Review of the Master Thesis

Reviewer: Mgr. Pavel Straňák, Ph.D.

The author sums up the approaches to acquiring lexical information from Wikipedia, experiments with them, analyses the extracted dictionaries, and evaluates their usefulness in a machine translation setup.

Thesis content

The main content of the thesis is divided in two chapters: Dictionary, and Evaluation on Machine Translation System.

The first of these chapters starts with a thorough introduction to Wikipedia, included the technical issues relevant for author's work. It continues with methods of extracting bilingual dictionaries from Wikipedia and analysing (identifying and categorising) named entities in the extracted dictionaries.

The second chapter is devoted to a task-based evaluation of the extracted dictionaries: trying to employ them in the statistical machine translation system to improve the translation quality (measured by the BLEU score).

Evaluation

I find the author's work very thorough. Clearly, she not only read the relevant papers and summarised them, but also found some deficiencies of the presented approaches and tried to improve on them, as we can see for example on pp. 17–18, where she provides additional rules to a presented heuristic for identification of named entities.

The choices of methods she decided to implement in her experiments as well as the evaluation of each partial result seem appropriate. However in some places I find explanation of the methods hard to follow, e.g. a start of the bootstrapping algorithm for classification of named entities on p. 21. I would prefer either a fuller explanation, or just a reference.

I appreciate the effort put into the task-based evaluation of the extracted dictionary, in addition to the intrinsic evaluation as measured by accuracy, precision, recall and the F1 score of each experiment.

The author's English is clear and generally very good. Grammatical errors are few and far between and do not impede understanding of the text. The most common errors are probably related to articles, or the lack of them, e.g. the title of the third chapter “Evaluation on Machine Translation System”. I was also surprised by the usage of the term "normalisation" meaning "lemmatisation" of word forms, even though lemmatisation certainly is a kind of normalisation.

The references are very thorough and the bibliography is both comprehensive and
up-to-date. However, using numbered references and at the same time putting the bibliography in the middle of the volume is truly a test of a reviewer's patience.

Conclusion
In my opinion the thesis fully complies with the requirements of a Master Thesis at the Faculty of Mathematics and Physics at the Charles University in Prague. I recommend the thesis to be accepted.

Prague, 1. September 2011 Pavel Straňák