

Abstract

Automatic analysis of syntax is one of the core problems in natural language processing. Despite significant advances in syntactic parsing of written text, the application of these techniques to spontaneous spoken language has received more limited attention. The recent explosive growth of online, accessible corpora of spoken language interactions opens up new opportunities for the development of high accuracy parsing approaches to the analysis of spoken language. The availability of high accuracy parsers will in turn provide a platform for development of a wide range of new applications, as well as for advanced research on the nature of conversational interactions. One concrete field of investigation that is ripe for the application of such parsing tools is the study of child language acquisition.

In this thesis, we describe an approach for analyzing the syntactic structure of spontaneous conversational language in parent-child interactions. Specific emphasis is placed on the challenge of accurately annotating the English corpora in the CHILDES database with grammatical relations (such as subject, objects and adjuncts) that are of particular interest and utility to researchers in child language acquisition. This work involves rule-based and corpus-based natural language processing techniques, as well as a methodology for combining results from different parsing approaches. We present novel strategies for integrating the results of different parsers into a system with improved accuracy.

One practical application of this research is the automation of language competence measures used by clinicians and researchers of child language development. We present an implementation of an automatic version of one such measurement scheme. This provides not only a useful tool for the child language research community, but also a task-based evaluation framework for grammatical relation identification.

Through experiments using data from the Penn Treebank, we show that several of the techniques and ideas presented in this thesis are applicable not just to analysis of parent-child dialogs, but to parsing in general.