Distributional Thesaurus vs. WordNet: A Comparison of Backoff Techniques for Unsupervised PP Attachment*

Hiram Calvo,[1] Alexander Gelbukh,[1] and Adam Kilgarriff [2]

[1] Center for Computing Research, National Polytechnic Institute, Mexico
hcalvo@sagitario.cic.ipn.mx, www.Gelbukh.com

[2] Lexical Computing Ltd., United Kingdom
adam@lexmasterclass.com

Abstract. Prepositional Phrase (PP) attachment can be addressed by considering frequency counts of dependency triples seen in a non-annotated corpus. However, not all triples appear even in very big corpora. To solve this problem, several techniques have been used. We evaluate two different backoff methods, one based on WordNet and the other on a distributional (automatically created) thesaurus. We work on Spanish. The thesaurus is created using the dependency triples found in the same corpus used for counting the frequency of unambiguous triples. The training corpus used for both methods is an encyclopaedia. The method based on a distributional thesaurus has higher coverage but lower precision than the WordNet method.

1 Introduction

The Prepositional Phrase (PP) attachment task can be illustrated by considering the canonical example *I see a cat with a telescope*. In this sentence, the PP *with a telescope* can be attached to *see* or *cat*. Simple methods based on corpora address the problem by looking at frequency counts of word-triples or dependency triples: *see with telescope* vs. *cat with telescope*. In order to find enough occurrences of such triples, a very large corpus is needed. Such corpora are now available, and the Web can also be used [4, 27]. However, even then some combinations of words do not occur. This is a familiar effect of Zipf's law: few words are very common and there are many words that occur with a low frequency [14], and the same applies to word combinations.

To address the problem, several backoff techniques have been explored. In general, 'backing off' consists of looking at statistics for a set of words, when there is insufficient data for the particular word. Thus *cat with telescope* turns into ANIMAL *with* INSTRUMENT and *see with telescope* turns into *see with* INSTRUMENT (capitals denote sets of instrument-words, animal-words, etc.) One way to identify

* Work done under partial support of Mexican Government (CONACyT, SNI, PIFI-IPN, CGEPI-IPN) and RITOS-2.