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 previous version of the thesis was not succesfully defended. The current version is a significant update of the previous one. Tt is written on the total of 46 pages, structured into 8 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), description of the proposed algorithm (Chapter Alhorithm) with details of its implementation added as a new chapter (Chapter Implementation), evaluation of the conducted experiments and conclusions (Chapters Evaluation and Conclusions, respectivelly).

Similarly to the previous version, 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 two main drawbacks of the previous version of the thesis (level of detail and extent of the work) have been resolved. The text contains sufficient level of details and also the experimental and evaluation parts have been impoved (e.g. in terms of analysis of the effect of lexicon size, the effect of cleaning methods employed and other parameters). The missing details regarding implementation have been added too.

The author submitted a significantly updated version of his work. The thesis demonstrates that he well understood the problem, studied 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 fulfilled and I recommend the thesis to be defended.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

In Prague, 31. 5. 2017

Podpis: