System Report for CCL24-Eval Task 7: Assessing Essay Fluency with Large Language Models

Haihong Wu,Chang Ao,Shiwen Ni
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
{ haihongw@mail.ustc.edu.cn, c.ao@siat.ac.cn,sw.ni@siat.ac.cn }

Abstract

With the development of education and the widespread use of the internet, the scale of essay evaluation has increased, making the cost and efficiency of manual grading a significant challenge. To address this, The Twenty-third China National Conference on Computational Linguistics (CCL2024) established evaluation contest for essay fluency. This competition has three tracks corresponding to three sub-tasks. This paper conducts a detailed analysis of different tasks, employing the BERT model as well as the latest popular large language models Qwen to address these sub-tasks. As a result, our overall scores for the three tasks reached 37.26, 42.48, and 47.64.

Keywords: Large Language Models, Assessing Essay Fluency, Fine-tuning

1 Introduction

CCL points out that with the development of education, the quantity of essay texts is gradually increasing. The cost and efficiency of manually grading essays have become a major challenge. Many research institutions aim to provide objective, accurate, and timely scoring and feedback by analyzing the language, content, and structure of essays. Among these factors, the fluency of expression is an important aspect of essay evaluation for teachers. Therefore, this competition revolves around evaluating the fluency of essay texts. There are three tracks in this competitionincluding:

Track 1: Identification of sentence types in primary and secondary school essays. the identification of incorrect sentence types in primary and secondary school essays is a multi label classification problem, which predicts which types of incorrect sentences a sentence is. The sick sentence type label contains both lexical and syntactic errors. In this evaluation task, a total of 5 coarse-grained error types and 14 fine-grained error types were defined.

Track 2: Revision of sentence types in primary and secondary school essays. the task of rewriting incorrect sentences in primary and secondary school essays is a text generation task, which involves inputting incorrect sentences and outputting modified ones.

Track 3: Evaluation of fluency in primary and secondary school essays. The fluency rating task for primary and secondary school essays is a multi classification task, which involves inputting an essay and outputting its level of fluency. This evaluation task defines three fluency levels: excellent, average, and failing.

Three tracks provide more basis for evaluating the fluency of primary and secondary school essays and offering higher-quality assessments.

2 Model and Methods

2.1 Model

In track1 and track2, fine-tuning the Qwen 7b model has proven highly effective, leveraging its superiority in processing Chinese texts to achieve excellent results. In addition, the BERT model (Devlin et al. (2018)) has a much smaller number of parameters than large language models, but it performs extremely well in classification tasks. In track 3, the BERT large model was employed for the classification task of essay fluency, and the results demonstrated the excellent performance of the BERT model as well.

2.1.1 Qwen

The Qwen-1.5-7B-Chat (Bai et al. (2023)) model is a large-scale language model based on the transformer (Vaswani et al. (2017)) architecture, designed primarily for conversational applications. With a parameter count of 7 billion, it belongs to the class of state-of-the-art transformer-based models, known for their ability to handle complex natural language understanding and generation tasks. The Qwen-1.5-7B-Chat model is built on the transformer architecture, which enables it to capture long-range dependencies in text data efficiently. The model is trained on diverse datasets encompassing various topics and language styles, ensuring robust language understanding and generation capabilities across different contexts.

2.1.2 Bert

The BERT-Large model (Devlin et al. (2018)) is a high-capacity language model based on the transformer architecture, designed for a wide range of natural language processing tasks. With a parameter count of approximately 340 million, it belongs to the class of state-of-the-art transformer-based models, renowned for their proficiency in handling intricate language understanding and generation tasks. The BERT-Large model leverages the transformer architecture's capability to efficiently capture long-range dependencies in text data. It is pre-trained on extensive datasets covering diverse topics and language styles, ensuring robust language understanding and generation abilities across various contexts.

To incorporate a formulaic description into this paragraph, we can highlight the pre-training process mathematically:

$$\mathbf{M}_{\text{BERT-Large}}(\theta) = \arg\min_{\theta} \sum_{(x,y)\in\mathcal{D}_{\text{pre}}} \mathcal{L}_{\text{BERT}}(\mathbf{M}_{\text{BERT-Large}}(x;\theta), y)$$

Here, $\mathbf{M}_{\text{BERT-Large}}(\theta)$ represents the BERT-Large model with parameters θ. $\mathcal{D}_{\text{pre}}$ denotes the pre-training dataset consisting of diverse texts. $\mathcal{L}_{\text{BERT}}$ is the loss function that measures the discrepancy between the model's predictions $\mathbf{M}_{\text{BERT-Large}}(x;\theta)$ and the true labels y.

This formulation emphasizes that the BERT-Large model is trained by minimizing the loss function $\mathcal{L}_{\text{BERT}}$ over a comprehensive pre-training dataset $\mathcal{D}_{\text{pre}}$, ensuring its readiness to comprehend and generate language across varied domains and styles.

2.2 Fine-tuning

Fine-tuning (Friederich (2017)) is a term commonly used in the fields of machine learning (Jordan and Mitchell (2015)) and deep learning (LeCun, Bengio, and Hinton (2015)). It refers to the process of further optimizing and adjusting a pre-trained model to enhance its performance on a specific task. Full-Parameter fine-tuning of large language models leverages the vast amounts of knowledge embedded in the pre-trained model while adapting it to perform better on specific tasks, such as text classification, question answering (Lv et al. (2023)). Previous work has also shown that fine-tuning the BERT model yields superior performance on multi-class classification tasks (Sun et al. (2019)).

The fine-tuning process can be described mathematically as follows:

1. Pre-trained Model: Let $\mathbf{M}_{\text{pre}}(\theta)$ be the pre-trained model with parameters θ. This model has been trained on a large corpus of general data.

2. Task-specific Dataset: Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be the task-specific dataset, where x_i represents the input data and y_i represents the corresponding labels.

3. Loss Function: Define a task-specific loss function $\mathcal{L}(\mathbf{M}_{\text{pre}}(x_i;\theta), y_i)$ that measures the difference between the model's predictions and the true labels.

4. Optimization Objective: Fine-tuning aims to minimize the loss function over the task-specific

dataset. This can be expressed as:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(\mathbf{M}_{\text{pre}}(x_i; \theta), y_i)$$

We conducted fine-tuning with training data on the Qwen-1.5-7B-Chat model. This fine-tuning process enhances the model's performance on specific tasks by adapting its parameters to the target dataset. Given the Qwen model's generally strong capability in understanding Chinese, we decided to fine-tune the Qwen model in the track1 and track2 to achieve assessment and modification of essay sentences.

Let $\mathbf{M}_{\text{Qwen}}$ denote the Qwen-1.5-7B-Chat model. The fine-tuning process can be formulated as:

$$\theta^*_{\text{Qwen}} = \arg\min_{\theta_{\text{Qwen}}} \sum_{i=1}^{N} \mathcal{L}_{\text{Qwen}}(\mathbf{M}_{\text{Qwen}}(x_i; \theta_{\text{Qwen}}), y_i)$$

Fine-tuning with training data on the BERT-Large model enhances its performance on specific tasks by customizing its parameters to suit the track3 dataset. This process adapts the model's learned representations to better align with the nuances and intricacies of the particular task at hand, resulting in improved task-specific performance and accuracy. Therefore, we decided to fine-tune the BERT-Large-Chinese model in the third track to achieve fluency assessment of essays.

Let $\mathbf{M}_{\text{BERT}}$ denote the BERT-Large model. The fine-tuning process for the BERT-Large model can be formulated as:

$$\theta^*_{\text{BERT}} = \arg\min_{\theta_{\text{BERT}}} \sum_{i=1}^{N} \mathcal{L}_{\text{BERT}}(\mathbf{M}_{\text{BERT}}(x_i; \theta_{\text{BERT}}), y_i)$$

In summary, the fine-tuning process involves adapting the parameters θ of pre-trained models $\mathbf{M}_{\text{pre}}$ to minimize the task-specific loss function over the dataset $\mathcal{D}$, thereby improving the model's performance on specific tasks.

2.3 multi-prompt context learning

Multi-prompt context learning, as discussed in Chen et al. (2024), is an innovative approach in natural language processing (NLP) that aims to enhance a model's capability to understand and process textual contexts. Traditional NLP models often operate with a single context prompt or focus on a single task, which can limit their ability to fully grasp the nuances of complex textual information. In contrast, multi-prompt context learning introduces multiple context prompts, enabling the model to better comprehend and adapt to various facets of the input data.

In the inference phase of the track2, a multi-prompt approach was employed. This method involves providing the model with several prompts or questions, which allows it to generate more comprehensive and diverse responses. By leveraging multiple prompts, the model can investigate different dimensions of the input data, leading to richer and more accurate inferences. The theoretical underpinning of this approach can be expressed with the following formalism:

Let $P = \{p_1, p_2, \ldots, p_n\}$ denote the set of prompts provided to the model, where p_i represents an individual prompt. The model's response R can be described as a function f of the input data D and the set of prompts P:

$$R = f(D, P) = f(D, \{p_1, p_2, \ldots, p_n\})$$

Each prompt p_i explores a different aspect of the input data D, contributing to the overall understanding and response generation. The individual responses r_i corresponding to each prompt p_i can be aggregated to form a final comprehensive response R:

$$r_i = f(D, p_i) \quad \text{for} \quad i = 1, 2, \ldots, n$$

The aggregation function g combines these individual responses r_i to produce the final response R:

$$R = g(r_1, r_2, \ldots, r_n)$$

This aggregation can be achieved through various techniques such as averaging, voting, or more sophisticated ensemble methods. The multi-prompt approach thereby allows the model to harness diverse perspectives from the input data, enhancing the overall quality and accuracy of its inferences.

By integrating multiple prompts, the model can effectively:

1. Capture a wider range of contextual information.
2. Mitigate the risk of bias associated with single-prompt models.
3. Improve robustness and reliability of the generated responses.

In summary, multi-prompt context learning represents a significant advancement in NLP, facilitating more nuanced and precise understanding of textual data through the use of multiple, complementary prompts. This approach not only broadens the scope of information the model can process but also enhances its adaptability to varied and complex contexts.

3 Experiments

During the experiment, we used the transformer library to load the Qwen-7b and Bert-large-chinese pre-training model, trained the model with the datasets announced by the competition organizer, and optimized the model parameters through backpropagation and gradient descent algorithms. The hyper-parameter settings during model training are shown in table1 and table2.

Hyperparameter	Value
Train Epochs	3.0
Learning Rate	5×10^{-5}
Batch Size	16

Table 1: Hyperparameters of Training Qwen

Hyperparameter	Value
Train Epochs	15
Learning Rate	1×10^{-5}
Batch Size	8

Table 2: Hyperparameters of Training BERT

3.1 Evaluation Results for Track 1

Metric	Score
Micro-F1	48.84
Macro-F1	25.68
Coarse-grained Micro-F1	58.06
Fine-grained Micro-F1	39.62
Coarse-grained Macro-F1	36.25
Fine-grained Macro-F1	15.10
Score	37.26

Table 3: Evaluation Metrics Scores

3.1.1 metric results

The Micro-F1 score of 48.84 indicates a moderate level of overall accuracy, considering both precision and recall across all instances. In contrast, the Macro-F1 score of 25.68 is significantly lower, suggesting that the model's performance varies widely across different classes, with some classes potentially having much poorer performance.

The Coarse-grained Micro-F1 score of 58.06 is higher than the overall Micro-F1 score, indicating that the model performs better when evaluated on broader categories. Similarly, the Coarse-grained Macro-F1 score of 36.25 is higher than the overall Macro-F1 score, reinforcing the idea that the model's performance is more consistent at a higher level of categorization.

The Fine-grained Micro-F1 score of 39.62 and the Fine-grained Macro-F1 score of 15.10 are both lower than their coarse-grained counterparts. This suggests that the model struggles with finer distinctions between categories, which is a common issue in complex classification tasks.

The overall score of 37.26 provides a single summary metric, but it should be interpreted in the context of the more detailed metrics. It reflects the model's average performance but does not capture the nuances revealed by the other scores.

3.1.2 Analysis

The evaluation metrics reveal a model that performs moderately well overall but has significant room for improvement, especially in distinguishing fine-grained categories. Future work should focus on improving the model's ability to handle finer distinctions between classes, as well as addressing any imbalances in class performance that are suggested by the low Macro-F1 scores.

3.2 Evaluation Results for Track 2

Metric	Score
EM	13.93
Bert PPL	15.82
Levenshtein	1.90
BLEU-4	89.60
BertScore	97.34
Precision	45.22
Recall	27.25
F0.5	39.95
Score	42.48

Table 4: Evaluation Metrics Scores

3.2.1 metric results

The result is shown on table 2,The evaluation metrics provide a comprehensive insight into the model's performance. The Exact Match (EM) score of 13.93 indicates that the model's output exactly matches the reference answers only a small fraction of the time, highlighting a need for improvement in generating fully correct answers. The perplexity (Bert PPL) score of 15.82 suggests moderate fluency and coherence in sentence generation, but there is room for refinement. A Levenshtein distance of 1.90 reflects a high similarity between the model's outputs and the reference answers, though not perfectly aligned. The BLEU-4 score of 89.60 demonstrates the model's excellent ability to capture the phrase structures of the reference answers. A BertScore of 97.34 shows that the model's generated answers are semantically very close to the reference, indicating strong contextual understanding. However, a precision score of 45.22 reveals that less than half of the model's generated answers are correct, and a recall score of 27.25 indicates significant omissions of correct answers. The F0.5 score of 39.95, which weighs precision more heavily, further underscores the need for improvement in balancing accuracy and completeness in the model's outputs.

3.2.2 analysis

The model performs well in semantic understanding and generating coherent text (with high BertScore and BLEU-4 scores), but there is significant room for improvement in exact matching and accurate generation (as indicated by lower EM and Precision scores). The relatively high Levenshtein score suggests a degree of similarity between the model's outputs and the reference answers, but they are not exactly the same. The perplexity score indicates moderate generation capability but not optimal performance. Overall, the model shows promise in generating high-quality text but needs to enhance its accuracy and comprehensive coverage of reference answers.

Finally, the overall score of 42.48 further highlights that while the model has some strengths, particularly in semantic similarity and structure capture, it requires significant enhancements in precision and recall to achieve better performance.

3.3 Evaluation Results for Track 3

Metric	Score
ACC	48.59
Precision	46.42
Recall	43.07
F1	43.44
QWK	0.1338
AvgScore	47.64

Table 5: Evaluation Metrics Scores

3.3.1 metric results

The evaluation metrics indicate that the model's overall performance is moderate but has significant room for improvement. The accuracy (ACC) of 48.59 means that slightly less than half of the model's predictions are correct, pointing to a need for better overall accuracy. The precision of 46.42 shows that less than half of the model's positive predictions are correct, suggesting a fair number of false positives that need reduction. With a recall of 43.07, the model correctly identifies a significant portion of the actual positive cases, but it still misses many true positives. The F1 score, which balances precision and recall, is 43.44, reflecting the need for improvement in both areas to achieve better overall effectiveness. The Quadratic Weighted Kappa (QWK) score of 0.1338 indicates poor agreement between the model's predictions and the true labels beyond random chance, highlighting the necessity for significant enhancements in prediction quality. Finally, the average score (AvgScore) of 47.64 summarizes the model's performance across different metrics, suggesting that while the model performs moderately well in some aspects, it generally underperforms and requires improvements in various areas to achieve better overall effectiveness.

3.3.2 analysis

The model demonstrates moderate performance with an accuracy of 48.59, indicating that slightly less than half of its predictions are correct. The precision of 46.42 and recall of 43.07 suggest that the model produces a fair number of false positives and misses a significant number of true positives. The F1 score of 43.44 highlights the need for balanced improvements in both precision and recall. The low QWK score of 0.1338 points to poor agreement with the true labels, indicating that the model's predictions are often inaccurate beyond what could be attributed to random chance.

Finally, the average score of 47.64 underscores the overall moderate performance of the model, suggesting that substantial enhancements are needed across all evaluated metrics to achieve better accuracy, precision, recall, and agreement with true labels.

4 Conclusion

In this experiment, we conducted a comprehensive assessment of the fluency of essays and validated it through a competition format. The results indicate that large language models perform exceptionally well in text correction for evaluating essay fluency, while BERT models are more effective in essay fluency classification. In future research, we will further explore evaluation methods for essay fluency, integrating theories from linguistics, psychology, and related fields, in order to seek more scientifically effective evaluation metrics and methods.

References

Bai, Jinze et al. (2023). *Qwen Technical Report*. arXiv: 2309.16609 [cs.CL].

Chen, Haoran et al. (2024). “Multi-prompt alignment for multi-source unsupervised domain adaptation”. In: *Advances in Neural Information Processing Systems* 36.

Devlin, Jacob et al. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.

Friederich, Simon (2017). “Fine-tuning”. In: *The Stanford encyclopedia of philosophy*.

Jordan, Michael I and Tom M Mitchell (2015). “Machine learning: Trends, perspectives, and prospects”. In: *Science* 349.6245, pp. 255–260.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). “Deep learning”. In: *nature* 521.7553, pp. 436–444.

Lv, Kai et al. (2023). “Full parameter fine-tuning for large language models with limited resources”. In: *arXiv preprint arXiv:2306.09782*.

Sun, Chi et al. (2019). “How to fine-tune bert for text classification?” In: *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*. Springer, pp. 194–206.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30.