Title: Neural Network Based Named Entity Recognition

Author: Jana Straková

Institute: Institute of Formal and Applied Linguistics

Supervisor of the doctoral thesis: prof. RNDr. Jan Hajič, Dr., Institute of Formal and Applied Linguistics

Abstract: Czech named entity recognition (the task of automatic identification and classification of proper names in text, such as names of people, locations and organizations) has become a well-established field since the publication of the Czech Named Entity Corpus (CNEC). This doctoral thesis presents the author's research of named entity recognition, mainly in the Czech language. It presents work and research carried out during CNEC publication and its evaluation. It further envelops the author's research results, which improved Czech state-of-the-art results in named entity recognition in recent years, with special focus on artificial neural network based solutions. Starting with a simple feed-forward neural network with softmax output layer, with a standard set of classification features for the task, the thesis presents methodology and results, which were later used in open-source software solution for named entity recognition, NameTag. The thesis finalizes with a recurrent neural network based recognizer with word embeddings and character-level word embeddings, based on recent advances in neural network research, which requires no classification features engineering and achieves excellent state-of-the-art results in Czech named entity recognition.

Keywords: named entity recognition, Czech Named Entity Corpus, artificial neural networks, recurrent neural networks, softmax, word embeddings, character-level word embeddings