An Approach to Acquire Word Translations from Non-parallel Texts*

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

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

Abstract. Few approaches to extract word translations from non-parallel texts have been proposed so far. Researchers have not been encouraged to work on this topic because extracting information from non-parallel corpora is a difficult task producing poor results. Whereas for parallel texts, word translation extraction can reach about 99%, the accuracy for non-parallel texts has been around 72% up to now. The current approach, which relies on the previous extraction of bilingual pairs of lexico-syntactic templates from parallel corpora, makes a significant improvement to about 89% of words translations identified correctly.

1 Introduction

In the last decade, many works have been carried out to automatically extract word and/or multi-word translations from bilingual parallel corpora [15, 1, 20, 13]. These works share a common strategy: they perform first the alignment of segments and then, on the basis of such an alignment, they compute word correspondences in each pair of segments. In some of these experiences, word-level translation accuracy achieved very high scores: about 99%. Unfortunately, the amount of available bilingual parallel texts is still small, specially in specific, academic or technological domains. Given a particular knowledge domain, the number of parallel texts is much lower than that of monolingual texts. This type of texts are known as comparable, *non-parallel* corpora. Nowadays, in the World Wide Web, comparable non-parallel corpora are more prevalent than parallel corpora. However, the highest rate to date in word-level translation from non-parallel corpora is relatively small, 72% [18], in comparison to the accuracy rate achieved from parallel corpora.

In this paper, we present a method to extract word translations, which takes advantage from the positive aspects of both parallel (high accuracy) and non-parallel corpora (high coverage). The strategy we propose is the following.

* This work has been supported by Ministerio de Educacin y Ciencia of Spain, within the project GARI-COTERM, ref: HUM2004-05658-D02-02.

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 600–610, 2005.

We first extract, from small parallel texts, a representative set of bilingual correspondences between unambiguous lexico-syntactic templates. Then, these pairs of bilingual templates are used as *local contexts* to extract word translations from comparable, non-parallel corpora. A word $w1$ in the source language can be the translation of a word $w2$ in the target language if $w1$ tends to occur in local contexts that are translations of the local contexts $w2$ occurs in. Using this method, we achieved 89% accuracy, which is close to the scores reached by the extraction approaches from aligned, parallel texts. In the following sections, we first recall previous approaches to word translations from non-parallel text, then we describe our own method, and finally, we provide and discuss two different experimental results.

2 Related Work

There are few approaches to extract bilingual lexicons from non-parallel corpora in comparison to those using a strategy based on aligned, parallel texts. The most efficient method to extract word translations from comparable, non-parallel corpora is described in [6, 7, 18]. The starting point of this strategy is as follows: word w_1 is a possible translation of w_2 if the words with which w_1 co-occurs within a particular window are translations of the words with which w_2 co-occurs within the same window. This strategy relies on a list of bilingual word pairs (*seed words*) provided by an electronic bilingual dictionary. w_1 is a candidate translation of w_2 if they tend to co-occur with the same seed words. There are three drawbacks to this method. First, it is not a knowledge-poor approach since it needs external lexical resources such as a bilingual dictionary. Second, not all the words of the dictionary seem to be reliable seed words. Polysemic words should be removed from the list of seed words since they may introduce semantic noise. And third, according to the Harris's hypothesis [12], counting co-occurrences within a window of size N is less precise than counting co-occurrences within local syntactic contexts. In the most efficient approaches to thesaurus generation [11, 14], word similarity is computed using co-occurrences between words and specific syntactic contexts. Syntactic contexts are less ambiguous and more sense-discriminative than contexts defined as windows of size N. In order to overcome these drawbacks, we will use as seed expressions, not word entries from an external bilingual dictionary, but pairs of bilingual lexico-syntactic templates which were extracted from a small parallel corpus. As the lexico-syntactic templates represent unambiguous local contexts of words, they are discriminative and confident seed expressions to extract word translations from non-parallel texts.

There exist other approaches to bilingual lexicon extraction which do not use external bilingual dictionaries [5, 17, 4]. Yet, [5] failed to reach an acceptable accuracy rate for actual use, [17] had strong computational limitations, and [4] was applied only to non-parallel texts in the same language. On the other hand, [3] describes a particular strategy based on a multilingual thesaurus instead of a external bilingual dictionary.