Master's thesis of Radoslav Klíč – the advisor's review

Title: Automatické osvojení vzorů s minimální supervizí (Acquisition of inflectional paradigms with minimal supervision)
Submission year: 2012
Name of the advisor: Jiří Hana (hana@ufal.mff.cuni.cz)
Affiliation of the advisor: Charles University in Prague, MFF, ÚFAL

The thesis describes a lightly-supervised system for the acquisition of morphological paradigms and a lexicon. It combines two components:

(1) an extensions of Paramor (Monson, 2009), an unsupervised system. Paramor is modified: (i) to accept a small seed of manually provided word inflections with marked morpheme boundaries; (ii) to handle basic allomorphic changes acquiring the rules from the seed and/or from previously acquired paradigms.

(2) a module for bottom up clustering of words with configurable metrics. Three metrics are provided: one using parametrized edit distance, one measuring similarity of morpheme segmentation and one comparing similarity of paradigms.

The algorithm has been tested on Czech, Slovene, Catalan and German tagged corpora. In comparison with the baseline Paramor results, F-measure improved for all languages, although for German only negligibly.

The thesis is clearly written and well structured. The first chapter provides motivation and an overview of linguistic terminology. The next chapter discusses all the major relevant work by other researchers, with appropriate references. A chapter briefly introducing the Paramor system follows. The last three chapters describe the original work of the student – his modification of the Paramor system, the clustering framework and the experiments and their results.

The text is very concise, but clear and generally easy to follow. However, some passages would benefit from more examples (e.g., for the clustering metrics in Chapter 5.1). Also some of the evaluation results could be discussed in more detail.

A supplemental file contains the code of the system, the seeds and the configuration files for all experiments (only the Slovenian corpus is included due to licensing restrictions). The actual code of both Paramor's extension and the clustering framework is very clean and easy to follow. The design is elegant and allows for easy extension and modification. More comments and a general readme file should have been added, although the structure and transparent naming conventions still make it easy to follow, definitely easier than Paramor's code.

In sum, the thesis presents a sound work with only minor problems. I recommend accepting it as a partial fulfillment of the requirements for the Master of Science degree.

Prague, May 3, 2012

Jiří Hana