

Supervisor's evaluation of doctoral thesis

Student: Mgr. Rudolf Rosa

Thesis title: Discovering the structure of natural language sentences
by semi-supervised methods

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

Rudolf Rosa's thesis aims at finding new ways to develop dependency parsers for languages for which no manually annotated treebanks are available and thus standard supervised approaches are not applicable.

The strategy adopted in his thesis lies in borrowing the needed information from languages for which the needed data resources exist. Two particular families of approaches for such cross-lingual dependency parser transfer are explored in detail in the thesis, namely delexicalized and lexicalized approaches. In the former case, sequences of part-of-speech labels are processed instead of sequences of words. In the latter case, the resources are transferred via Machine Translation across the boundary between source and target languages.

In fact there are often no POS-tagged data too for languages that are under-resourced from the treebanking point of view. That is why the thesis pays attention also to cross-lingual tagging.

In general, I appreciate that Rudolf systematically explored multiple different directions whenever he faced a decision to make (although it often led to quite complex and computationally demanding experimental scenarios), he described all the experiments in a technically sound way, and he carefully compared his achievements with the latest state-of-the-art context. What I consider the main contribution of this thesis is the newly introduced KL-based language similarity measure that proved to be useful not only for choosing the nearest source language for a given target language, but it worked well also as a mixture coefficient in the multi-source transfer scenario.

Besides working on his own research topic Rudolf was active also in several other research endeavors carried out in the institute. Above all, he participated in several

Machine Translation projects, and he is a co-author of the widely acknowledged treebank collection called Universal Dependencies. In total, Rudolf is an author or co-author of more than 30 papers published in reviewed proceedings of international conferences, and his research has already received a remarkable number of citations.

Last but not least, Rudolf currently teaches three courses related to Natural Language Processing at the faculty, and I am happy he conveys his extensive practical experience on processing multi-lingual data to younger students.

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

Jahodov, 28th May 2018

doc. Ing. Zdeněk Žabokrtský, Ph.D.