Title: Creating a Bilingual Dictionary using Wikipedia

Author: Angelina Ivanova

Department/Institute: Institute of Formal and Applied Linguistics (32-ÚFAL)

Supervisor of the master thesis: RNDr. Daniel Zeman Ph.D.

Abstract:

Machine-readable dictionaries play important role in the research area of computational linguistics. They gained popularity in such fields as machine translation and cross-language information extraction. In this thesis we investigate the quality and content of bilingual English-Russian dictionaries generated from Wikipedia link structure.

Wiki-dictionaries differ dramatically from the traditional dictionaries: the recall of the basic terminology on Mueller’s dictionary was 7.42%.

Machine translation experiments with Wiki-dictionary incorporated into the training set resulted in the rather small, but statistically significant drop of the the quality of the translation compared to the experiment without Wiki-dictionary. We supposed that the main reason was domain difference between the dictionary and the corpus and got some evidence that on the test set collected from Wikipedia articles the model with incorporated dictionary performed better.

In this work we show how big the difference between the dictionaries developed from the Wikipedia link structure and the traditional dictionaries is. Wiki-dictionaries may reduce the quality of the machine translation if used on the improper domain. The advantage of such dictionaries is that that they provide a great number of named entities and that they can be easily developed automatically for any language pair which is present in Wikipedia.

Keywords: machine-readable dictionary, Wikipedia, named entity recognition and classification, parallel corpus, machine translation