Diploma Thesis Review

Thesis title: Large-Scale Discriminative Training for Machine Translation into Morphologically-Rich Languages

Thesis author: Miloš Stanojević

Opponent: Zdeněk Žabokrtský

Thesis description

The aim of the thesis is to develop discriminative models for choosing best translations from candidate translations offered by state-of-the-art (generative) machine translation models.

The thesis is structured as follows. After Introduction, Chapter 1 gives a quick overview of the contemporary SMT approach. Chapter 2 brings motivation for using discriminative training in SMT and presents selected previously published techniques capable of such training. Chapter 3 summarizes metrics that are used for evaluating MT quality. Chapter 4 describes features which are hoped to discriminate between good and bad candidates. Chapter 5 presents performed experiments. The Conclusion chapter and Bibliography follows. The thesis consists of 93 pages.

Comments

The thesis is well structured. The background knowledge (especially the training algorithms and evaluation metrics) are described in a unified fashion and with a very good insight. The fifth chapter, which constitutes the center of gravity of the thesis, describes and evaluates a large number of experiments. I appreciate that the author often tries to compare several different alternative (e.g. different mutations of objective functions, different learning algorithms, different feature sets) and to find some interpretation of the results achieved. On the other hand, to me it seems that feature engineering sections might deserve more attention (but according the the conclusion, this was left for future work).

Although the experiments are relatively complex and well designed, the overall achieved improvement is very modest. However, this is often the case in the contemporary SMT.

The previous version of the thesis was rejected, with the biggest concerns being the insufficient extent of experimental work and abundance of formal flaws. I am very happy to see that the current version of the thesis has been radically improved in both these aspects. The experimental part is several times bigger now, and the number of formal flaws is much smaller than before (actually I have found only a few typos such as “breath-first search”, “prevent form...”; a few pages with references are missing in my hard copy, however, one can find them in the electronic version; the table of contents gives a wrong page number for Bibliography).

I would like to ask a question. The author argues (p.45) that target-language dependency trees could be built much faster by projecting dependency trees from the source side through alignment, than by proper parsing. He seems to be ready to accept quite big amount of errors introduced by such a projection. However, wouldn't it be better to introduce a highly simplified target-side parser (in analogy to the author's “quick and dirty”, but sufficient unigram tagger), instead of the tree projection? Perhaps even a drastically down-scaled supervised parser model could lead to better accuracy than tree projection, while retaining comparable speed. Or does the tree projection have some other advantage?

Conclusion

Miloš Stanojević has shown that he is able to analyze and implement a complex research task. I recommend to accept the thesis for the defense.

In Rychnov nad Kněžnou, 14th January 2013

doc. Ing. Zdeněk Žabokrtský, Ph.D.
Institute of Formal and Applied Linguistics
Charles University in Prague