LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

Using Unsupervised Word Sense Disambiguation to Guess Verb Subjects on Untagged Corpora

PAULA CRISTINA VAZ
DAVID MARTINS DE MATOS
Spoken Language Laboratory, INESC-ID-Lisboa, Portugal

ABSTRACT

This article explores the use of subject lists extracted from an annotated corpus to find subject-verb pairs in untagged corpora. Our goal is to identify verb syntactic functions (subjects and direct objects) to characterize verb arguments. Since identifying syntactic functions on corpora using parsers is time-consuming, it is desirable to automate the annotation process of the syntactic functions without parsing the corpus. We present a method that uses a small annotated corpus to cluster sentences with synonymous verbs. We observe that verbs in the same cluster have the same list of nouns as subject in the test corpus, even though the specific pair subject/verb does not appear in the annotated corpus. The result shows that annotating the subject/verb pair using the subject lists extracted from the clusters is quicker than syntactically parsing the corpus.

1. INTRODUCTION

Our goal is to find nouns that can be used as subjects or objects using a small annotated manually-corrected corpus and without having to parse large amounts of corpora. Syntactically parsing usually means to lemmatize, disambiguate, chunk or find the parse tree, and, finally, connect the subject, object, and prepositional object with the predicate of the phrase or sentence. After the parsing process is complete, the corpus is ready to be searched for (subject, predicate), (object, predicate), or (preposition, predicate) pairs. Performing all these tasks is time-consuming. It becomes desirable to find a method of extracting syntactic functions without parsing the corpus.

A syntactically annotated corpus for the Portuguese language, Bosque Sintáctico (Afonso et al. 2002), is publicly available. To our knowledge, this corpus is the only one manually-corrected. This corpus is small (180k words) and the number of phrases available for each