Supervisor's review of master thesis

Author of the review:	RNDr. Pavel Pecina Ph.D.
Author of the thesis:	Ilana Rampula
Title of the thesis:	Semantic Relation Extraction from Unstructured Data in the Business Domain

The thesis presented by Ilana Rampula deals with extraction of structured information from unstructured data. The unstructured data is text. The structured information is in a form of semantic relations of named entities appearing in the text. The input data is automatically processed, the relations extracted and inserted into a database.

Comparing to other diploma theses defended at the Institute of Formal and Applied Linguistics, this work is quite unique. This is mainly due to the data the experiments have been performed on. The dataset was provided by an industrial partner with the goal to enrich (or check) an existing client database by mining short textual notes (complaints) associated with the clients.

The thesis is structured into six chapters including introduction and conclusion. After the introduction, the author presents theoretical background of the work. Then she specifies the data set used in the experiments and continues with description of the methods employed in the work. This is followed by a chapter presenting and discussing the experimental results. The text spans a total of 68 pages.

The author focused on two diverse methods for relation extraction from text. Both rely on contextual information but one is based on statistical classification, the other one is a heuristic bootstrapping method iteratively extracting contexts which are likely to represent the relation. The two methods were originally proposed for binary relations and modified by the author to be applicable to unary relations. Both methods were tuned, evaluated and their results analysed.

The methods were applied to the unary relation *country of origin*. This was motivated by the practical need of test data. The author focused on relations already existing in the database with the goal to test the results against some existing data. Unfortunately, it turned out that *country of origin* was basically the only possible relation which allowed that. Extraction of other relations would be impossible to evaluate in such a way or they were not present in the data frequently enough.

The thesis is very well written and readable, the methods were applied correctly, are well described and illustrated using appropriate examples, the results are clearly presented and analysed. The author showed her ability to find a solution for the given task and delivered a good piece of work (very precise and thorough in same aspects – e.g. the evaluation of POS tagging and NE extraction – which is usually neglected). I recommend the thesis to be defended.

Prague 1. 6. 2016