English Language Proficiency Evaluation

Executive Summary

Mentor: Kash Bari
Team Members: Abhinav Chand, Arpan Pal, Moyi Tian
Mengting Chao, Zhaobidan Feng

Dec 09, 2022

1 Motivation and Goal

The group of students learning English as a second language, known as English Language Learners (ELLs), has rapidly grown. While automated feedback tools make it easier for teachers to assign more writing tasks, they are not designed with ELLs in mind. Existing tools are unable to provide feedback based on the language proficiency of the student, resulting in a final evaluation that may be skewed against the learner. We hope to improve automated feedback tools to better support the unique needs of these learners. Our goal is to utilize a dataset of essays written by ELLs to develop English language proficiency evaluation models, which help ELLs receive more accurate feedback on their language development and expedite the grading cycle for teachers.

2 Dataset

The dataset (the ELLIPSE corpus) comprises argumentative essays written by 8th-12th grade ELLs. The essays have been scored according to six analytic measures: cohesion, syntax, vocabulary, phraseology, grammar, and conventions. Each measure represents a component of proficiency in essay writing, with greater scores corresponding to greater proficiency in that measure. The scores range from 1 to 5 in increments of 0.5. Our task is to predict the scores for each of the six measures for the essays in the test set with train test split 80 : 20.

3 Methodology

Our evaluation metric is the mean column-wise root mean squared error (MCRMSE). We use the mean score of each analytic measure from the train set as our baseline. Firstly, we use pretrained BERT to train through partial LSTM decoder and dense output layer. Secondly, we try pretrained BERT plus XGBoost regression. In addition, we propose a model tailored to the grammar measure. We first train a BERT model on the Corpus of Linguistic Acceptability (CoLA) dataset, and then use this pretrained model as a single sentence grammar checker to obtain ratios of grammatically correct sentences for the essays, and lastly use the preprocessed data to train against grammar measure through XGBoost regressor. Next, we replace the grammar column training in the second model with this proposed grammar model. Finally, we train DeBERTa model using its disentangled attention mechanism, apply mean pooling on the last hidden state and then add dense output layer.

4 Conclusion

- DeBERTa plus pooling has the best performance. It improves MCRMSE by 32.6% compared with the baseline.
- The transformer models (e.g. BERT and DeBERTa) perform better than traditional neural network architectures (e.g. LSTM and feedforward dense layers) and tree classification models (e.g. XGBoost) in English language proficiency evaluation. To potentially improve the performance of the transformer models, we can explore more variations of pooling like LSTM pooling, attention pooling, weighted layer pooling, etc. We can also extend training time or epochs on BERT and DeBERTa, which will require more computing resources.
- It would be better to apply data augmentation, which helps deal with imbalance of data. Furthermore, different preprocessing methods tailored to our measures may help extract important features for training.
- Better interpretability of our models can provide more detailed features and feedback on essays to help language learners and teachers.