COMPARISON OF TRANSFER METHODS FOR LOW RESOURCE MORPHOLOGY

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

22.05.2017

Motivation:

- It is about the paradigms of languages as well as allocation of lemmas to their inflected form
- High Resource languages vs Low Resource languages. It is difficult for low resource, because they are limited resources.
- Can a similar language be used in order to solve the problem?
- High Resource: Bulgarian Low Resource: Macedonian

As an approach we use 3 different methods for the paradigms:

- Cross-language paradigm completion sets. We search for a low resource language a similar (the similarity is really important) high resource language. The data from both languages is mixed and trained together. The biggest issue is to receive useful results.
- Auto Encoding. It is a pretty simple process where the input is simultanously an output. The aim is to have a maximum number of word inflections that are the same.
- Combination of both methods

In order to make the model process possible we need to have data in a certain format. The models get the source data as input and the result data as output. The training section consists of:

- Test-src, Test-trg
- Dev-src, Dev-trg
- Train-src, Train-trg
- Vocabulars for ”Source” and ”Target”

Error analysis:

- Often occurrence of the wrong ending
- Autoencoding delivers a number of mistakes, because of the method
- It is difficult to evaluate mistakes, when one does not have certain language skills

Other tasks:

- Analyzing other error sources
- Model
- Emphasize on the existing procedures