Master's Thesis Review

Author: Miloš Stanojević
Title: Large-Scale Discriminative Training for Machine Translation into
              Morphologically Rich Languages
Supervisor: RNDr. Ondřej Bojar, Ph.D.

The thesis submitted by Miloš Stanojević examines machine learning techniques applied in machine translation (MT) that aim to improve translation quality by considering large sets of features. These features are then supposed help to discriminate better and worse candidate translations.

The thesis is well structured and contains an acceptable number of grammatical or typesetting errors. All the experiments are clearly described in appropriate detail and the text reads well.

The main contributions of the thesis are:

- a solid introduction to generative vs. discriminative models as applied in MT,
- a detailed description of state-of-the-art optimization algorithms,
- a thorough discussion of objective functions used,
- Miloš's own variant of one of the objective functions, ROUGE-SX, that corrects some unfortunate properties of the original ROUGE-S,
- identification of technical problems with using rich features in phrase-based MT and proposed (but not tested) solutions for them,
- a summary of possible feature functions applicable in discriminative models,
- experiments with English-to-Czech MT comparing various optimization algorithms, scoring functions, and utilizing some of outlined features; the most complex model uses 42 thousand features.

Sadly, none of the tested setups performed significantly better than the baseline, but this can still be considered a good result because e.g. the largest model includes 42 thousand features and despite of that, the learning algorithms are able to reach the state-of-the-art score on a rather small tuning set.

To conclude, the thesis documents that Miloš Stanojević was able to explore discriminative training methods, bring in his own suggestions, empirically test them and clearly describe the results. Despite the relative short text overall and mainly the experimental section, I suggest the thesis to be accepted as a M.Sc thesis at Charles University in Prague.

Prague, August 29, 2012.

                              RNDr. Ondřej Bojar, Ph.D.
                              Charles University in Prague, ÚFAL