Leipzig University
Institute of Computer Science
Summer Term 2021

Prof. Dr. Andreas Maletti
Dr. Erik Paul

Machine Translation

Course Assignments

The practical aims to teach the generation of machine translation systems using the frameworks Moses and Tensorflow. More precisely, you will train

- a phrase based system using Moses,
- a syntax based system using Moses,
- a system based on recurrent neural networks using Tensorflow.

For the extended practical, you will in addition

- implement and document an approach to improve one of the above systems,
- train a transformer architecture neural translation system using Tensorflow.

You can freely choose the language and corpus used for training and you are free to change them between assignments.

Introduction — *Language model creation and corpus preparation*

Our first meeting serves as an introduction and aims to teach the following skills.

- working on a remote server using the command line
- compiling Moses
- preparing a corpus for use in an MT system: tokenization, recasing, cleaning
- creating a KenLM language model

Assignment 1 — *Phrase based translation system using* Moses

Your first task is to train and tune a phrase based translation system using Moses.

- In our live meeting, we will train and tune a small phrase based translation system using the French-English corpus of Europarl. If you choose to use the same corpus for the assignment, you will be provided with a fragment of the corpus.
- If you choose to use a different corpus, the resulting translation system must be non-trivial.
- If you choose to use tools or resources different from those presented in the live meeting (e.g. language model, training data, parser or wrapper in Assignment 2), please document this *shortly*.

Assignment 2 — *Syntax based translation system using* Moses

Your second task is to train and tune a syntax based translation system using Moses. *Please keep in mind the remarks from Assignment 1!*

Assignment 3 — *RNN based translation system using* TENSORFLOW

Your third task is to train an RNN based translation system using TENSORFLOW. For this, you will be provided with a PYTHON script which partially implements a suitable Encoder-Decoder RNN architecture. Your task is to add the missing architecture of the encoder and use the script to train a translation system.
Please ensure that the script retains its functionality to load trained systems!

Extension Assignment 1 — *Improved translation system*

Your first task for the extended practical is to implement and document an approach to improve one of the systems from the first three assignments. The focus of this assignment lies on the correct implementation and documentation of the approach. Achieving an actual improvement is secondary. So do not hesitate to try your own ideas!
You should document

- a description of the approach and how you implemented it
- the training data used
- a quality comparison between the improved system and the original system

To give a better idea, the following techniques are examples of approaches commonly employed to improve machine translation systems.

- data augmentation, for example through synonym replacement
- subword tokenization
- stemming und lemmatization
- coreference resolution
- pre-trained word-embeddings (only TENSORFLOW)

Extension Assignment 2 — *Transformer based system using* TENSORFLOW

Your second task for the extended practical is to train a neural translation system based on the transformer architecture using TENSORFLOW. As in Assignment 3, you will be provided with a PYTHON script which partially implements a suitable network architecture. Your task is to add the missing architecture of the decoder and use the script to train a translation system.
Please ensure that the script retains its functionality to load trained systems!