Evaluation of doctoral thesis

Student: RNDr. David Mareček

Thesis title: Unsupervised Dependency Parsing

Supervisor: doc. Ing. Zdeněk Žabokrtský, Ph.D.
Institute of Formal and Applied Linguistics, MFF UK
Malostranské náměstí 25
118 00 Praha 1

To rephrase the focus of David Mareček's thesis, it aims at developing a dependency syntactic analyzer of written natural language sentences, without using manually parsed data. Unsupervised dependency parsing is a relatively new, but growingly popular subfield of Natural Language Processing. The thesis gives overview of the current methods used for the task and presents the data sets containing thirty languages, which are used for experiments. Then, several probabilistic models are elaborated and used in the inference procedure based on Gibbs sampling. The parsing performance achieved by the newly implemented system is evaluated using several performance metrics for all languages under study.

The main technique used in the work – Gibbs sampler – has found its place in Bayesian branch of Natural Language Processing already several years ago. Some of the partial models used in the work have conventional shapes as well, even if there are certainly several novel ideas too (such as the reducibility model or some solutions in the sampling procedure) . However, what I consider the biggest value of David's work is the wide range of performed experiments, which allowed him to gradually combine the most promising shreds and to obtain excellent accuracy for many languages at the end. Needless to say there were countless unsuccessful experiments, many of which did not find their way into the final thesis text, but it never made him tired of designing new experiments. It paid off, as the performance which David's parser reaches for some of the languages constitutes the state of the art in unsupervised dependency parsing now. I can only hope that his experience with the modern unsupervised approaches will be conveyed to other students soon.

Besides working on his own research topic, David has been active also in several other research endeavors carried out in the institute. He has been one of the most fertile developers of our syntax-based machine translation system. He has contributed to building language data resources, especially the parallel treebank CzEng. Currently he supervises two younger colleagues working with him on the European machine translation project FAUST.

David is an author or co-author of 20 papers, most of them published in reviewed proceedings of international conferences. For those interested in scientometrics, I can add that the value of his h-index currently equals 4, which – given that he is only in the fourth form – indicates good publication quality. By the way, some of David's most recent results has been publicly cited on an international conference even before being actually published.

Last but not least, let me mention that his thesis belongs to the most quickly submitted PhD theses in our institute in the last years.

To conclude, in my opinion David's work presents a valuable scientific contribution and I do find it fully adequate for a PhD defense.

Jahodov, 20. 8. 2012

doc. Ing. Zdeněk Žabokrtský, Ph.D.