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 8 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 4, that lists data sources. 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 important data source are also the lists of linking morphemes, presented in a special table on p.16.
The core of the thesis are the chapters 5-7:
The chapter Algorithm presents the rules and other procedures used to achieve a correct decompounding of a given word. Several ways of cleaning the results of the algorithm are explained. Final result - the correct split - is necessary to select from potentionaly more results. Thus, the procedure of ranking the splits is also important.
The chapter 6 - Implementation - concerns the documentation. It contains also brief instructions for those who would need to add a language or a method to the present system.
The chapter Evaluation compares and discusses results of several system settings and all three languages.
In the Conclusion, the author summarises the presented work and gives several hints for more experiments that could provide better results.

The results presented in the Evaluation suggest, that the decompounding tool really works and gives satisfactory results. There is a wide range of different settings compared, together with a solid discussion.
I would suggest one more experiment with another lexicon - not only a "cheap" one from the Wikipedia, but with a real big morphological lexicon, at least for German. It would be interesting to uncover, how much the lexicon influences the results.

I have several small comments:
In chapter 5, the rules of cleaning and ranking methods are sometimes complicated - more

examples would help a lot.
In the method 5.4.2. last_parts, why do you think that "base forms of words are more likely to be known than inflected ones"? Known from where? From Wikipedia texts?
Another not very precise statement is in the following method 5.4.3. "if ... and is not much larger than it ...", especially with examples of -iv and -iver, being twice larger (longer) - what is "much larger"?
The method 5.4.4.: "If any part of a proposed split is a prefix, we disregard the split." How can such a situation occur?
There are two conditions for the method 5.4.5, but it is not clear if they should be connected with "and" or "or".
The name of the section 5.6 should be rather "Modifications" than "Improvements" - their evaluations is yet to come.
In the evaluation of Swedish results for the V+S setting (p.33), there is a hypothesis that the low Recall may be caused by the fact that Swedish words are generally shorter than German ones. It would be interesting to check it by setting a different limit for Swedish.
A caption of Table 7.3 is not appropriate - it was probably cut and pasted from the Table 7.1.
The Table 7.5 should include a column "total", for clarity.

The goal of the work was fulfilled, I recommend the thesis for the defense.

Práci doporučuji 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.května 2017 **Podpis**