Master's Thesis Review

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

The thesis submitted by Miloš Ercegovčević aims at proposing a method for machine learning of phenomena centered around verbs at the syntax-semantics boundary. The bold goal would be to automatically learn complex structures like FrameNet or PropBank frames at an appropriate and automatically established level of detail and automatically annotate a given corpus with them. In supervised setting, this task is often called *semantic role labeling (SRL)*.

After an introduction, Chapter 2 presents FrameNet and PropBank, two mainstream approaches to lexical semantics. Chapter 3 introduces a supervised and an unsupervised approach to SRL. The formal means that the author will use for SRL in the presented work, namely Latent Probabilistic Context-Free Grammar, is very briefly described in Chapter 4. Chapter 5 provides a broad discussion on supervised and unsupervised methods for various tasks such as finding verb classes, modifier roles, word classes or word sense. The proposed model is also introduced there. Experimental results are in Chapter 6, followed by the discussion of future goals and conclusion.

The second submission of the thesis brings a significant improvement over the first attempt. The main objective has now been described clearer, the notation is explained and most importantly, the input to the PCFG-LA is introduced in detail in Section 5.4.1.

I have two questions to the model:

- Does it ever happen that the parser (given only the string of terminals and preterminals) would construct a nonsensical tree, not following model 1 or model 2 specification, e.g. a tree rooted F? Why (hopefully) not?

- The author says that the model does not use any features (p.47). What is a feature then? The model obviously relies on standard parses of sentences because the input to the model are strings like "*nsubj she blame dobj government prep-for failing*". Why the language-dependent syntactic roles do not constitute a 'feature'?

The experimental evaluation has also been bettered by far. The measured score is at least briefly illustrated. I have three questions:

- The evaluation of the proposed models excludes the sense of the verb. Do also the scores of the baseline/benchmark systems Zhao and Nugues exclude the sense?

- The directory Models on the CD contains scripts for each language separately, and the scripts differ across languages. What are the differences?

- Your evaluation script actually emits a rather detailed report of errors. Since there is no error analysis in the thesis, could you provide us with something interesting that we could learn from the reports?

From the formal point of view, the text has been corrected in many places. A relatively high number of typos or errors in English still remain, e.g. "*does it means*", "*Thus should corresponds to*" on page 28, "*An example sentence we show in Figure 5.12*", "*we need so specify*" on page 34 or *word->world*, *threat->treat* in the Conclusion on page 47. Some regression since the first submission has also happened, e.g. "*cross - -verb*" instead of "*cross-verb*" on page 27; the typesetting in general is far from perfect but unambiguous except perhaps "*verb.01 : A1 A0 AM - LOC*" on page 43 where the spaces should not appear around the dash. The provided CD also contains some confusing formal errors, e.g. two README files where only one corresponds to the actual (abbreviated) content of the CD.

In summary, the thesis brings up a very interesting idea of using PCFG-LA to jointly learn classes of frames and lemmas of verbs, and semantic roles, syntactic realizations and lexical values of their arguments. The idea is evaluated in contrast to supervised semantic role labellers and promises a great potential. The description of the idea as well as the evaluation are unfortunately rather terse and they would deserve both more details as well as a clearer presentation, but in general, the thesis is understandable with some effort.

To conclude, I recommend the thesis to be accepted as a M.Sc. thesis at Charles University in Prague.

Prague, January 17, 2013.

RNDr. Ondřej Bojar, Ph.D.
Charles University in Prague, ÚFAL