Robust Parsing, Error Mining, Automated Lexical Acquisition, and Evaluation

Gertjan van Noord
University of Groningen
vannoord@let.rug.nl

Abstract

In our attempts to construct a wide coverage HPSG parser for Dutch, techniques to improve the overall robustness of the parser are required at various steps in the parsing process. Straightforward but important aspects include the treatment of unknown words, and the treatment of input for which no full parse is available.

Another important means to improve the parser's performance on unexpected input is the ability to learn from your errors. In our methodology we apply the parser to large quantities of text (preferably from different types of corpora), and we then apply error mining techniques to identify potential errors, and furthermore we apply machine learning techniques to correct some of those errors (semi-)automatically, in particular those errors that are due to missing or incomplete lexical entries.

Evaluating the robustness of a parser is notoriously hard. We argue against coverage as a meaningful evaluation metric. More generally, we argue against evaluation metrics that do not take into account accuracy. We propose to use variance of accuracy across sentences (and more generally across corpora) as a measure for robustness.