Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Daniel Kondratyuk
Název práce	Multilingual Learning using Syntactic Multi-Task Training
Rok odevzdání	2019
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	David Mareček **Role** oponent
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

The goal of Daniel's master thesis is to use pretrained multilingual BERT model for developing a universal multi-task learning of taggers and parser trained on all treebanks included in Universal-Dependencies that would be able to analyse any language with only one single model.

The thesis is divided into 6 chapters. The introduction comprises the main contributions and research questions.

The first chapter is devoted to the work background, it describes the Universal-Dependencies annotation scheme and then goes through the evolution of deep neural network architectures, from feedforward networks, recurrent networks, to transformer self-attention networks. Further, the author explains pretrained contextualized word representations such as ELMo or BERT, and some related works concerning the fine-tuning of such representations and training highly-multilingual models.

The second chapter describes the task itself, which is predicting Universal Dependencies form raw texts, according to the CoNLL 2018 shared task. The task is to predict part-of-speech tags, lemmas, morphological features, and dependency trees with dependency labels. Daniel describes some approaches solving these tasks, such as RNN-based tagging or graph-based biaffine attention parser on bidirectional LSTM (by Dozat and Manning). Daniel also introduces his own model LemmaTag, learning jointly to tag and lemmatize, and UDPipe Future by Milan Straka, a multi-task learner of all UD annotations.

In the third chapter the model itself is described. The UDify model is based on the UDPipe Future model, but using the BERT model as the base instead of recurrent neural networks. Daniel also proposes layer attention method which enables each learned task (POS tagging, lemmatization, parsing) to take individual weighted average of BERT layers that is most suitable for the task. Because the BERT architecture is a very strong network inclining to be easily overtrained, Daniel proposes amounts of regularization methods like dropout or input words masking.

Experiments and results are shown in Chapter 4. Daniel chooses six high-resourced and six low-resourced treebanks and compare the results of his UDify model trained on only the one particular language with the model trained on all UD languages and with the UDPipe Future model. The results show that the multilingual UDify model outperforms UDPipe in terms of tree attachment score and in many cases also with respect to the tagging accuracy. Only the task of lemmatization is consistently a bit worse than for UDPipe Future. UDify model also shows good performance in zero-shot learning, e.i. when analysing languages that were not in the training data at all. Daniel also provides self-attention visualisations of BERT for pretrained model and after training on Universal Dependencies tasks showing that attention distributions of many self-attention heads become more sparse and more resembles dependency relations. In conclusions the results are summarized and some future work is proposed.

The thesis is 63 pages long, it is written in very good English, and it is clear and well organized into chapters. I found only a very low number of typos and other mistakes. The results of this thesis are very significant and important. The single multilingual model trained by Daniel seems to outperform majority of the previous parsers, which were trained or tuned specifically for individual languages. This work also seem to beat all recent approaches of parsing low-resourced and no-resourced languages. Even though the work done may look simple (in fact, Daniel only replaced the RNN base in UDPipe Future by the BERT self-attention network), there was a lot of work done on the regularization strategies that prevent the network from overtraining on a relatively small treebank data. I also appreciate the analysis of how the fine-tuning changed the BERT hidden representations. Daniel used probing syntax approach (Hewitt and Manning, 2019) and self-attention visualisation to show that the BERT representations resemble dependency relations after the fine-tuning.

In Chapter 3, I was expecting a detailed description of the whole network architecture that produced the final model. Daniel thoroughly described the layer attention mechanism and the regularization strategies used, which is all right, because these are the key and novel things that made possible to train such a strong model. However, I am missing detailed description of the final parts of the network, e. g. the taggers and parsers theirself. I suppose that the networks are equal to those used in the LemmaTag or UDPipe Future tools, however, the thesis should be self-contained and it should not be necessary to read e.g. the Dozat's and Manning's paper for finding the parser's details. The short description in Section 2.5 is insufficient and there seem to be mistakes, for example the variable r_i referred at the beginning of page 31 is missing in the formula (2.1).

I would also like to see more comparison with the CoNLL shared task systems. The author claims that such a comparison would not be fair since he uses gold segmentation and tokenization, but it would be easy to use e.g. the segmentation produced by UDPipe. I would be curious for

what languages the shared UDify model is able to beat all other systems. I would expect it could be at least for the low-resourced languages. I would also like to know, what improvement would bring a BERT model pre-trained only on single language. It might be compared at least for English and Chinese, for which the single-language models exist.

To conclude, I find the output of this thesis very exceptional with a very strong contribution for the NLP community. What seemed almost impossible 5 years ago is now very easy. One model that is able to analyse any language and majority of them with superior quality. The thesis itself could be more detailed in some ways and little more evaluation could be provided, but the overall quality is very high. I definitely recommend this work to be defended.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 4. 6. 2019

Podpis: