Opponent's review of the Master thesis

Title: Robust Parsing of Noisy Content
Author: Joachim Daiber
Supervisors: RNDr. Daniel Zeman Ph.D., Prof. Dr. Gertjan van Noord

Content

This thesis describes methods for parsing out-of-domain texts, specifically the user-generated content on Twitter. Such noisy texts contain many misspellings and abbreviations and therefore require a preprocessing step, in which they are normalized. The author discusses several techniques how to normalize the sentences and which tagger to use to improve the parsing performance.

The thesis is written in very good English and is 67 pages long (including the initial pages and the bibliography). It is clearly divided into 7 chapters. After a short introduction, the parsing background and the related work is introduced. The third chapter discusses existing data sets and describes development of a new testing data, which was necessary for evaluation of parsers on noisy texts. The annotations consist of two parts: text normalization including the word alignment and dependency structure production on the normalized sentences. The author introduces a special evaluation metric for parsing, which is necessary due to the normalization. The fourth chapter discuss different part-of-speech taggers, their combination, and impact on the parsing performance. The fifth chapter describes two techniques for normalization of noisy texts. The first one is based on unsupervised lexical normalization rewriting out-of-vocabulary words into in-vocabulary words. The second one is based on phrase-based machine translation from noisy to normal text. The last chapter discusses the results, analyses parsing errors and concludes.

Evaluation

The author chose a difficult topic, which required to understand different NLP tasks: tagging, parsing, alignment, and machine translation. He proved to conduct his experiments reasonably and effectively. I appreciate mainly the development of the testing data and its very precise documentation. I think that such user-generated data annotation will be very useful in the future.

Questions

1) Your data was divided into development and test set. However, all the results were measured on the development set only. Did you try to evaluate the final “best” configuration on the test set as well?

2) It seems that you trained the parser on WSJ treebank only. Did you try some kind of self-training on in-domain data? I would also suggest to use a part of annotated data for training. I think that even hundred sentences would be useful and would reduce some errors you came across in your error analysis.

Conclusion

This is a very well written thesis. It is evident that the author have made a plenty of experiments, which led to promising results. I definitely recommend it for the defense.

Prague, August 26, 2013

David Mareček
Institute of Formal and Applied Linguistics
Charles University in Prague