Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Eric A. Lief		
Název práce	Deep contextualized word embeddings from character language models for neural sequence labeling		
Rok odevzdání	2019		
Studijní program	Informatika	**Studijní obor**	Matematická lingvistika

Autor posudku	Ing. Tom Kocmi	**Role**	Oponent
Pracoviště	UFAL, MFF UK		

Text posudku:

The thesis presents solid study of using different embeddings setups in three sequence labeling tasks: part-of-speech tagging, named entity recognition, and verbal multiword expression. This work is primarily concerned with a neural sequence tagger combining various word and character embeddings for the Portuguese language.

The thesis clearly shows that the author is familiar with the newest development in the deep learning. He implemented the models by himself and conducted all the experiments as well as manual analysis of the errors.

The text of the thesis is divided into 5 chapters plus the introduction and the conclusion. The division is clear and makes the text easy to follow. The work is written in very good English, with some typos. However, the structure of individual sections is rather poor. For instance, the indexing of figures is shifted (probably one figure was removed, but the references in the text were not updated). The author usually does not refer to the tables and figures by their number but rather by words “above” or “below”, which can be confusing especially in cases when the tables are on different pages or several tables are on one page. There are many unnecessary duplicates of equations, figures and tables, for example equations 15 and 16, table 6 and 9 or the description of experiment architectures on pages 31-33, 41-42, 58-59 and 76-77. The tables and figures are often placed without explicit order and it is not easy to notice differences in hyperparametrs between experiments. I would suggest to put all hyperparameters and experiment description into the appendix. Also, I miss a thorough analysis of hyperparameters and their effect on training (for example total number of parameters in each setup, total time of training, etc).

The author uses many illustrations, which I appreciate, to explain the various architectures, but he uses original pictures from published papers which often use a different notation. For example, figure 4 and 5 use the same graphical representation for different concepts: for instance in the first figure the arrow represents a simple connection of components and in the second figure it represents an operation like matrix multiplication; square symbol represents word embedding in the former figure and one-hot representation in latter case, etc. I would suggest more thorough explanation of the architectures or re-drawing of the illustrations into one style and notation. The figure 6 is not from a work of Bengio at al. 2003.

There are other minor issues like missing abbreviations in the list (VMWE, PWE, WE, LVC, VPC, …), some terms are used without any definition (word2vec), there are many inconsistencies (equations on the page 19), etc. There is a duplicate in reference (Józefowicz 2016) and several preprint references instead of original publication references. All could be solved by proof-reading.

The evaluation of the approach is reasonable thorough, including both automatically computed scores as well as manual analysis of errors. The author showed that the architecture improves the performance over the baseline and gets close to the state of the art.

Despite the mentioned problems, the author implemented the architectures by himself, he showed understanding of current approaches in the field, wrote a thorough section on the related work and did a proper evaluation. This is what matters the most and thus I recommend the thesis for defense.

Questions for the defense:

1) In Chapter 4, the task of Part of Speech Tagging is examined. There are missing any results to support the claim that the first configuration (in Formula (30) on page 43) outperformed the others. After the selection of BiLSTM character embeddings on the same page there is a claim that a number of experiments were run in a hyperparameter search, but there are no results to support the claim. What type of hypersearch have you run, do you have results supporting your claims of which run was the best? Also there are missing results for RNN-CLE in Table 33, can you provide them?

2) How did you define the maximal number of epochs per training? Can you explain why VMWE was trained for less epochs than NER despite having five times more training data?

3) There is a claim that “state of the art results were achieved for PoS tagging (F-score of 97.49)” [sic], however the improvement is only by 0.02, have you run any significance tests to support your claim? If not, explain why do you claim state-of-the-art result and how do you prove it?

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 21. ledna 2019 **Podpis**