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	Milan Straka **Role** Vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

The goal of the Master thesis was to explore the possibility of improving morphological and syntactic analysis using multilinguality, i.e., benefiting from morphological and syntactic annotations available in multiple languages. While appealing, designing massively multilingual models has been a challenging task so far.

The thesis starts with describing Universal Dependencies, which is a framework for cross-linguistically consistent grammatical annotations, providing lemmas, POS tags, morphological features and labeled dependency trees. The recent version UD 2.3 used by the author provides 124 freely available treebanks in 75 languages. The deep learning prerequisites are then described in the rest of Chapter 1, notably word embeddings, contextualized embeddings and feedforward, recurrent and Transformer neural networks. The Chapter 2 reviews previous work on Universal Dependencies in lemmatization, POS tagging, morphological feature generation and dependency parsing.

The main contribution of the thesis, the UDify model, is described in Chapter 3. Building upon BERT contextualized embeddings, the UDify model is simultaneously trained on all 124 treebanks of UD 2.3, obtaining astonishing results. The model is purely multilingual, obtaining no language identification during training nor inference. Just as a matter of interest, the model was trained for 25 days on a single GPU.

Several model evaluations are presented in Chapter 4. When comparing to one of the winning system of CoNLL 2018 shared task in UD parsing, the UDify achieves comparable UPOS tagging accuracy and better unlabeled and labeled parsing accuracy (by +1.32 and +0.67 pp). The multilingual training is experimentally shown to really improve performance, especially for low-resource languages.

I consider the thesis to be of very high quality. The author was to my best knowledge the first to design and successfully train a massively multilingual model – a single model achieving results comparable to state-of-the-art performance on UD, and performs several experiments showing the influence of multilinguality, language-specific finetuning, and zero-shot evaluation. He has demonstrated independent, high-quality scientific work in an unexplored area, on an international level.

I recommend the thesis to be defended.

Práci doporučuji **k obhajobě.**

Práci nenavrhuji **na zvláštní ocenění.**

Datum 3. června 2019 **Podpis**