Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Jonathan Oberländer		
Název práce	Splitting word compounds		
Rok odevzdání	2017		
Studijní program	Informatika	**Studijní obor**	Matematická lingvistika

Autor posudku	RNDr. Pavel Pecina Ph.D.	**Role**	vedoucí
Pracoviště	ÚFAL MFF UK		

Text posudku:

The thesis by Jonathan Oberländer deals with the problem of word compounds. Some languages (such as German, Dutch, Swedish, etc.) tend to form compounds which are combinations of several single words without orthographical separation. Such words are typically not lexicalized, they are created ad-hoc and therefore pose a problem for many tasks of natural language processing. This problem is often approached by automatic splitting of the compounds into their components (single words). The goal of the presented thesis is to explore existing tools and algorithms for this task and implement a new one which can be easy adaptable to new languages.

The thesis is written on the total of 28 pages, structured into 7 chapters plus a list of references and two attachments in a form of software and data packages, which have been uploaded to the Student Information System. The thesis is experimental and contains all the required parts: an introduction to the research problem, motivation for the work, and description of the thesis goals (Chapter Introduction), theoretical definition and analysis of the problem (Chapter Compounds), overview of related work (Chapter Related Work), description of the data and methods used in the experiments (Chapters Corpora and Methods), evaluation of the conducted experiments and conclusions (Chapters Evaluation and Conclusions).

The text of the thesis is in English, well written, dense and concise, and mostly easy to read. The author proposed a method for splitting word compound, implemented it for three languaes (German, Swedish, Hungarian), evaluated its performance on own test set, and compared the results with other (state-of-the-art) methods. From the research/experimental point of view, the work that has been done is very good and above average (implementation of the method, design of the experiments, evaluation, result analysis, etc.).

The thesis has two drawbacks: level of detail and extent of the work. The dense and concise text is sometimes too brief and missing details which makes certain parts of the text harder to understand, especially for a reader no familiar with the research methods and terminology used in the work (e.g., page 13-14: what is “a binary split”, “the smallest possible split”?). In some cases, the terminology is not properly defined and the author relies on the intuition of the reader (e.g., “split” – does it include the linking morphemes or not?) Some details important to replicate the experiments are not mentioned (e.g., in the construction of the evaluation corpus, page 10-11: how exactly was the source data shuffled – document/sentence/word level? How did the author make sure that the shuffling did not “introduce difference between languages”? How was the minimum word length set? How the list of words was created? How long was it? How was the Kappa value on the Hungarian set calculated?

Regarding the extent of the work, both in terms of programming and experimentation, the amount of the of the work that has been done is below average. The author identifies several options for future work and ways how to improve effectiveness of the method but none of them has not been explored (e.g., the effect of the size of training data/list of single words or learning of easy compounds first). Exploring some of these areas and conduction additional experiment would have made the work richer and stronger.

Still, the author demonstrated that he well understood the problem, studied most of the related work, presented his own solution for splitting word compounds and compared its performance with existing state-of-the-art tools. The goals of the theses were therefore fulfilled and I recommend the thesis to be defended.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

In Prague, 24. 1. 2017

Podpis: