Task-Related Pretraining with Whole Word Masking for Chinese Coherence Evaluation

Ziyang Wang[1,2], Sanwoo Lee[1,3], Yida Cai[1,2], and Yunfang Wu[1,3(✉)]

[1] MOE Key Laboratory of Computational Linguistics, Peking University, Beijing, China
[2] School of Software and Microelectronics, Peking University, Beijing, China
[3] School of Computer Science, Peking University, Beijing, China
wuyf@pku.edu.cn

Abstract. This paper presents an approach for evaluating coherence in Chinese middle school student essays, addressing the challenges of time-consuming and inconsistent essay assessment. Previous approaches focused on linguistic features, but coherence, crucial for essay organization, has received less attention. Recent works utilized neural networks, such as CNN, LSTM, and transformers, achieving good performance with labeled data. However, labeling coherence manually is costly and time-consuming. To address this, we propose a method that pretrains RoBERTa with whole word masking (WWM) on a low-resource dataset of middle school essays, followed by finetuning for coherence evaluation. The WWM pretraining is unsupervised and captures general characteristics of the essays, adding little cost to the low-resource setting. Experimental results on Chinese essays demonstrate that this strategy improves coherence evaluation compared to naive finetuning on limited data. We also explore variants of their method, including pseudo labeling and additional neural networks, providing insights into potential performance trade-offs. The contributions of this work include the collection and curation of a substantial dataset, the proposal of a cost-effective pretraining method, and the exploration of alternative approaches for future research.

Keywords: Coherence evaluation · pretraining · low-resource · Chinese computing

1 Introduction

In Chinese National College Entrance Examination and Senior High School Entrance Examination, evaluating the writing quality of essays has been a time-consuming task whose results might lack consistency when evaluated by human raters. Previous essay assessment tasks have focused on leveraging linguistic features of essays, such as those related to rhetoric and idioms.

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14304, pp. 314–322, 2023.
https://doi.org/10.1007/978-3-031-44699-3_28

Coherence is a fundamental concept in essay assessment and is particularly useful to assess how well an essay is organized. Coherence can be broken down to the cohesion between sentences and the fluency of transitions between paragraphs. It plays a vital role in ensuring clarity, conciseness and fluency in an essay, which is also crucial in improving the overall writing quality.

While early works on coherence evaluation can trace back to entity grid model(Barzilay and Lapata 2008; Guinaudeau and Strube 2013), recent works (Farag and Yannakoudakis 2019; Mesgar and Strube 2018; Moon et al. 2019; Nguyen and Joty 2017) have focused on utilizing neural networks for modeling coherence with varied structures such as CNN, LSTM and transformers. These models have achieved noticeable performance when sufficient amount of labeled data is provided. As emphasized above, manually labeling the coherence of essays relies on expert knowledge, requiring significant amount of time and cost. Hence modeling coherence in a low-resource setting can be crucial in many real-world scenarios and applications. However, most previous coherence models assume sufficient labeled data available while the low-resource setting is less explored.

In this paper, we present our approach which pretrains RoBERTa with whole word masking (WWM) on Chinese middle school student essays collected from an external source. WWM is performed in an unsupervised way which adds little cost to the original low-resource setting. In addition, WWM is effective in capturing general characteristics of middle school essays. Subsequently, the pretrained RoBERTa is finetuned on a small set of training data for coherence evalutaion.

Though we pretrain RoBERTa, our method is easy-to-employ and universal across most of the transformer-based language models. Experiment results on Chinese essays written by middle school students provided by NLPCC2023 Shared Task7 demonstrates that this simple strategy can achieve a fair performance. We also illustrate the performance of some variants of our method including pseudo labeling and adding additional neural network on top of RoBERTa, which provides insights into potential methods that are likely to result in performance drop.

The contributions of this work are as follows:

- We collected and curated a substantial amount of middle school student essay data relevant to the task.
- We propose a simple yet effective pretraining method that comes with little additional cost under a low-resource setting.
- We carry out experiments on several methods beyond pretraining to provide future works with evidence on effective approach for coherence evaluation.

2 Related Work

For Coherence Evaluation, there had been several theories that characterize coherence(Asher and Lascarides 2003; Grosz et al. 1995; Mann and Thompson 1988). Inspired by the Centering Theory (Grosz et al. 1995), some early coherence evaluation models (Barzilay and Lapata 2008; Guinaudeau and Strube 2013)

were proposed to distinguish a coherent from incoherent texts with the entity grid model.

Later works have designed neural network architectures for coherence modeling: the Neural Local Coherence Model (Nguyen and Joty 2017) which uses CNN to capture local coherence features in an essay; LSTM variants for modeling potentially longer coherence relationships (Farag and Yannakoudakis 2019; Mesgar and Strube 2018; Moon et al. 2019); multi-task learning which jointly trains Bi-LSTM to score coherence and predict the type of grammatical role (GR) of a dependent with its head. With transformer-based models becoming widespread across various NLP tasks, some recent works utilized transformer-based architectures for coherence evaluation. For instance, Jeon and Strube (Jeon and Strube 2022) proposed an entity-based neural local coherence model which encode an essay with XLNet.

Coherence evaluation can be incorporated into other tasks to boost the performance of the target task. One model for automated essay scoring (AES), for instance, can take coherence evaluation as one of its components for assessing organization score of an essay (He et al. 2022; Song et al. 2020), which greatly improves the effectiveness of essay scoring.

3 Method

When pretraining a language model, general corpora such as Chinese Wikipedia are typically used to capture the linguistic knowledge that is universal across various NLP tasks. However, essays written by middle school students might differ substantially from the corpora on which the language model is pretrained. In general, grammatical and logical errors are frequently found in those essays, which poses a gap between language models and the downstream coherence evaluation task for middle school students' essays.

To this end, we pretrain RoBERTa on middle school student essays with whole word masking (WWM) strategy so that RoBERTa has a better understanding of the general content and structure of the essays. Pretraining with WWM is performed in an unsupervised way, hence it could be easily adopted in our setting where little labeled examples are available. We choose whole word masking as it outperforms individual character masking in various Chinese NLP tasks (Cui et al. 2021).

Whole Word Masking (WWM) primarily changes the training data generation strategy during the pre-training phase. In simple terms, the original tokenization based on Word Piece would split a complete word into several subwords, and during the generation of training samples, these separated subwords would be randomly masked. In WWM, if some of the Word Piece subwords of a complete word are masked, then other parts belonging to the same word will also be masked, which means the whole word is masked.

It's important to note that the term "mask" here refers to different actions, such as replacing with [MASK], keeping the original vocabulary, or randomly replacing with another word. It is not limited to the case where a word is replaced with the [MASK] label.

Subsequently, we finetune the pretrained RoBERTa on the labeled dataset. Specifically, we add a linear classifier head on top of the [CLS] token representation produced by RoBERTa, and finetune the model with the standard cross entropy loss:

$$\mathcal{L}_{class} = \sum_{i=1}^{n} \sum_{c=1}^{|C|} p(y_c^{(i)}|x_c^{(i)}) \log q(y_c^{(i)}|x_c^{(i)}) \quad (1)$$

where $p(y_c^{(i)}|x_c^{(i)})$ and $q(y_c^{(i)}|x_c^{(i)})$ are the true and predicted probability of c-th class of i-th training instance, respectively.

4 Experiments

4.1 Dataset and Evaluation Metric

We carry out experiments on the dataset provided by NLPCC2023 Shared Task7 Track1: Coherence Evaluation[1] This dataset consists of Chinese essays written by middle school students, where the coherence of each essay is evaluated on a three-level scale of excellent, moderate and poor. Within the dataset, 60 essays are train set, while another 5000 essays serve as test set.The data statistics are shown in Table 1.

Table 1. Statistics of Coherence Evaluation dataset

	paragraphs/article	sentences/paragraph	words/paragraph	words/sentence
Median	6.0	2.0	68.0	27.0
90%ile	10.0	5.0	234.0	64.0
Mean	6.79	2.65	88.50	33.38
Maximum	49	42	1010	550

For pretraining dataset, we crawl essay data from website Lele Ketang[2] The Chinese essays are written by middle school students from 7 to 12 grade. We split the dataset during pretraining so that all data is utilized while also ensuring appropriate text length for the language model. This yielded approximately 200,000 essays. The statistics of the length of the above data is provided in Table 2.

The performances of our method and baselines are evaluated on macro precision(P), recall(R), F1-score(F1) and accuracy (acc). We perform 5-fold cross validation on training set, and report the test set performance of models that have best validation accuracies.

[1] https://github.com/cubenlp/NLPCC-2023-Shared-Task7.
[2] http://www.leleketang.com/zuowen/.

Table 2. Statistics of the length of the pretraining data

	Median	75th Percentile	90th Percentile	Average	Maximum	Minimum
Value	329.0	344.0	361.0	292.35	3563	50

4.2 Implementation Details

We implement our model with Pytorch and transformers library. For pretraining, We used the Roberta-chinese-wwm-ext-large as the baseline pre-trained model and trained it for 10 epochs using AdamW optimizer with default parameters. The batch size was set to 16 and the learning rate was set to 2e-5. The training was performed on two NVIDIA RTX 3090 GPUs. Next, we finetune the pretrained RoBERTa with the Adawm optimizer for 20 epochs. The learning rate was set to 1e-5 and the batch size was fixed at 8. All other parameters were set to their default values.

4.3 Baselines

We consider several variants of our method as baselines which we compare our proposed method against:
PFT our proposed method; pretraining on task-related data with WWM followed by finetuning.
PFT+HAN a hierarchical attention pooling network (HAN) on top of RoBERTa pretrained on task-related data; Attention pooling layer with RoBERTa map the essay into a sequence of paragraph representations, and another attention pooling layer maps paragraph representations into an essay representation for final classification. Punctuations in each paragraph are embedded as a single vector that is concatenated to the corresponding paragraph representation.
PFT+HAN+pseudo assign pseudo labels on unlabeled test set using PFT model, and augment original train set with pseudo dataset to train HAN.

Table 3. Performance comparison on the test set. Best results are in bold.

Model	Precision	Recall	F1	Accuracy
PFT	**38.50**	**43.54**	32.54	**43.99**
PFT+HAN	34.91	35.14	**34.83**	35.15
PFT+HAN+pseudo	33.68	32.20	28.46	38.78

5 Results

Experiment results are shown in Table 3. We can observe that even when provided with a small amount of labeled data, combining finetuning with the task-related pretraining is an efficient strategy which can outperform random guess by a large margin, achieving an accuracy of 43.99%. Contrary to our expectation, PFT experiences a big drop in its performance when it is added with an auxiliary hierarchical attention pooling network, reaching an accuracy of 35.15%. Augmenting PFT+HAN further with pseudo labeled test data slightly improves the accuracy (38.78%) yet it also worsens precision, recall and F1 of PFT+HAN. It shows that knowledge learnt from PFT is not good enough to transfer to the test set, since the pseudo-labeled dataset has a harmful effect on the PFT+HAN model. In short, both adding auxiliary network or pseudo-dataset to PFT have failed to make further improvements over a simple PFT.

6 Conclusion

Coherence is a fundamental concept in essay assessment in that it plays a vital role in ensuring clarity, conciseness and fluency of an essay. Due to the prohibitive cost for manually assigning coherence labels to essays, developing coherence models under low-resource settings is of importance in various real-world scenarios. In this paper, we propose an effective approach for Chinese coherence evaluation task. Specifically, we address the challenge of a small amount of labeled data through pretraining RoBERTa with a large amount of task-related data in an unsupervised manner and finetuning the pretrained model on labeled data.

Experiment results on the Chinese essays written by middle school students demonstrate that our simple approach can outperform a random guess by a large margin despite of limited amount of labeled data. In addition to the simplicity, our method is also applicable to transformer-based coherence evaluation models other than RoBERTa. However, both adding auxiliary network or pseudo-dataset to our original method had negative effects on the performance, indicating that more investigations are necessary to carefully design auxiliary network or self-training strategy.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (62076008) and the Key Project of Natural Science Foundation of China (61936012).

Appendix

During the process of improving overall accuracy, we have also experimented with some new model architectures, including the Cross Task Grader model mentioned below.

PFT+HAN

We proposed a multi-layer coherence evaluation model, depicted in Fig. 1, which firstly utilized pre-trained RoBERTa to extract features from the articles, followed by an attention pooling layer. Then, we concatenated punctuation-level embeddings and passed them through another attention pooling layer. Finally, we obtained the ultimate coherence score by using a classifier.

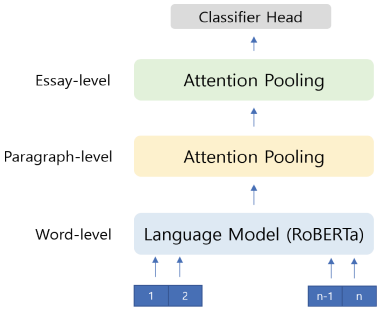

Fig. 1. PFT+HAN

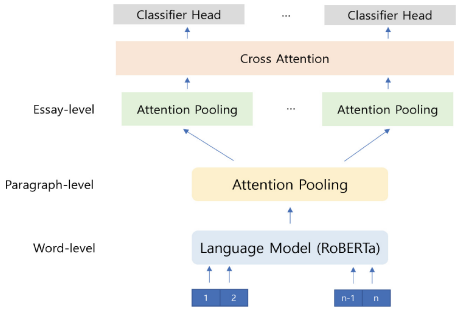

Fig. 2. Cross Task Grader

Pre-trained Encoder. A sequence of words $s_i = \{w_1, w_2, \ldots, w_m\}$is encoded with the pre-trained RoBERTa.

Paragraph Representation Layer. An attention pooling layer applied to the output of the pre-trained encoder layer is designed to capture the paragraph representations and is defined as follows:

$$m_i = tanh(W_m \cdot x_i + b_m) \tag{2}$$

$$u_i = \frac{e^{w_u \cdot m_i}}{\sum_{j=1}^{m} e^{w_u \cdot m_j}} \tag{3}$$

$$p = \sum_{i=1}^{m} u_i \cdot x_i \tag{4}$$

where W_m is a weights matrix, w_u is a weights vector, m_i is the attention vector for the i-th word, u_i is the attention weight for the i-th word, and p is the paragraph representation.

Essay Representation Layer. We incorporated punctuation representations to enhance the model's performance. We encoded the punctuation information for each paragraph, obtaining the punctuation representation pu_i for each paragraph. Then, we concatenated this representation pu_i with the content representation p_i of each paragraph:

$$c_i = concatenate(p_i, pu_i) \tag{5}$$

where c_i represents the representation of the concatenated i-th paragraph. Next, we use another layer of attention pooling to obtain the representation of the entire essay and is defined as follows:

$$a_i = tanh(W_a \cdot c_i + b_a) \tag{6}$$

$$v_i = \frac{e^{w_v \cdot a_i}}{\sum_{j=1}^{a} e^{w_v \cdot a_j}} \tag{7}$$

$$E = \sum_{i=1}^{a} v_i \cdot c_i \tag{8}$$

where W_a is a weights matrix, w_v is a weights vector, a_i is the attention vector for the i-th paragraph, v_i is the attention weight for the i-th paragraph, and E is the essay representation.

Cross Task Grader

We also used Multi-task Learning(MTL) in our experiment, which is depicted in Fig. 2.

We used both target data and some pseudo-labeled essays from various grade and created a separate PFT+HAN model for each. To facilitate multi-task learning, we adopted the Hard Parameter Sharing approach, sharing the pre-trained encoder layer and the first layer of attention pooling among all the models.Additionally, we added a cross attention layer before the classifier.

Cross Attention Layer. After obtaining the essay representation, we added a cross attention layer to learn the connections between different essays, defined as follows:

$$A = [E_1, E_2, \ldots, E_N] \tag{9}$$

$$\alpha_j^i = \frac{e^{score(E_i, A_{i,j})}}{\sum_{i}^{l} e^{score(E_i, A_{i,l})}} \tag{10}$$

$$P_i = \sum \alpha_j^i \cdot A_{i,j} \tag{11}$$

$$y_i = concatenate(E_i, P_i) \tag{12}$$

where A is a concatenation of the representations for each task $[E_1, E_2, \ldots, E_N]$, and α_j^i, is the attention weight. We then calculate attention vector P_i through a summation of the product of each weight α_j^i and $A_{i,j}$. The final representation y_i is a concatenation of E_i and P_i.

References

Asher, N., Lascarides, A.: Logics of Conversation. Cambridge University Press, Cambridge (2003)

Barzilay, R., Lapata, M.: Modeling local coherence: an entity-based approach. Comput. Linguist. **34**(1), 1–34 (2008)

Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z.: Pre-training with whole word masking for Chinese BERT. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3504–3514 (2021)

Farag, Y., Yannakoudakis, H.: Multi-task learning for coherence modeling. arXiv preprint arXiv:1907.02427 (2019)

Grosz, B.J., Joshi, A.K., Weinstein, S.: Centering: a framework for modelling the local coherence of discourse. Comput. Linguist. **21**(2), 203–225 (1995)

Guinaudeau, C., Strube, M.: Graph-based local coherence modeling. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 93–103 (2013)

He, Y., Jiang, F., Chu, X., Li, P.: Automated Chinese essay scoring from multiple traits. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 3007–3016 (2022)

Jeon, S., Strube, M.: Entity-based neural local coherence modeling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 7787–7805 (2022)

Mann, W.C., Thompson, S.A.: Rhetorical structure theory: toward a functional theory of text organization. Text-interdisciplinary J. Study Discourse **8**(3), 243–281 (1988)

Mesgar, M., Strube, M.: A neural local coherence model for text quality assessment. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4328–4339 (2018)

Moon, H.C., Mohiuddin, T., Joty, S., Chi, X.: A unified neural coherence model. arXiv preprint arXiv:1909.00349 (2019)

Nguyen, D.T., Joty, S.: A neural local coherence model. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1320–1330 (2017)

Song, W., Song, Z., Liu, L., Fu, R.: Hierarchical multi-task learning for organization evaluation of argumentative student essays. In: IJCAI, pp. 3875–3881 (2020)