Master's thesis review

Title: Joint Learning of Syntax and Semantics
Author: Miloš Ercegovčević
Supervisor: RNDr. Ondřej Bojar Ph.D.

The aim of the thesis is to develop language-independent method for identifying semantic frames of verbs as they are in ProbBank or FrameNet. The author uses a joint model for modeling syntax and semantics and split-merge method to find the best subcategorization of words (or verbs). The results are evaluated across eight languages included in the CoNLL 2009 shared task.

After a short introduction into semantics in Chapter 1, the annotation schemes used in ProbBank and FrameNet are described in Chapter 2. Chapter 3 compares supervised and unsupervised approaches on the semantic role labeling task. Chapter 4 describes a generative modeling of latent probabilistic context-free grammars which are used for modeling semantics. Chapter 5 discuss all the possibilities and introduces the model. The evaluation is summarized in Chapter 6, Chapter 7 discuss a future work, and Chapter 7 concludes.

The thesis is only 30 pages long and in many cases quite unintelligible. Although it seems that there are a lot of possible methods and possibilities discussed, it is not clear which method have been finally used and evaluated. It is even not very clear which features were used for predicting the semantic roles. Are there used only the word forms or the POS-tags as well? Are there uses any kind of previously annotated syntactic or semantic dependencies? How the split-merge method is used and on what? In which manner is the proposed method semi-supervised?

The experimental part (Chapter 6) is only 2 pages long. Experimental setting and the evaluation metric used is not described at all. It only refers the reader to the CoNLL shared task paper. The author mentions that he use also the outputs of a state-of-the art system by Nugues for the prediction, specifically the argument identification, POS tags, dependencies and the labeling of non-verbal predicates. However, the final score of the proposed method is worse than the Nugues' one, which seems like that the proposed method only worsens the annotation produced by Nugues. Can the author comment on this?

It seems to me that the thesis was written in a hurry. There are a lot of mistakes and misspellings. A couple of sentences end abruptly in the middle. Some paragraphs are completely unintelligible. Many mistakes occur in formulas and grammar description in Chapter 4. Many of the variables used in the formulas are not described in the text at all.

The thesis is written in English, however, there have been some Czech labels left (e.g. “Obrázek” or “Literatura”). The Czech abstract seems to be translated from English by GoogleTranslate.

I would recommend mainly to extend the experimental part, to describe exactly the final method so that it can be easily reimplemented by anyone, and of course to revise the text an make it more understandable. This thesis is unfortunately not sufficient for obtaining the Master degree and I recommend it to be rejected.

Prague, September 3, 2012

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