Extraction of Translation Equivalents from Parallel Corpora Using Sense-sensitive Contexts

Pablo Gamallo Otero

Departamento de Língua Espanhola
Universidade de Santiago de Compostela
pablogam@usc.es

Abstract. The paper proposes an unsupervised method to extract translation equivalents from parallel corpora. The strategy we use takes into account the context of words. Given a word of the source language and a particular context, we learn its word translation within an equivalent context. We first extract pairs of similar contexts and, then, we compare the similarity between words appearing in each pair. This allows us to use a very low threshold to identify correct translation equivalents. Moreover, as polysemic words tend to have different senses in different context pairs, we are able to associate several translation equivalents to the same polysemic word. The main contribution of this paper is precisely to learn the correct translation equivalent of a word in a specific context. On the other hand, we do not align texts by detecting sentences or other small linguistic units. We identify natural boundaries by detecting explicit parts or segments of the corpus. Most text corpora contain natural boundaries to explicitly separate basic parts such as chapters, articles, receipts, legal documents, letters, etc. We use these explicit and natural parts to align parallel corpora. To compute similarity within these large segments, we define a particular version of the Dice coefficient.

1. Introduction

Parallel corpora are a huge resevoir of bilingual lexical information. The extraction of translation lexicons relies, to a certain extent, on parallel text alignment often to the sentence level. Most extraction methods use the co-occurrence frequencies and locations of expressions in aligned sentences to compute the translation correlations between expression types in the two languages (Gale and Church, 1991; Melamed, 1997; Ahrenberg and Andersson 1998; Tiedemann, 1998; Vintar, 2001; Kwong, 2004). These methods have at least two drawbacks: first, although a number of automatic sentence alignment methods have been proposed, they are not very reliable when corpora have unclear sentence boundaries or with noisy bilingual texts (Fung and McKeown, 1996). Second, as these methods are mostly based on one-to-one word translations, they are not adapted to dealing with word polysemia.

To avoid the first drawback, we use a type of aligment based, not on sentence identification, but on identifying natural boundaries by detecting the explicit parts of the corpus. Most text corpora contain natural boundaries to explicitly separate basic parts such as chapters, articles, receipts, legal documents, letters, etc. We will use these explicit and natural parts to align parallel corpora. Regarding the second drawback (i.e., polysemia), we will propose a method that takes into account word sense disambiguation in context. More precisely, the extraction method we propose in this paper aims at identifying bilingual correspondences between sense-sensitive contexts. Equivalent sense-sensitive contexts will allow us to extract correspondences between types of word senses. This is the main contribution of our work. Moreover, sense-sensitive contexts will also be used to generate multi-word translations in a compositional way.

The paper is organized as follows. First, section 2 describes pre-processing and natural align-

ment. Then, section 3 introduce the measure used to compute translation correlations between expression types in the two languages. Section 4 will define the notion of *sense-sensitive context*. In section 5, we will present the extraction algorithm. And finally, in section 6, a evaluation protocol will be outlined.

2. Pre-processing the Corpus

First, the texts of both languages are tokenized, lemmatized and tagged using TreeTagger (Schmid 2002). No manual correction was made on the tagged texts. So, the bilingual lexicon extractor will inherit errors caused by the tagger. Then, the texts are superficially parsed by simple pattern matching, where the objective is to extract sense-sensitive contexts of words. In section 4, we will explain the notion of sense-sensitive context.

The following pre-processing step is to align the source and target texts by detecting natural boundaries such as chapters, specific documents, articles, etc. Our experiments were made on a corpus constituted by the English and French versions of the European Legislation in Force. The natural boundaries used to divide this corpus are the beginning and the end of legal documents such as agreements, directives, and regulations. We detected 1,050 English legal documents and their corresponding French translations. Each pair of legal document constituted by the English source and its translation is considered as an aligned segment. The main drawback of this type of alignment is that it only allows to select very large 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 corpus. And second, this alignment does not need manual correction.

3. A Particular Version of the Dice Coefficient

We estimate the probability that a candidate target expression is a translation by counting both occurrences of expressions in the corpus as a whole and co-occurrences of the expressions within pairs of aligned segments.

Following (Smadja and McKeown, 1996), we selected the Dice coefficient to measure translation correlations. Given a source expression type e_1 and a candidate translation e_2, our particular version of the Dice coefficient is defined as follows:

$$Dice(e_1, e_2) = \frac{2 * F(e_1, e_2)}{F(e_1) + F(e_2)}$$

where

$$F(e_1, e_2) = \sum_i \min(f(e_1, s_i), f(e_2, s_i))$$

and

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

Note that $f(e_n, s_i)$, represents the frequency of the expression type e_1 occurring in segment s_i. Unlike most approaches to bilingual lexicon extraction, we consider that the frequency of an expression 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, an expression type of the target language, e_2, is likely to be a translation of a source expression, e_1, if both expression tend to have a similar frequency in each segment s_i. This is an important difference with regard to standard approaches. In most approaches, two expressions are linked if they tend to appear in the same aligned segments. However, as in our approach many different expressions can appear in all segments, we need a more informative feature, namely the number of times an expression appears in each segment.

4. Sense-sensitive Contexts

The main contribution of this paper is to use sense-sensitive contexts to extract word translations. Consider the following two expressions:

1. vehicle registration
2. registration of the notification

Expression (1) is translated into French as *"immatriculation du véhicule"*, whereas (2) is translated as *"enregistrement of the notification"*. In (1), the modifier *"vehicle"* behaves as a sense-sensitive context that selects a specific sense of *"registration"*: a part of a vehicle. Note that this sense is slightly different from that selected in the context introduced by "of the notification" in expression (2), where *"registration"* refers to a specific action. The two sense-sensitive contexts of "registration'" in expressions (1) and (2) are:

<vehicle [NOUN]>
<[NOUN] of the notification>

These contexts seem to be useful to distinguish particular word senses. Note that a sense-sensitive context can be one of the two positions underlying any Head-Modifier dependency. Given a binary dependency, for instance:

of (registration, notification)

where *"registration"* is the head word and *"notification"* is its modifier, we can select two sense-sensitive contexts:

<registration of [NOUN]>
<[NOUN] of the notification>

So, we consider that not only the modifier can select a sense of the head but also the head can select for a particular sense of the modifier (Pustejovsky, 1995; Gamallo, 2005).

The extraction algorithm we will outline in the following section is focused on the identification of bilingual correspondences between sense-sensitive contexts.

5. The Algorithm

The algorithm is divided in three steps. First, bilingual links between sense-sensitive contexts are extracted. In the second step, the information learnt in the first step is used to extract links between single words, both monosemic and polysemic words. And third, using the information extracted in the previous steps, we generate multi-word translation equivalents.

5.1. Step 1: Extracting Bilingual Links between Sense-sensitive Contexts

Using appropriate syntactic patterns, a large list of sense-sensitive contexts is selected for each of the two compared languages. The patterns used in our experiments were:

NOUN - PREP - NOUN
NOUN – NOUN
ADJ - NOUN
NOUN – ADJ

Then, we compute the Dice score between pairs of contexts. Each context of the source language will be linked to those contexts of the target language whose Dice coefficient is larger than *50%*. Some examples of bilingual links between contexts are given in Table 1, Table 2, and Table 3.

ENGLISH	***FRENCH***	**SIM**
<[NOUN] acetate>	*<acetate de [NOUN]>*	*0.66*
<[NOUN] activation>	*<activation de [NOUN]>*	*0.82*
<[NOUN] air>	*<air de [NOUN]>*	*0.74*
<[NOUN] alignment>	*<alignement de [NOUN]>*	*0.64*
<[NOUN] alloy>	*<alliage de [NOUN]>*	*0.81*
<[NOUN] aluminium>	*<alluminium de [NOUN]>*	*0.92*
<[NOUN] anchorage>	*<ancrage de [NOUN]>*	*0.62*
<[NOUN] atlantic>	*<atlantique de [NOUN]>*	*0.53*

Table 1. An excerpt of correlations between sense-sensitive contexts extracted from the English pattern NOUN–NOUN.

ENGLISH	***FRENCH***	**SIM**
<competition between [NOUN]>	*<concurrence entre [NOUN]>*	*0.73*
<difference between [NOUN]>	*<différence entre [NOUN]>*	*0.71*
<distance between [NOUN]>	*<distance entre [NOUN]>*	*0.52*
<trade between [NOUN]>	*<échange entre [NOUN]>*	*0.70*
<[NOUN] between belgium>	*<[NOUN] entre belgique>*	*0.87*
<[NOUN] between government>	*<[NOUN] entre gouvernement>*	*0.84*
<[NOUN] between manufacturer>	*<[NOUN] entre fabricant>*	*0.69*
<[NOUN] between producer>	*<[NOUN] entre producteur>*	*0.63*
<coal in [NOUN]>	*<charbon dans [NOUN]>*	*0.62*
<cognac in [NOUN]>	*<cognac en [NOUN]>*	*0.84*
<conduct in [NOUN]>	*<comportement sur [NOUN]>*	*0.62*
<convention in [NOUN]>	*<convention en [NOUN]>*	*0.72*
<preparation of [NOUN]>	*<préparation de [NOUN]>*	*0.63*
<principle of [NOUN]>	*<principe de [NOUN]>*	*0.76*
<profitability of [NOUN]>	*<rentabilité de [NOUN]>*	*0.73*
<promotion of [NOUN]>	*<promotion de [NOUN]>*	*0.61*
<protection of [NOUN]>	*<protection de [NOUN]>*	*0.61*
<province of [NOUN]>	*<province de [NOUN]>*	*0.92*
<provision of [NOUN]>	*<disposition de [NOUN]>*	*0.53*

Table 2. An excerpt of correlations between sense-sensitive contexts extracted from the English pattern NOUN–PREP–NOUN.

ENGLISH	*FRENCH*	SIM
<legal [NOUN]>	*<[NOUN] juridique>*	*0.51*
<legitimitate [NOUN]>	*<[NOUN] legitime>*	*0.85*
<light [NOUN]>	*<[NOUN] léger>*	*0.75*
<linear [NOUN]>	*<[NOUN] linéaire>*	*0.90*
<liquide [NOUN]>	*<[NOUN] liquide>*	*0.68*
<local [NOUN]>	*<[NOUN] local>*	*0.57*
<long-term [NOUN]>	*<long [NOUN]>*	*0.79*
<longitudinal [NOUN]>	*<[NOUN] longitudinal>*	*0.65*

Table 3. An excerpt of correlations between sense-sensitive contexts extracted from the English pattern ADJ–NOUN.

Bilingual links between sense-sensitive contexts will be used to extract word links in the second step of the algorithm.

Note that these links can be viewed as "translation templates". Each translation template contains two components that were generalized by replacing them with two bound variables: both NOUN and NOUN in the two expressions (Güvenir & Cicekli, 1998). However, unlike most approaches in Example-Based Machine Translation (known as EBMTs), we do not induce translation templates from example translations. Templates (i.e., correlations of sense-sensitive contexts) are extracted by comparing pairs of contexts using the Dice coefficient.

5.2. Step 2: Extracting Bilingual Links between Single Words

This step consists of two processes: the extraction of links between monosemic words and the extraction of links between potentially polysemic words.

- **First Process:** We assume that a word link with an association score higher than *80%* shows that the two compared words are monosemic throughout the training corpus. So, we consider that word links with the highest Dice values can be perceived of as one-to-one links (i.e., there is at most one link for each source language word). As in the *competitive linking approach* proposed by Melamed (Melamed, 1996), the extracted links are removed from the search space before starting a new extraction process.
- **Second Process**: The search space of the second extraction process is now constituted by those words that both have not been considered as monosemic and appear in pairs of equivalent sense-sensitive contexts. Every word occurring in a context of the source language is compared with all words occurring in the corresponding equivalent context of the target language. As the candidates to be word translations are chosen from sense-sensitive contexts, we assume that it is possible to reduce the threshold in a significant way. More precisely, we consider that a target language word is likely to be a context-sensitive translation of a source language word if the pair of words are at least *40%* similar. This method allows us to identify several-to-several word links, i.e., word translations for words that are potentially polysemous. For instance, *"fuel"* should be translated either as *"carburant"* or *"combustible"*. As the Dice coefficient assigns only *50%* score to the link *"fuel"/"carburant"* and *48%* to *"fuel"/"combustible"*, these links were not selected in the first process. However, in the second process, *"fuel"* can be linked to *"carburant"* because this is the highest association score between *"fuel"* and any French word in the context pair:

ENGLISH	*FRENCH*	SIM
<aid on [NOUN]>	*<aide sur [NOUN]>*	*0.66*

By contrast, the link *"fuel"/"combustible"* is the highest link between *"fuel"* and any French word in a slightly different type of context pair:

ENGLISH	*FRENCH*	SIM
<nuclear [NOUN]>	*<[NOUN] nucléaire>*	*0.66*

Likewise, the French word *"président"* is translated as *"chairman"* when the two words co-occur in:

ENGLISH	*FRENCH*	SIM
<[NOUN] of commitee>	*<[NOUN] de commité>*	*0.45*

whereas it is translated as *"president"* in contexts such as:

ENGLISH	*FRENCH*	SIM
<[NOUN] of council>	*<[NOUN] de conseil>*	*0.49*

5.3. Step 3: Generating Links between Multi-Words

Given a set of links between equivalent sense-sensitive contexts (step 1) and a set of links between equivalent word types (step 2), we can easily generate equivalent translations of composite expressions. Let's suppose that each pair of equivalent contexts, for instance, *<aid on [NOUN]> / <aide sur [NOUN]}>*, is associated with a list of word links: e.g., *"fuel"/"carburant"*, *"export"/"exportation"*, etc. So, we can generate the following correspondences:

"aid on fuel"/"aide sur carburant"
"aid on export"/"aide sur exportation"

Note that this method is fully compositional. It could be easily used to generate new translation equivalents from non-parallel corpus.

6. Experiments and Evaluation

The algorithm was tested on an English and French parallel corpus containing over 2 million token words, which was built from the Legislation in Force of the European Commission. The corpus consists of 1,050 aligned segments. Each segment is a quite long document containing a rate of 2,000 words.

The extraction algorithm allowed us to compile a bilingual dictionary containing 1,598 word entries (lemmas of nouns and adjectives), 1,797 bilingual entries of sense-sensitive contexts, and 4,393 entries of multi-words.

In order to evaluate our extraction method, a test corpus of 150 English words which were tagged either as a noun or adjective was selected at random. Each selected word was extracted with its immediate context in order to allow evaluators to make appropriate decisions. The results are summarized in Table 4. We call *Precision* the number of correct translations proposed by the system divided by the number of all translations which has been suggested. *Recall* is the number of correct translations proposed by the system divided by the number of all test instances.

	Precision	***Recall***
NOUNS	0.94	0.78
ADJECTIVES	0.93	0.66
TOTAL	0.94	074

Table 4. Results concerning the equivalent translations of nouns and adjectives.

In Table 5, precision was measured at two different threshold values. The *>= 80%* threshold is the value we use to identify what we call monosemic words. At this level, the system achieves over *97%* precision. Potentially polysemic words are extracted using the threshold situated between *40* and *80%* Dice similarity. Given that we only compare words appearing in equivalent sense-sensitive contexts, precision remains still quite high. That is, equivalent sense-sensitive contexts allow us to find correct word translations with low Dice values.

Threshold	***Precision***
>= 80%	0.97
>= 40% <= 80%	0.92

Table 5. Precision of monosemic and polysemic words.

A protocol to evaluate multi-words was not defined yet. As in (Fung and Mckeown, 1996), we aim at designing a way to measure the increase in efficiency that can be observed when translators are using a particular list of equivalent translations.

Although the experiments herein are on English and French, we believe the model is equally applicable to other language pairs.

7. Conclusions

In this paper, we have described a method to retrieve translation equivalents from bilingual corpora aligned at a very row level. In our experiments, the aligned segments were documents containing between 1 and 50 Kbytes. To overcome the problems caused by such large segments, we used a similarity coefficient, which takes into account frequency of expressions in each segment.

On the other hand, we dealt with polysemy by comparing word similarity within sense-sensitive contexts. For this purpose, we first extract pairs of similar contexts and, then, we compare the similarity between words appearing in each pair. This allows us to use a very low threshold to identify correct translation equivalents. Moreover, as polysemic words tend to have different senses in different context pairs, we are able to associate several translation equivalents to the same polysemic word. The main contribution of this paper is precisely to learn

the correct translation equivalent of a word in a specific context.

Finally, in future work, we aim at dealing with correlations between multi-words and single words. The main drawback of our current work is to not learn such correlations.

8. Acknowledgements

This work has been supported by Ministerio de Educación y Ciencia, within the project Garicoter, ref: HUM2004-05658-C02-02, and by Xunta de Galicia, by means of the program Isidro Parga Pondal.

9. References

AHRENBERG, M. Anderson, and MERKEL, M. (1998). 'A simple hybrid aligner for generating lexical correspondences in paralell texts'. In COLING-ACL'98, Montreal, pages 29-35.

FUNG, Pascale, and McKEOWN, Kathleen (1996). 'A technical word and term translation aid using noisy parallel corpora across language groups'. Machine Translation Jounal, pages 53-87.

GALE, Willian A., and CHURCH, Kenneth W. (1991). 'Identifying Word Correspondences in Parallel Texts'. In DARPA Speech and Natural Language Workshop, California, pages 152-157.

GAMALLO, Pablo, AGUSTINI, Alexandre, and LOPES, G. (2005). 'Clustering sintactic positions with similar syntactic requirements'. Journal of Computational Linguistics, 31(1).

GÜVENIR, H. Altay, and CICEKLI, Ilyas. (1998). 'Learning Translation Templates from Examples'. Information Systems, 23, pages 353-363.

KAJI, Hiroyuki, and AIZONO, Toshiko (1996). 'Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information'. In COLING'96, pages 23-28.

KWONG, O., TSOU, B., and LAI, T. (2004). 'Alignment and extraction of bilingual legal terminology frmo context profiles'. Terminology, 10(1), pages 81-99.

MELAMED, Dan (1997). 'A word-to-word model of translational equivalence'. In ACL'97, Madrid, Spain.

SCHMID, Helmut (2002). 'A language independent part-of-speech tagger'. In http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

SMADJA, F., McKEOWM, K., and HATZIVASSILOGLOU (1996). 'Translating collocations form bilingual lexicons'. Journal of Computational Linguistics, 22(1).

TIEDEMANN, Jorg (1998). 'Extraction of translation equivalents from parallel corpora'. In 11th Nordic Conference of Computational Linguistics.

VINTAR, Spela (2001). 'Using parallel corpora for translation-oriented term extraction'. Babel Journal, 47(2).