Diploma thesis review

Duong Thanh Long : Universal POS Tagger

In this thesis, Long Duong investigate the task of training a Part-of-Speech tagger by exploiting parallel data and projecting POS information from the tagged source language side into the (untagged) target language side. The work is motivated by the need of having POS taggers for languages which are widely used but lack annotated data needed for training supervised models.

Chapter 1 of the thesis presents introduction and motivation for the work. In Chapter 2, the author presents a comprehensive overview of relevant work – monolingual and multilingual tagging, covering both supervised and unsupervised approaches. Chapter 3 includes description of the main algorithm, experiments and results. Chapter 4 deals extends the work by a proposal of method for predicting the optimal selection of the source language. Chapter 5 concludes the work and presents ideas for future work. The thesis is written in a good English (with some grammatical/stylistic errors) on 56 pages plus comprehensive references on 5 pages and an appendix with the list of universal tags.

The thesis is well-structured. The algorithms and experiments are well-described and analyzed. The results of POS tagging are comparable to the state-of-the-art method by Das and Petrov (2011), but Long's algorithm is much more efficient and the code was made publicly available. This method was presented at the ACL 2013 conference and published in the proceedings. The method for optimal selection of the source language will be published at the IJCNLP conference later this year. These two facts are a good sign of the quality of the work.

Conclusion

Long Duong has submitted an impressive work and managed to publish its results in two top-notch international conferences within less than one year which is exceptional compared to other theses submitted to our department. I recommend this diploma thesis to be defended.

Pavel Pecina, ÚFAL MFF UK   Praha, 23.8. 2013