Learning Spanish-Galician Translation Equivalents using a Comparable Corpus and a Bilingual Dictionary

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
[2] Departamento de Tecnologia Linguística da Imaxin|Software
Santiago de Compostela, Galiza

Abstract. So far, research on extraction of translation equivalents from comparable, non-parallel corpora has not been very popular. The main reason was the poor results when compared to those obtained from aligned parallel corpora. The method proposed in this paper, relying on *seed patterns* generated from external bilingual dictionaries, allows us to achieve similar results to those from parallel corpus. In this way, the huge amount of comparable corpora available via Web can be viewed as a never-ending source of lexicographic information. In this paper, we describe the experiments performed on a comparable, Spanish-Galician corpus.

1 Introduction

There exist many approaches to extract bilingual lexicons from parallel corpora [8, 16, 1, 22, 14]. These approaches share the same basic strategy: first, bitexts are aligned in pairs of segments and, second, word co-ocurrences are computed on the basis of that alignment. They usually reach high score values, namely about 90% precision with 90% recall. Unfortunately, parallel texts are not easily available, in particular for minority languages. To overcome this drawback, different methods to extract bilingual lexicons have been implemented lately using non-parallel, comparable corpora. These methods take up with the idea of using the Web as a huge resource of multilingual texts which can be easily organized as a collection of non-parallel, comparable corpora. A non-parallel and comparable corpus (hereafter "comparable corpus") consists of documents in two or more languages which are not translation of each other and deal with similar topics. However, the accuracy scores of such methods are not as good as those reached by the strategies based on aligned parallel corpora. So far, the highest values have not improved an 72% accuracy [18], and that's without taking into consideration the coverage of the extracted lexicon over the corpus.

This paper proposes a new method to extract bilingual lexicons from a POS tagged comparable corpus. Our method relies on the use of a bilingual dictionary to identify bilingual correlations between pairs of lexico-syntactic patterns. Such