COMPARISON OF TRANSFER METHODS FOR LOW-RESOURCE MORPHOLOGY

Anton Serjogin
Centre for Information and Speech Processing, LMU
anton.serjogin@gmail.com

15.05.2017

- What can we make use of if there is not enough language data?
- Would it help to add data of similar language?

Morphological analysis and generation of previously unseen word forms is a fundamental problem in many areas of natural language processing (NLP). Its accuracy is crucial for the success of downstream tasks like machine translation and question answering. The work features the use of the SIGMORPHON 2016. which was a Shared Task on Morphological Reinflection as well as an extended analysis of how different design choices contribute to the final performance. The reasons for using the SIGMORPHON is that small amounts of data are sufficient in order to obtain successful results.

The main idea is to make the model learn morphological forms by running through a large number of combinations (starting from 50 and ending at 1600) of annotated/non-annotated language samples. For instance, the model may deliver a correct form of word enginds in plural forms and so on. The model features an encoder and decoder. The basic idea of the model is relatively simple: we have an RNN language model, but before starting calculation of the probabilities of E, we first calculate the initial state of the language model using another RNN over the source sentence F. The name "encoder-decoder" comes from the idea that the first neural network running over F "encodes" its information as a vector of real-valued numbers (the hidden state), then the second neural network used to predict E "decodes" this information into the target form. The scheme is as follows:

Lemma $\rightarrow$ Target Form $\rightarrow$ Morphosemantic description

The different forms are correlated with meanings which are labeled as 'features'. However, not all features that are identified through inflectional morphology are morphosyntactic. The most basic definition of a morphosyntactic feature is a feature which is relevant to syntax. Gender, number, and person are involved in agreement in a large number of languages, therefore they are typical morphosyntactic features. However, while in many familiar languages the feature 'tense' encodes regular semantic distinctions, it is not required by the syntax through the mechanisms of either agreement or government: syntax is not sensitive to the tense value of the verb. Therefore, many familiar instances of the feature 'tense' are morphosemantic, but not morphosyntactic.

The work includes following steps:

- Combine annotated Russian samples with annotated Ukranian samples
- Combine annotated Russian samples with non-annotated Russian samples
- Combine samples of all three categories
- Evaluate a graphical representation of the results
- Search for errors and improvements