Kolloquium, 22. Mai 2017

Dayyan Smith

Comparing representation learning over character-level, word-level or a combination of both in NLP tasks

Speaker: Iuliia Khobotova

Advisor: Wenpeng Yin

Iuliia is exploring the effect of different representations on the two tasks: sentiment analysis and part of speech (POS) tagging. How do the commonly used word embeddings differ from character level embeddings? The character level embeddings can be learned with a recurrent neural network (RNN) or a convolutional neural network (CNN). On a basic level CNNs are good at modelling data hierarchically and RNNs at modelling input in sequence. Iuliia uses a CNN for the tasks of sentiment analysis and part of speech tagging. To improve the accuracy on these tasks, hyperparameters such as the hidden size of the network (how many neurons do the hidden layers have?), the embedding size (how big is the context of the used embeddings), and the batch size (how many sample are looked at before adjusting the weights). By comparing different parameter combinations (probably by using a simple grid search) some good parameters can be found. The data the experiments are performed on is from the Stanford Sentiment Treebank, which contains annotated data from movie reviews, that comprises around 215,000 tokens. The implementation of the sentiment classifier and the POS tagger is done in Python using the Deep Learning library Theano. Iuliia next step is evaluating her results.