Review of the Master Thesis

Author: Angelina Ivanova  
Title: Creating a Bilingual Dictionary using Wikipedia  
Supervisor: RNDr. Daniel Zeman, Ph.D.

The goal of the thesis is exploring the possibilities of using Wikipedia as a bilingual lexical resource, creating a bilingual dictionary and evaluating the dictionary. The thesis consists of the Introduction, two main sections, the Conclusion, the lists of references, tables, figures and abbreviations, and an appendix with samples from the acquired dictionaries. The two main sections span almost 50 pages and cover the two main experimentation areas: dictionary acquisition (including experiments with named entities) and extrinsic evaluation in machine translation environment. The attached CD contains, besides the text of the thesis, a collection of software and data files (including the acquired dictionaries) that will simplify building upon the author’s work.

The author’s contribution can be divided into the following areas: 1) implementing several algorithms of dictionary acquisition from Wikipedia (including necessary engineering tasks to assess efficiency of various existing software packages designed to help accessing Wikipedia); 2) implementing and evaluating a heuristic-based algorithm to recognize and classify named entities among the acquired translation pairs; 3) semi-automatic evaluation of the acquired dictionary against an existing English-Russian dictionary; 4) setting up and analyzing several statistical machine translation experiments as an extrinsic evaluation of the usability of the acquired dictionary.

The text of the thesis is well structured, written in good English (although a few typos inevitably made it to the final version). The experiments are well documented and, even more importantly, analyzed. The work shows that the author has gained a good insight into several areas of natural language processing, such as lexical acquisition, named entity classification and statistical machine translation. She provides the reader with an ample theoretical background and with references to relevant literature. She also extensively works with existing tools and proved the ability to install and quickly gain confidence with new programming languages, software packages and environments.

Conclusion
The thesis proves the author’s ability to independently and creatively solve problems and conduct research in the area of NLP. On my opinion, the thesis complies with the requirements for Master Thesis at MFF UK. I recommend the thesis to be accepted.

Praha, August 30, 2011

RNDr. Daniel Zeman, Ph.D.  
Institute of Formal and Applied Linguistics  
Charles University in Prague