Title: Multilingual Learning using Syntactic Multi-Task Training

Author: Daniel Kondratyuk

Institute: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Milan Straka, Ph.D., Institute of Formal and Applied Linguistics

Abstract: Recent research has shown promise in multilingual modeling, demonstrating how a single model is capable of learning tasks across several languages. However, typical recurrent neural models fail to scale beyond a small number of related languages and can be quite detrimental if multiple distant languages are grouped together for training. This thesis introduces a simple method that does not have this scaling problem, producing a single multi-task model that predicts universal part-of-speech, morphological features, lemmas, and dependency trees simultaneously for 124 Universal Dependencies treebanks across 75 languages. By leveraging the multilingual BERT model pretrained on 104 languages, we apply several modifications and fine-tune it on all available Universal Dependencies training data. The resulting model, we call UDify, can closely match or exceed state-of-the-art UPOS, UFeats, Lemmas, (and especially) UAS, and LAS scores, without requiring any recurrent or language-specific components. We evaluate UDify for multilingual learning, showing that low-resource languages benefit the most from cross-linguistic annotations. We also evaluate UDify for zero-shot learning, with results suggesting that multilingual training provides strong UD predictions even for languages that neither UDify nor BERT have ever been trained on. Finally, we provide evidence to explain why pretrained self-attention networks like BERT may excel in multilingual dependency parsing.

Keywords: syntax multilingual universal dependencies BERT