Deep Text Complexity Metric Model

This report documents our model submission for the **Deep Text Complexity Metric Challenge** of the *Quality & Usability Lab of the Technical University of Berlin* in the summer term of 2021.

Task description

The task in this challenge was to predict the perceived complexity of German sentences with a (deep) machine learning model. For that, a **dataset of 900 German sentences** and their respective ratings was given. The ratings were given in form of MOS (Mean Opinion Score) values, which are **continuous numerical values from 1-7**. Those were collected by surveying L2 (second language) learners of German and further post-processing and expert assessment. Details about the dataset creation process can be found in the following paper: 'Subjective Assessment of Text Complexity: A Dataset for German Language'

In the field of task categories which are being worked on in machine learning and specifically natural language processing (NLP), this task falls in the category of **sequence classification** or more exactly **sequence regression**.

BERT encodings

A central step in most current NLP methods is the numerical encoding of natural language (i.e. word, sentences, text documents) into a representation that can be further used in e.g. a classification model. One state-of-the-art approach to this problem is BERT, which uses the Transformer architecture to produce said representations.

By now (July 2021) there exist several adaptations of the BERT model which try to improve it or cater to certain needs. Apart from the differences in model architecture, the performance of a specific model depends heavily on the training data, especially when it comes to its application for different languages. Due to the high resource demands in data, time and computing power to train a well performing language model with an architecture like BERT, it is common to build upon pre-trained models which are made publicly available.

To our knowledge the state-of-the-art general purpose language models for German are the GBERT and GELECTRA models which have been published in October 2020. The huggingface library offers a convenient way to integrate Transformer-based models into custom models and applications and pre-trained models can easily accessed via their model hub.

Submitted model

Our submitted model follow the architecture for sequence classification as it was described in the original BERT paper. The processes of predicting the MOS value for an input sentence consists of the following steps:

1. Tokenization:
   a) Tokenizing the sentence.
   b) Adding [CLS] and [SEP] tokens.
   c) Converting the tokens to integer ids.
2. Passing the token ids through the BERT layers.
3. Pooling:
   a) Extracting the last hidden sate of the [CLS] token.
   b) Passing it through a dense linear layer.
   c) Applying a tanh activation layer.
4. Regression:
   a) Applying a dropout layer.
   b) Projecting to a single float value with another linear layer.

Diagram of model architecture

As our base BERT model we used `gbert-base`, the smaller of the two GBERT models. We found, that with the larger model and the GELECTRA models there was a significantly higher variance in evaluation error when using different random seeds for weight initialization and data splitting/shuffeling.

Training process

epochs, learning rate, optimizer, huggingface trainer -> ez life

Discarded models

Mumbo Jumbo

Contributors

This work was a collaboration of Faraz Maschhur (@fmaschhur), Chuyang Wu and Max Reinhard (@wuxmax).