Evaluation of doctoral thesis

Student: Loganathan Ramasamy, M.A., M.Sc.

Thesis title: Parsing under-resourced languages:
                Cross-lingual transfer strategies for Indian languages

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

Dependency parsing is a very popular discipline in the nowadays Natural Languages Processing. Typically, the best results are achieved by using supervised parsers that try to exploit linguistic expertise and man-years of manual annotation effort invested into building a treebank for a particular language. However, alternative and less expensive approaches are constantly being searched for too, because for vast majority of languages there are simply no treebanks.

Loganathan explored both ways in his experiments with parsing Indian languages. He went the hard way and built his own treebank for Tamil from scratch. But in the same time, he experimented with a number of alternative methods aimed at creating a parser without using a particular treebank, by distilling and synthetizing the needed information from other language data resources.

Even if the Tamil treebank might look as "just another treebank", I think it deserves respect not only because of the fact TamilTB is a "one-man" treebank, which is quite exceptional in the treebanking world. Besides all the standard problems that one expects when building a treebank for one of the languages of "the old continent", Loganathan had to resolve many other issues, both technical (related e.g. to the encoding of Tamil script) and linguistic (because of crossing the border between very different linguistic traditions). I am sure that the existence of the Tamil treebank will be appreciated by the steadily growing community of Indian NLP researchers.

Loganathan's alternative parsing methods are implicitly based on the assumption there is something language universal about dependency syntax, and thus it should be possible to transfer the parsing models (in the very wide sense) from one language to another. He designed several transfer strategies for various resource-poor situations, and carefully evaluated them. Loganathan's work shows that sharing syntactic structures among equivalent translations is much harder task than I would expect, but still he shows several positive experimental results (for instance the surprisingly competitive results reached by replacing a parallel corpus with a bitext resulting from machine translation).

Last but not least, I would like to mention that Loganathan is a co-author of the multilingual collection of treebanks called HamleDT, which is becoming a popular resources for multilingual parsing experiments.

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

Rychnov nad Kněžnou, 20. 8. 2014

                                        doc. Ing. Zdeněk Žabokrtský, Ph.D.