Project Description

Project title:
BERT2Punc

Goal:
This project aims to use NLP transformers to predict where to insert punctuation in non-punctuated texts. Specifically, a pretrained BERT model from the Transformers repository made by Hugging Face will be modified s.t. it can predict punctuation in English Wikipedia texts (which have been stripped from punctuation). Initially, the modification of the BERT model will simply entail adding a linear layer with softmax activation to the head of the BERT model s.t. the model outputs a probability distribution over the set of possible punctuations.

If we prove successful in this endeavor, and time allows for it, we will also try to use a Nordic BERT model to predict punctuation in Danish Wikipedia texts. The Nordic BERT model is a BERT model which has been trained on Nordic texts. The reason why punctuation prediction in Danish texts is not our first priority is that it is likely to require more data preprocessing and model code as the Nordic BERT model is not included in the Transformers repository. Therefore, we would rather get something "simple" up and running first and then add other functionalities and models later.

Framework, Data, and Models:
As specified above, we will use the Transformers repository made by Hugging Face for loading the pretrained BERT model. Furthermore, the dataset consisting of English Wikipedia texts will be downloaded using the Transformers repository as well. This data will be processed such that training and test data become texts stripped from punctuations, and targets/labels the original placements of the punctuations. Finally, functionalities implemented in the Transformers repository such as tokenizers, training loops, and metrics will to a large extent be used for data preprocessing, fine-tuning/training and evaluation of our modified BERT model.

If time allows, the weights of the Nordic BERT model will be downloaded from the Nordic BERT repository: https://github.com/botxo/nordic_bert, and the Danish Wikipedia texts will be downloaded from https://da.wikipedia.org/wiki/Forside. Due to the Transformers repository not including Danish Wikipedia texts, we will most likely have to implement code for preprocessing the Danish data ourselves.