Using Aligned Word Embedding Spaces for Orthography and Script Normalization

- **Supervisor**:
- **Examiner**: Prof. Hinrich Schütze
- **Open**: MSc
- **Summary**: Sometimes one language could be written in various conventions and different sources of text could employ different scripts and/or orthographies. These variations render data from different sources non-comparable. Recent study [1] has shown that normalizing different orthographies and scripts is beneficial to the performance of various downstream tasks. To train a model that performs such normalization, parallel word pairs of different scripts and orthographies are essential. However, these kinds of pairs can be difficult to obtain, especially when parallel corpora are unavailable. To tackle this problem, this project proposes to make use of unsupervised dictionary induction methodology such as MUSE [2], and retrieve parallel word pairs by aligning monolingual word embedding spaces, without using any parallel corpora. Prior linguistic knowledge or hand-crafted rules can be used to further filter a subset of correspondence from the initial word pairs. The expected outcome is a model that can automatically normalize differences in orthographies and scripts for a wide range of languages which only have monolingual data.

- **Prerequisites**: enthusiasm, interest for multilingual NLP, good programming background (preferably Python), good command of DL framework

[1] https://aclanthology.org/2023.acl-long.809.pdf
[2] https://arxiv.org/pdf/1710.04087.pdf