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 a discriminative model for word form choice (within an existing Statistical Machine Translation system) whose training can be driven by a sentence-level MT metric and which is capable of using a large number of features.

The thesis is structured as follows. After Introduction, Chapter 1 gives a very 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 evaluation of MT quality. Chapter 4 describes features which are hoped to help the model, and Chapter 5 presents evaluation of experiments. Conclusion chapter follows and Bibliography follows. The thesis consists of 45 pages.

Comments

As it is clear from the introductory chapters, the author understands the principles of the contemporary SMT. However, in my opinion, the overall thesis quality does not comply with the standards of the faculty and of the institute. While the parts which are based on literature research are written relatively carefully (Chapters 2 and 3 and majority of Chapter 4), the original contribution of the author himself is very modest. In fact, there are only a few relatively straightforward ideas mentioned in Chapter 4 and there are only 3.5 pages of text describing the new experiments in Chapter 5. I do not criticize the fact that no substantial improvement was achieved, but I believe that more energy should have been invested into trying it.

There are also some issues that remained unclear to me. Is the search space really infinite, even if the number of translation pairs (at any level) observed in the training data is finite? Treating morphological richness is one of the proclaimed features of the work, but did it all really shrink just to several-letter prefixes at the end, as suggested in Chapter 4? Is the impact of latent variables on the loss function convexity really specific (in any sense) for machine translation, if the existence of local extremes is a known risk for almost any optimization task?

Not only the passages that should have constituted the core of the work are short, but the thesis as a whole is also not flawless from the formal point of view. Several letters of the Czech alphabet are systematically wrongly encoded in the Czech version of the abstract. There are many mistakes in the mathematical typesetting and several in-text formulas are even not “latexed” at all (e.g. “phrase fi bi-1” on page 3, as well as itemization on page 8). There are wrongly used quotes, many missing spaces in front of references to literature, mixed format of references to literature, function τ without explanation on page 9. One can find several language errors ("We first a translate", "unprobable", "The derivation … represesent", "splitting tuning data in to shards", "is is", "Working … allow us", "acheaved", "opetimizing", several missing determiners). The maximizing variable y' on page 10 does not appear in the maximized formula. There are missing double braces in the bibtex file (which leads to undesired lowercasing of titles several times). There are missing first names in some bibtex entries.

Conclusion

I do not recommend to accept the current version of the thesis for the defense, but I believe that Miloš Stanojević will be able to prepare a more elaborate version if he is given more time.

In Rychnov nad Kněžnou, August 20, 2012

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