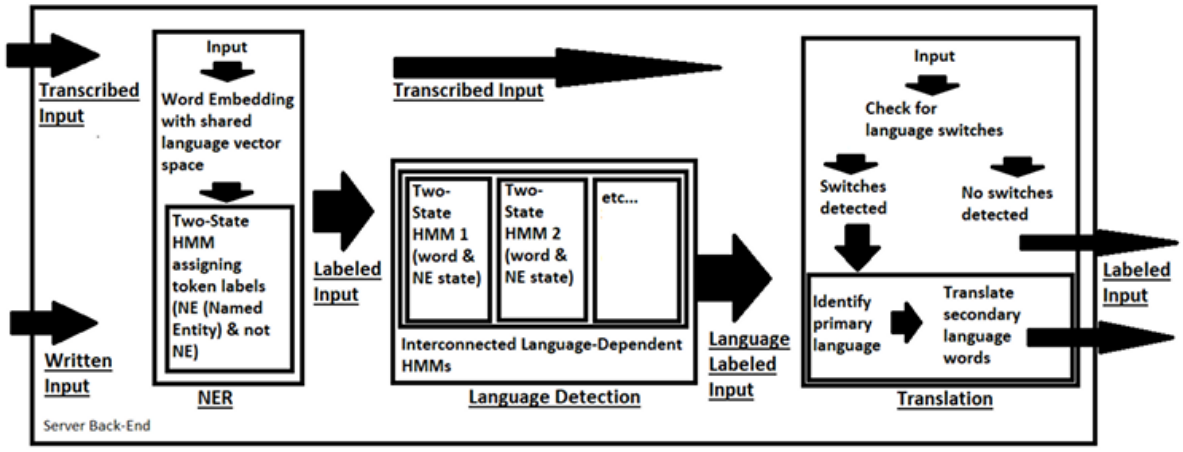
Proposal for a Master's Thesis in the Language Technology track

In my Bachelor's thesis I created a theoretical framework for a preprocessing module in a multilingual personal digital assistant. A sketch of it is included below.



In this preprocessing module I divided the task of processing possible multilingual and code-switched input into three steps: Named-Entity-Recognition, Language border detection and partial translation.

The three steps are based on tested models presented in peer-reviewed papers and have been shown to work well for their respective tasks.

As this work was purely theoretical, my proposal for a master's thesis topic is therefore to create a part of this module in practice. It is important for my proposal to note, that the models used as the base for each of the steps do not use the possibilities of Machine Learning and Neural Networks so far.

I would like to try to create the NER step of my module and compare different approaches to the task on their performance in accuracy and processing speed as well as needed storage space, as this is an equally as important feature to consider in a PDA.

My plan is to first use my bachelor's thesis as a base instruction to create the HMM-based model for named entity recognition in a multilingual context. This includes creating a

multilingual vector space for embeddings and building as well as training the HMM to recognize the named entities reliably.

In the next step, I would like to apply the task to a Feed-Forward Neural Network. This step includes optimization steps in order to achieve the highest possible accuracy in labelling as well as keeping the processing speed as high as possible.

Finally, as transformer architectures are the current state of the art for a variety of NLP tasks, I would like to use the Huggingface library to fine-tune an existing transformer architecture (for example a BERT system) in order to perform the NER task.

Finally, the performance for all the optimization steps and final models will be compared and ranked based on their accuracy, required storage space and speed.

The result of my master's thesis would be a decision which model could be used in a real-life application for the multilingual PDA with the given parameters and whether my initial proposal still holds up with the current neural network technologies.