DS-UA 203: Machine Learning for Language Understanding
Course Project Proposal

Charles Chang, Tinlong Liao, Kristine Zeng

This research project aims to evaluate the accuracy of Masked Language Models in predicting words in texts with different Flesch Reading Ease scores. The Flesch Reading Ease (FRE) score is adopted to distinguish the texts as it is a formula that measures readability and linguistic complexity based on syllables per word and words per sentence. The range of the possible FRE scores is 0 – 121.22, with smaller values indicating lower readability and higher difficulty. The FRE score of a plain English text that could be easily understood by 8^{th} and 9^{th} Grade students typically lies in the range of 60 – 70. FRE score is widely used to ensure political and legal documents are comprehensible to the general public. Therefore, it is a standardized formula that can properly quantify the difficulty of the texts in the dataset. With it, this project could show whether the performance of machine learning models is consistent with humans in dealing with easy and hard texts. The results could be helpful in deciding on specific models or human resources to be used to restore information missing in significant documents.

The dataset to be used is RACE, which is a large-scale reading comprehension dataset with more than 28,000 passages and nearly 100,000 questions created by English instructors. However, only the passages will be used for this research paper. The passages cover a wide range of topics that are designed for middle school and high school students in China. This RACE dataset was collected from English examinations in China and published and presented by five college students from Carnegie Mellon University in their empirical paper RACE: Large-scale ReAding Comprehension Dataset From Examinations. Furthermore, their goal for creating the dataset was to “serve as a valuable resource for research and evaluation in machine comprehension”. We choose this dataset because we are interested in how well the Masked Language Modeling will do in predicting a random sample of input tokens that have been replaced by a mask placeholder in a variety of Flesch Reading Ease scores.

While most pertaining language models are suitable for performing this experiment, we choose three particular models: BERT, RoBERTa, and T5. These three widely-cited models all use masked language modeling in the pre-training period, so we assume they are suitable to perform our experiment. We intend to split our dataset into two portions: train set and test set. After training our masked language models in a self-supervised manner, we then report the accuracy of the test set. We are interested in two things: a) Does the accuracy on the test set have anything to do with the readability of text? b) Is that previous relation consistent across three pre-trained language models?