus partition reaches the 50% of the whole corpus (see Figure 1). If we use sentence alignment, precision stabilises when the partition contains $10,000 - 20,000$ aligned segments. Yet, if alignment is made with longer portion of texts (as in natural alignment), then stability is achieved using less segments: $2,000$. In both experiments, given a specific corpus size, the fact of increasing the number of segments does not improve precision in a significant way.

However, these experiments are not enough to discover other relevant information. In particular, we need to perform further experiments to find the relationship between the size of aligned segments and the number of segments that are required to precision be stabilised.

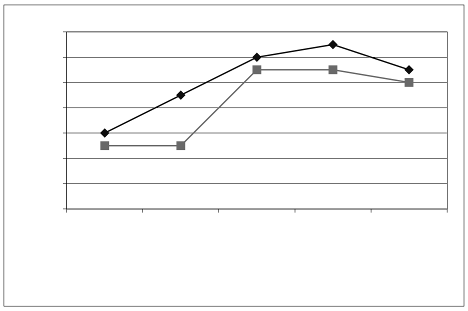

Fig. 1. Variation of precision as a function of corpus size.

6 Conclusions and Further Research

The experiments described above lead us to conclude that it is not necessary to align parallel corpora at the sentence level if the aim is only to extract bilingual translation equivalents. For this specific purpose, what we need is to identify natural boundaries and use the particular version of the Dice coefficient defined above. This type of strategy is robust to noise (insertions and deletions) in the parallel corpus and produces results that are at least as useful as those obtained by sentence alignment. Moreover, the bilingual lexicon extracted using our method could be considered as a set of reliable anchor points if the aim is to incrementally search for smaller aligned segments (until reach the word level).

Taking into account these conclusions and the evaluation results, nothing prevents us from building a huge amount of parallel texts whose basic unit of alignment is a book or a novel. For instance, consider that we are provided with a huge electronic library containing thousands of old Spanish novels and their English translations[8]. Then, a computationally realistic task will be to use our strategy to extract a bilingual lexicon taking as basic segment each single novel and its translation. Since such an aligned corpus is considerably bigger than those corpora aligned at the sentence level, the bilingual lexicon extracted from this enormous parallel corpus could reach an interesting coverage. As soon as big digital libraries are available, it will be possible to perform this type of experiment.

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 (COLING-ACL'98)*, pages 29–35, Montreal, 1998.
2. P.F. Brown, J. Lai, and R. Mercer. Aligning sentences in parallel corpora. In *29th Conference of ACL*, 1991.
3. K. Church. Char_align: A program for aligning parallel texts at the character level. In *31st Conference of the Association for Computational Linguistics (ACL)*, pages 1–8, Columbus, Ohio, 1993.
4. Pascale Fung and Kathleen McKeown. Finding terminology translation frmo non-parallel corpora. In *5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong, 1997.
5. Willian Gale and Kenneth Church. Identifying word correspondences in parallel texts. In *Workshop DARPA SNL*, 1991.
6. Pablo Gamallo. Extraction of translation equivalents from parallel corpora using sense-sensitive contexts. In *10th Conference of the European Association on Machine Translation (EAMT'05)*, pages 97–102, Budapest, Hungary, 2005.
7. Philipp Koehn. Europarl: A multilingual corpus for evaluation of machine translation. In *http://people.csail.mit.edu/koehn/publications/europarl/*, 2003.
8. 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.
9. Dan Melamed. A word-to-word model of translational equivalence. In *35th Conference of the Association of Computational Linguistics (ACL'97)*, Madrid, Spain, 1997.
10. Dan Melamed. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1), 1999.
11. A. Ribeiro, G. Dias, G. Lopes, and J. Mexia. Cognates alignment. In *Machine Translation Summit VIII*, pages 287–293, Santiago de Compostela, Spain, 2001.
12. A. Ribeiro, G. Lopes, and J. Mexia. Using confidence bands for parallel texts alignment. In *38th Conference of the Association for Computational Linguistics (ACL)*, pages 432–439, 2000.

---

[8] Some promising projects, such as the Google Library and the Gutenberg Project, are working to collect electronic books without Copyright in free libraries.

13. M. Simard and P. Plamondon. Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation*, 13(1):59–80, 1998.
14. F. Smadja, K. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons. *Computational Linguistics*, 22(1), 1996.
15. Jorg Tiedemann. Extraction of translation equivalents from parallel corpora. In *11th Nordic Conference of Computational Linguistics*, Copenhagen, Denmark, 1998.
16. Ŝ. Vintar. Using parallel corpora for translation-oriented term extraction. *Babel Journal*, 47(2):121–132, 2001.