Opponent's review of Master thesis

Title: Exploring Higher Order Dependency Parsers
Author: Pranava Swaroop Madhyastha
Supervisors: RNDr. Daniel Zeman Ph.D., Prof. Michael Rosner

Content

In this thesis, the author describes an approach of improving higher order dependency parsing by incorporating semantic and morphological features. He uses configurable Koo and Collins' implementation of second and third order dependency parser. The experiments are made on Czech and English. For Czech, some positions of the Czech morphological tag are added as individual features. Another experiment incorporates semantic part-of-speech tags from the tectogrammatical layer. For English, the possible senses of each word are fetched from WordNet, an then disambiguated using Ted Pedersen's disambiguation tool, which is based on context similarity measure, another experiment uses synsets. The results are summarized in tables and shows that the morphological and semantic features improves the higher order dependency parsing.

The thesis has 52 pages in total and 40 pages of the pure text. It seems that the thesis was written in a hurry. The theoretical part makes up majority of the text and the experiments are described very briefly and many important things are missing there. The thesis is written in English. At several places it is less understandable. I would also recommend not to write a comma after the word 'that'.

It seems, that the improvement of parsing was done using the gold standard morphological and semantic part-of-speech tags of the testing data, which is useless in practice. In practice, all the annotation of new data is done automatically. It can happen, that the automatically assigned semantic part-of-speech tags will not improve the parsing accuracy at all. Using the automatic disambiguation of WordNet senses for English is correct.

The enclosed CD contains three folders: the parser implemented by Koo and Collins, word-sense disambiguation tool implemented by Ted Pedersen, and a couple of preprocessing and evaluation scripts implemented by the author. Documentation is missing and the README file is insufficient for understanding the scripts.

Questions

It is not clear how much data was used for training and testing. 15000 sentences for training, 1000 for validating and 2000 for testing are mentioned first, but then there is a remark about extracting only thousand sentences from each corpora.

There is not clearly described which features were used in the particular experiments. For example, in the experiment adding individual morphological features of Czech. Were there used only this features? Or also Czech parts-of-speech? Or the full Czech tags?

One experiment adds morphological features of English. Does it mean that only fine-grained tags were used instead of coarse-grained ones?

Was there any tagger used for tagging the testing data? If not, the results are not comparable which previously reported state-of-the-art results. It would be nice to rerun all the experiments with automatically tagged testing data.

Conclusion

This thesis fulfilled the assignment and I recommend it for the defense.

Prague, August 30, 2011

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