15.05.17

Protokoll zur Sitzung vom 15.05.2017 – Computerlinguistisches Arbeiten

3. Referat: Kristina Smirnov, Comparison of transfer methods for low-resource morphology (Katharina Kann)

The motivation for Kristina's dissertation is that there is not enough data for morphology tasks for all languages. The idea is to add data of a similar language to a low-resource data set in order to train the model. A variation of that approach is to add non-annotated data of the same language or the combination of both methods.

The model is written by her tutor Katharina Kann and Prof. Schütze and is based on a shared task (SIGMORPHON 2016 Shared Task). It is defined as a morphological encoder-decoder and is relatively successful with even small amounts of data. For her work Kristina, as a Russian native speaker, uses the languages Russian and Ukrainian to evaluate the model.

Her thesis can be subdivided into three main tasks: Combining annotated Russian samples with annotated Ukrainian samples. Combining annotated Russian samples with non-annotated Russian data and lastly combining the samples of all three categories.

Kristina uses the data set given in the CoNLL-SIGMORPHON 2017/2016 Shared Task, which are in the format: Lemma – target form – morph syntactic description. At this point Mr. Schulz notes that this format can lead to complications due to ambiguity.
As input for the model both low-resource annotated samples and high-resource annotated samples and their combinations are used. She mentions that for each sample group a separate training set, test set and development set is created.

Besides evaluating the system with statistical methods Kristina aims at analyzing the results on a morphological level, where she can use her expertise in Russian.