Deep Case: data and probing

- **Supervisor**: Yihong Liu
- **Examiner**: Prof. Schütze
- **Bsc, Msc, Open**: MSc/Bsc
- **Summary**: Noun case is an interesting and important linguistic feature which explicitly indicates the function of a noun in a noun phrase, or more globally, in a sentence. Some languages, e.g., English and Chinese, do not have cases. In contrast, some languages, e.g., German and Russian, do use case markers. However, the case categories in those languages often differ in their numbers. Different case numbers mean that some functions of cases in different languages can overlap. The mapping of the case, however, is often not strictly one-to-one, one-to-many or many-to-one, because each language can form language-specific characteristics on the case system. Therefore, mapping different cases across languages would be a challenging task. Interestingly, we see that supervised machine translation models, although not trained in case mapping, can often translate to correct nouns with correct case markers. Therefore, in this task, we would like to explore the following (not necessarily cover all of them):

  (1) Since we are interested in noun cases cross-lingually, we should explore the cases in different languages. Unfortunately, as far as we know, there are no such a dataset for sequentially tagging the noun cases. Therefore, the first task is to build such a dataset for multiple languages.
  (2) Based on (1), set up baselines for the sequential case tagging in each involved language.
  (3) Probe if PLMs are able to correctly map the noun cases in one language to cases in another language. As a simple example, if we train a sequential case tagger by fine-tuning some PLM layers on one language, e.g., Finnish (15 noun cases), or Hungarian (18 noun cases) can we use the same model to perform case tagging on other languages, e.g., German (4 noun cases) and Russian (6 noun cases).

- **Prerequisites**: enthusiasm (publishing the results of the thesis at a conference/workshop), good programing background (preferably python), basic knowledge of NLP, a basic command of DL framework (preferably PyTorch), probably knowledge in a language that has noun cases