Kaggle Competition: Automated Essay Scoring 2.0

Yuquan Zhou Qin Ma Bo Li

Abstract

AES technology is meaningful in the field of education. We use deep learning methods to enhance this technology and test our results in the Kaggle competition (https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2).

Project Background

Rising education costs have led to pressure to hold the educational system accountable for results. The advance of AES technology promises to measure educational achievement at reduced cost.

Project Aim

We hope to achieve high scores in Kaggle competitions by allowing the model to fully comprehend the content of the articles and provide corresponding ratings.

Literature Review

A lot of traditional work on AES has focused on feature engineering. McNamara et al.'s hierarchical classification approach (McNamara et al., 2015[1]) uses linguistic, semantic and rhetorical features. But many recent AES systems are neural-based. Dasgupta et al.'s work (Dasgupta et al., 2018[2]) augments textual features in deep convolution recurrent neural network, which inspires us.

Due to page limitations, detailed investigation of related work will be presented in the PPT.

Method

1. Dataset: We use the dataset provided on Kaggle competition AES 2.0. To prevent participants from overfitting the test set, the competition does not offer a reliable test set, making the task challenging. The train set comprises about 24000 student-written argumentative essays and each essay was scored on a scale of 1 to 6.

2. Basic Approach: We obtain the meta feature of the entire article through the pre-trained large model DeBERTa, then obtain the features of lexical diversity, cohesion, causality, and informativeness of a text through feature engineering. By combining the features from both, we get the representation of each article. Finally, we use LightGBM to replace the fully connected layer to classify the articles according to their scores (1-6). We may use LoRA or other fine-tuning techniques that require fewer computational resources to fine-tune DeBERTa and improve the accuracy of our approach.

3. Technical challenge: Fine-tuning large models is computationally expensive, and due to the limited number of training samples, selecting appropriate features through cross-validation may not be reliable.

Expected Results

We aim to achieve good results on the quadratic weighted kappa evaluation metric in the Kaggle competition, which measures the agreement between true scores and predicted scores.

References

[1] Danielle S McNamara, Scott A Crossley, Rod D Roscoe, Laura K Allen, and Jianmin Dai. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59.

[2] Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. 2018. Augmenting Textual Qualitative Features in Deep Convolution Recurrent Neural Network for Automatic Essay Scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 93–102, Melbourne, Australia. Association for Computational Linguistics.