Title: Robust Parsing of Noisy Content

Author: Joachim Daiber

Department: Institute of Formal and Applied Linguistics

Supervisor:
RNDr. Daniel Zeman, Ph.D. (Univerzita Karlova)
Prof. dr. Gertjan van Noord (Rijksuniversiteit Groningen)

Abstract: While parsing performance on in-domain text has developed steadily in recent years, out-of-domain text and grammatically noisy text remain an obstacle and often lead to significant decreases in parsing accuracy. In this thesis, we focus on the parsing of noisy content, such as user-generated content in services like Twitter. We investigate the question whether a preprocessing step based on machine translation techniques and unsupervised models for text-normalization can improve parsing performance on noisy data. Existing data sets are evaluated and a new data set for dependency parsing of grammatically noisy Twitter data is introduced. We show that text-normalization together with a combination of domain-specific and generic part-of-speech taggers can lead to a significant improvement in parsing accuracy.

Keywords: dependency syntax, parsing, domain adaptation