

# **Exploiting Treebanking Decisions for Parse Disambiguation**

by

**Md. Faisal Mahbub Chowdhury**

**M.Sc. Thesis**

**European Masters Program in Language and Communication Technologies**

**September 2009**

## **Abstract**

Treebanks play an increasing role in computational linguistics for training parsers. Some treebanking environments allow annotators to quickly navigate through the parse forest and identify the correct or incorrect or preferred analysis in the current context by selecting or rejecting discriminants. Although, these treebanking decisions are recorded in log files or databases, but, to our best knowledge, until now nobody has inspected potentiality of incorporating such fine-grained decisions made by human annotators for automatic parse disambiguation. This thesis examines this new potential research direction by developing a novel approach for extracting discriminative features using treebanking decisions. The thesis presents comparative analyses of the performance of discriminative disambiguation models built using the treebanking decision features and the state-of-the-art features which indicate features extracted using treebanking decisions are more efficient and informative compared to their traditional counterparts.

We highlight how these different types of features scale when their corresponding models are tested on out-of-domain data. The result suggests that, treebanking decision features are more robust. Analyses from different perspectives such as impact of different types of decisions on the disambiguation model, or using the disambiguation model of the treebanking decisions feature as a re-ranker are also included.

The study also develops a method to extract patterns of correlated discriminant from human decisions and use them for parse forest reduction. The empirical results indicate that, finding such patterns that yields substantial reduction of parse forest preserving the preferred analyses is not an easy task.

The thesis argues that, the discriminative nature of the treebanking decisions allows them to be highly effective features to contribute to an efficient disambiguation model. This is demonstrated by a number of experiments that also reveal some open research questions for future works.