Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

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

Autor posudku	Hlaváčová Jaroslava	**Role**	Oponent
Pracoviště	ÚFAL		

Text posudku:

The thesis solves a problem of splitting compound words in languages where the compounding forms an important part of word formation process. The author works with three “compounding languages” - German, Swedish and Hungarian.
The thesis has 7 chapters plus Bibliography.
In the first chapter, Introduction, he gives a few examples of compound words in German, including main problems of their splitting. He also mentions the reason why the decompounding is important.
The second chapter, Compounds, brings a slightly more detailed insight into the problem of compounds.
The chapter Related Work gives a brief survey of recent approaches to the problem.
Corpora is the name of the chapter that lists data sources. Here we learn that the work deals with a limited - medical - domain. The important data source is the list of linking morphemes, which are presented in a special table on p.12. Apart from corpora used for making a lexicon (Wikipedia medical articles) and as evaluation texts (EMEA corpus), there are also files with stopwords and affixes for all three languages.
The core of the thesis is the chapter Methods presenting the procedures used to achieve a correct decompounding of a given word, together with the chapter Evaluation comparing results of several system settings and all three languages.
In the Conclusion, the author gives several hints for more experiments that could provide better results.

I appreciate that the author included the recent method of word embedding in the Methods chapter, though a more detailed explanation would be appropriate.
The results presented in the Evaluation suggest, that the decompounding tool really works and gives satisfactory results. It would probably work for other languages with a similarly minimal data resources, which was one of the main requirements.

However, the thesis has severe weaknesses:
The descriptions are too brief and not precise enough. (The work has only 24 pages of text.) The main terms are not explained properly and the reader is sometimes confused what they really mean. The first problem is the meaning of the basic term “split” itself. Starting with the chapter Methods, it should be used in a strict sense, but the reader hesitates, if it concerns the procedure of splitting a word, or the set of split parts of the word, or the place where the splitting takes place. This makes the understanding difficult.
Another inaccuracy concerns the terms “dictionary” and “lexicon”. Is it the same?
What is a “tie” (first occurring at p.19)?

Presenting examples would help a lot, but there are only few of them in the whole thesis, all for German only (except for the table of the linking morphemes).
The description of the algorithm used (chapter Methods) should be more precise, and more formal. For instance, what are “possible” binary splits in the 2nd sentence of the section Generation of candidate splits (p.13)?
The main weakness of the work is the absence of a detailed user as well as technical documentation. The only user documentation is a section Interface at p.17, but there is even not given the name of the script to run.
The text presented in the chapter Methods cannot substitute for a proper technical documentation.
There is no exact description of the Attachments uploaded to the SIS. The arrangement of the files is really strange - there are paths of several empty directories, the last of which contains a gzipped file with the data itself. There is no “read.me” file, no description of the individual files, their formats.
The insufficient documentation is the main reason why I do not recommend the thesis for the defense.

Minor questions:
A hypothesis is presented, that the system could probably work better on general, not expert medical data. According to my experience, a limited domain gives usually better results. Why does the author think that it woud be vice versa in the case of word decompounding?

The lists of linking morphemes are surprisingly short - only 1 linking morpheme for Swedish, 8 morphemes each for German and Hungarian. Is it a general feature of those languages, or is it caused by the limited medical domain? What was the source of the lists?

Práci nedoporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 23. ledna 2017 **Podpis**