Automated Poetry Scoring Using BERT with Multi-Scale Poetry Representation

Mingzhi Gao[1*] **Selin Ahipasaoglu**[2] **Kristin Schuster**[3]

[1]University of Southampton, Southampton, UK
gmz100@163.com
[*]Corresponding Author
[2]University of Southampton, Southampton, UK
s.d.ahipasaoglu@soton.ac.uk
[3]Writing Through, Cambodia
info@writingthrough.org

Abstract: Automated poetry scoring is an emerging task in automated text scoring, which is receiving increasing attention in AI for education. Poetry is distinct from other text in its complexity and specialty in language feature Moreover, poems are usually rated from multiple criteria besides the overall impression. However, few existing methods to the best of our knowledge have considered a tailored text representation model for encoding poetry. Moreover, the lack of large poetry corpus and extensive labelled data is another major constraint to construct an effective poetry scoring model. To address such problems, we proposed BERT-based models with multi-scale poetry representation. In addition, we employ multiple losses and R-Drop strategy to align the distribution of manual and model scoring and mitigate the tendency of consistent score in poems. Experiment results demonstrate that our model with multi-scale poetry representation stands out when comparing with single-scale representation model.

Keywords: automated poetry scoring, pre-trained language model, multi-scale text representation.

1. Introduction

With the widespread application of online education in the last few years, automated text scoring (ATS) is receiving more and more research attentions, which assesses written works. An ATS model typically consists of two parts: one for text representation and another for scoring. In previous research, various types of text have been involved including essays, arguments, stories, and question answers etc. In contrast to other types of text, the syntactic structure of poetry is more flexible, forming unique linguistic feature. The quality of a poem is determined by different aspects including creativity, coherence, mood, novelty etc. Exploring a specified model to learning informative poetry representation for accurate scoring is necessary and remains unexplored.

In general, ATS has been developed from traditional machine learning methods and pre-training-based methods. In the early stage, automated text scoring systems mainly adopt regression or ranking systems with pre-processing text and human extracted domain-independent features like text length, text complexity, sentiment intensity etc. [7] [8] [9].

After that, more and more recent ATS systems were transformed by deep neural networks of various architectures and achieved comparable results with traditional ATS. For example, Taghipour & Ng[1] proposed a stacked model that involving both convolutional layers and recurrent layers. Dong and Zhang[3] designed a hierarchical CNN model and proved the ability of domain-adaptation of neural networks. Moreover, Dong et al.[4] provided new insights into attention mechanism combined with CNN and LSTM. Nadeem et al.[10] explored, for the first time, the effect of incorporating contextual embeddings and concatenating discourse structure information through new design of pre-training task. Ridney et al.[11] predicted the overall as well as specific traits score in cross-prompt setting. The superior performance of deep neural networks for text representation learning contributes to the development of ATS. However, large scale of annotated text corpus with artificial scores is the prerequisite for constructing a well-performed neural network based ATS system, which does not fit the case of the scarcity of large-scale poetry corpus.

In the past few years, pre-trained language models have been applied to more and more ATS systems as backbone model to obtain more informative text representation. In automated essay scoring, Rodriguez et al.[12] is one of the earliest research discussing and analysing the effect of BERT and XLNet[13] for generating text representation. Song et al.[14] applied adaptive fine-tuning to score essays of target prompt. Chang et al.[15] applied sentence-BERT plus neural network to generate sentence representation for assessing quality through clustering essays. Kumar et. al [16] designed stacked essay representation model in multi-task pattern for scoring essay traits. Wang et al.[17] is one of the most competitive models, which demonstrated that learning from word, section and document scales based on pre-trained BERT could achieve state-of-the-art results in essay scoring. In automated answer scoring, Steimel et al.[18] proposed a BERT-based model and demonstrated considerable adaptation performance for content scoring of answers to free-response questions. Fernandez et al.[19] developed in-context BERT fine-tuning approach to score the answer of reading comprehension adaptable to various items. In addition, Rodriguez et al. [12] and Dong et al.[20] tried RoBERTa[21], and ALBERT[22] for ATS while failed to stand out among original BERT models. With only few research related to variant of BERT in ATS, such findings deserve future verification.

To alleviate the difficulties of this novel text scoring task, we propose to apply pre-training-based BERT model with joint learning of multi-scale poetry representation, which is suitable for the case with limited training data. To our knowledge, no prior work has investigated the effectiveness of pre-trained language model on poetry scoring. Moreover, words or sentences in a poem contribute differently to the score, and certain key words that represent imagery, mood, moving etc. play an important role in determining the level of a poem and are the parts need to be more focused on. Therefore, we incorporate attention pooling operation to calculate word and sentence weights as complements to our pre-trained based model.

The contributions of our research are as follows:

(1) We explore multi-scale poetry representation using BERT-based model and build a model ensembled with neural network and attention mechanism that achieve considerable performance.

(2) We introduce a new similarity loss function and employ R-Drop[6] which is a universal and widely applied new training strategy, to train the model to score poems under the convention of human annotators.

(3) We conduct comparison experiments and demonstrate the advantages of our multi-scale poetry representation learning model, attention mechanism and R-Drop strategy.

2. Methodology

The automated poetry scoring task is defined as follows: Given a poem with n words $X = \{x_i\}, i = 1,2, \ldots, n$, we need to output one score y as the result of measuring the level of this poem. In this section, we illustrate the methodology and model architectures we use for our task.

2.1 Multi-scale Poetry Representation

We explored the poetry representation from word, sentence, and poem scale respectively.

Our task is oriented to scoring poems written in English, while both Chinese and English poems are included for training. The tokenizer and model for encoding poetry we choose is multilingual BERT since it performs well in few-shot and zero-shot cross-lingual transfer scenario.

In the first stage, the input $X = \{x_1, \ldots, x_m\}$, a sequence of words of length $m.$ will be tokenized into a token sequence $T_1 = [t_1, t_2, \ldots\ldots t_n]$, where t_i is the ith token. All input will be padded with $[PAD]$ token or truncated to the specified maximum input size of the BERT model, 510. After that, the tokenized input will be fed into the BERT model to generate the representation of poem, and we use the output of the $[CLS]$ token, a fusion of linguistic information of the whole input, as the poem-scale representation. Besides that, the sequence output, which is a sequence of representations of each word of a poem, also incorporate the contextual information. To aggregate all the word representation, deep neural networks like RNN and LSTM are not the best choice due to the gradients vanishing problem especially in the case of long poems. Instead, we apply the attention pooling operation, introduced by Xu et al.[23] and Luong et al.[24] and first applied in text scoring by Dong et. al[4], on those sequence outputs to obtain the combined word-scale poetry representation.

To obtain sentence scale representation, we divided each input poem into sentences and input one sentence a time into the multilingual BERT for a poem and obtain the $[CLS]$ outputs which represent the embedding of those sentences. We then input those $[CLS]$ outputs in a sequence to a LSTM model, which is commonly used in encapsulating discourse structure, to further explore the connection among the sentences, as LSTM model is designed with superior ability to capture long term memory among deep neural network models. The LSTM layer is followed by attention pooling operation on the output hidden states. In this way, we finally obtained a combined sentence-scale representation.

2.2 Model Architecture

We implemented two models using pre-trained multilingual BERT model on the top of Wang et al.[17], a leading approach in essay scoring. The first model is focused on word and poem-scale representation, where we concatenate the poetry and combined word-scale representation of input poem and put it into task-specific Multi-Layer Perceptron to output a score. Different from the first model, the second one aims at encoding poems in sentence scale and the input are sentences in a poem. The output of $[CLS]$ tokens of those sentences will be further processed by a LSTM layer, followed by an attention pooling layer, which takes in the hidden states output of LSTM and produce the combined sentence-scale poetry representation.

The attention pooling operation for both poetry scoring models we use could be defined as follows:

$$\hat{\alpha}_t = \tanh(Q_a \cdot h_t + b_a) \qquad (3-1)$$

$$\alpha_t = \frac{e^{q_a \cdot \hat{\alpha}_t}}{\sum_j e^{q_a \cdot \hat{\alpha}_j}} \qquad (3-2)$$

$$o = \sum_t \alpha_t \cdot h_t \qquad (3-3)$$

Specifically, for document and token scale representation, h_t is the sequence output of multilingual BERT model and for sentence scale representation, h_t is the hidden states of LSTM layer. o is the result of attention pooling. Q_a, q_a and b_a are the parameter of weight matrixes and bias vector respectively.

Our designed model architecture is illustrated in Figure 1.

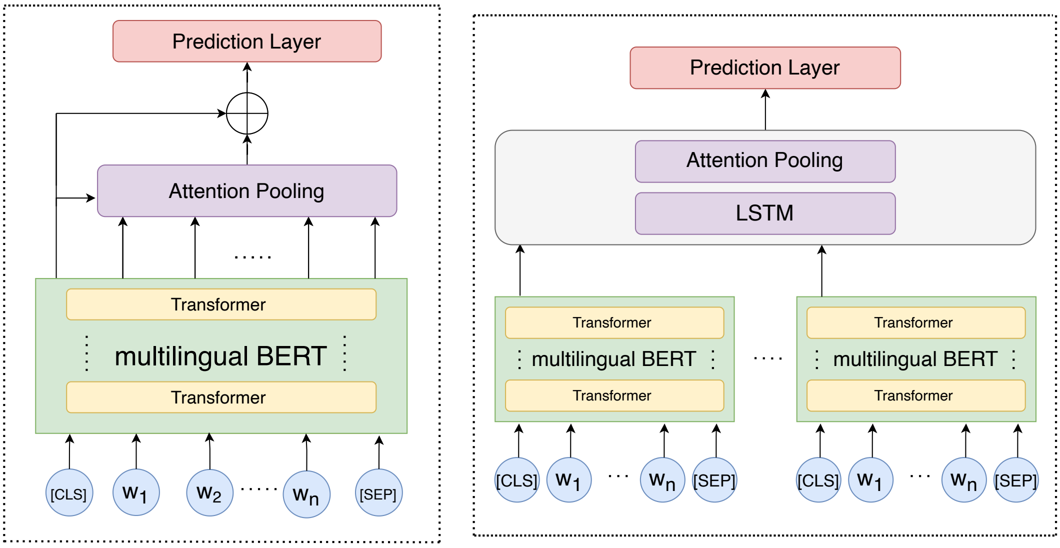

Figure 1: Our proposed automated poetry scoring architecture based on multi-scale poetry representation. The left part illustrates the model for document and token scale poetry representation, and the right part illustrates the model for sentence scale poetry representation.

3. Experiment

3.1 Data Collection and Description

The first part of dataset we use for building the model comes from an anthology of poems written by students from in-person or online workshops held by an NGO last summer, containing poems scored manually by workshop teachers. The second part is the THU Poetry Quality Evaluation DataSet (THU-PQED), developed by THUAIPoet group of THUNLP[25]. It collects 173 Chinese quatrains scored from criteria of fluency, coherence, meaningfulness and provides the overall score. Besides that, we collect a small amount of rated English poems from online resources[26] as supplement for training.

3.2 Model Training

We use the mean square error (MSE) loss, which is the most used measurement for regression tasks, to measure the average value of squared error between target scores and prediction scores among all the poems. Given N poems, we calculate MSE according to the following equation:

$$MSE(y,\hat{y}) = \frac{1}{N}\sum_{i} (y_i - \hat{y}_i)^2 \qquad (4-1)$$

, where y_i and $\hat{y}_i$ are the predicted score and the label for the ith poem respectively. Moreover, the scores assigned by teachers usually follows a normal distribution. Based on the ideas above, we adopt similarity loss to measure the similarity in distributions between the score by teachers and by model. In each training step, we take the predicted scores of a batch of poems as the prediction vector y, and a batch of human-rated scores as the label vector $\hat{y}$. We use cosine similarity as the similarity measure and expect to minimize the similarity loss to enable the model to learn more about the correlation and score distribution among the poems in a batch. The similarity loss is defined as below:

$$\text{similarity}(y,\hat{y}) = 1 - \cos(y,\hat{y}) \qquad (4-2)$$

Then we could obtain the combined loss:

$$loss_{combined}(y,\hat{y}) = MSE(y,\hat{y}) + \text{similarity}(y,\hat{y}) \qquad (4-3)$$

In addition, we also incorporate R-drop strategy, introduced in the research of Wu et al. [6], by minimizing the bidirectional KL-divergence of the output distributions of any pair of sub models sampled from dropout to mitigate the mismatch problem between training and inference model. In our training settings, we added the bidirectional KL-divergence loss to our original combined loss and jointly perform back propagation and parameter update.

In conclusion, the expression of our loss function is defined as below:

$$\mathcal{L}_i = \mathcal{L}_i^{(Loss)} + \alpha\mathcal{L}_i^{(KL)} \qquad (4-4)$$

$$\mathcal{L}_i^{(KL)} = \frac{1}{2}\left[KL\left(P'_\theta(y) \parallel P_\theta(y)\right) + KL\left(P_\theta(y) \parallel P'_\theta(y)\right)\right] \qquad (4-5)$$

$\mathcal{L}_i^{(Loss)}$ is the combined loss of a batch of input calculated by the combined loss function of expression $(4-3)$. $P'_\theta(y)$ and $P_\theta(y)$ are two probability distribution mapped by the "SoftMax" operation on two output batches, and the KL-divergence between two distributions is represented by $KL\left(P'_\theta(y) \parallel P_\theta(y)\right)$ and $KL\left(P_\theta(y) \parallel P'_\theta(y)\right)$. After averaging, the bidirectional KL divergence between these two distributions of the model prediction batch and label score batch is represented by $\mathcal{L}_i^{(KL)}$. The parameter α is tuned within the range of 1 to 10 and we select the best value, 5, according to the training performance.

3.3 Model Comparison

To further study the effectiveness of our model based on multi-scale poetry representations, we conduct the following experiments comparing performances of our baseline models:

(1) Following the classic fine-tuning pattern, we first directly input the representation of $[CLS]$ token obtained from BERT module, which is the poem-scale poetry representation, into the task-specific dense layer for predicting score.

(2) We explored the performance of another cross-lingual pre-trained language model: XLM-Roberta [27] for poem and word-scale representation, as our poetry scoring model used bilingual (Chinese and English) poems for training, although the original BERT-based model already has shown competitive benchmark performance in many NLP tasks. As the multilingual version of Roberta, XLM-Roberta follows the same configuration and architecture of Roberta and inherited the training pattern of XLM. However, the superiority of Roberta-based model mainly demonstrated on large-scale downstream task and may vary among tasks. Whether it could achieve better performance on poetry scoring task still requires experiment.

3.4 Implementation Details

The multilingual BERT part in figure 1-(a) is shared by the document and word scale poetry representations, and the ones in figure 1-(b) are shared among all sentence scale poetry representations. We use the "bert-base-multilingual-cased" as the base PLM for both proposed models, which includes 12 transformer layers with the hidden size being 768. In the overall training process, we freeze all the layers except the LSTM, attention, and the task specific dense layer.

We applied 60/20/20 split for train, validation, and test datasets to the scored dataset after pre-processing and normalizing the overall scores in the dataset to the same range. For fine-tuning the parameters, we use mini batch-based Adam optimizer [28] with the learning rate of $2e-5$, $\beta1 = 0.9$, $\beta2 = 0.999$, and $L2$ weight decay of 0.005 in the end-to-end pattern. The Dropout rate is set to 0.1 according to the settings of BERT.

We use m-BERT-DOC-TOK to represent scoring poems from poetry scale features based on multilingual BERT, m-BERT-SEN to represent scoring poems from sentence scale features, m-BERT-DOC to represent the original pattern of fine-tuning pre-trained based models, which only scores poems from single poem scale. Based on our proposed m-BERT-DOC-TOK and m-BERT-SEN model, we explore the effect of attention mechanism and R-Drop strategy on pre-training models for our task.

4. Result

4.1 Overall Results

We summarize the performance of our proposed models and model comparison results in Table 1.

ID	Model	Combined Loss	RMSE
1	m-BERT-DOC	0.3814	0.6176
2	m-BERT-DOC-TOK	**0.1139**	**0.3374**
	w/o R-Drop	0.2291	0.4786
	w/o attention pooling w max pooling	0.1585	0.3892
	w/o BERT w XLM-Roberta	0.2163	0.4651
3	m-BERT-SEN	0.2902	0.5387
	w/o R-drop	0.4834	0.6953
	w/o attention pooling w max pooling	0.3412	0.6022

Table 1：Experiment results of all models in terms of the combined loss and RMSE as the evaluation metrics. The names of our proposed models are denoted in bold. The values of combined loss and the RMSE in bold are the best performance for our poetry scoring task.

All models included in Table 1 are implemented with the same model settings introduced in 3.4. As the poetry scoring task is a regression task for predicting values, we choose RMSE, a common evaluation metric for assessing the scoring accuracy and the expression is defined as follows:

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{N}\sum_{i}(y_i - \hat{y}_i)^2} \qquad (5-1)$$

, where y_i and $\hat{y}_i$ are the predicted score and the label for the ith poem respectively. Moreover, we calculated the combined loss introduced in 4.2 as a supplement.

It is concluded from Table 1 that our model m-BERT-DOC-TOK obtains the best result of on both evaluation metrics. In terms of the performance of different scales of representation, models encoding poetry from poetry and word scale (model 2) outperform those models from sentence scale (model 3). The model m-BERT-DOC-TOK outperforms not only m-BERT-DOC, which only use the document scale feature, and but also m-BERT-SEN, which encodes poems from sentence scale and applies the LSTM and attention pooling to further process the sentence scale features. Such results demonstrate that BERT based model could benefit from multi-scale poetry representation though only two scales are combined.

The following plot (Figure 2) makes contrast between the performance of our two main models (m-BERT-DOC-TOK and m-BERT-SEN) tested on some poems produced from the workshop mentioned in 3.1. It is intuitive that the distribution of predicted scores by model 2 (our best model) is closer to that of ground truth scores compared with model 3.

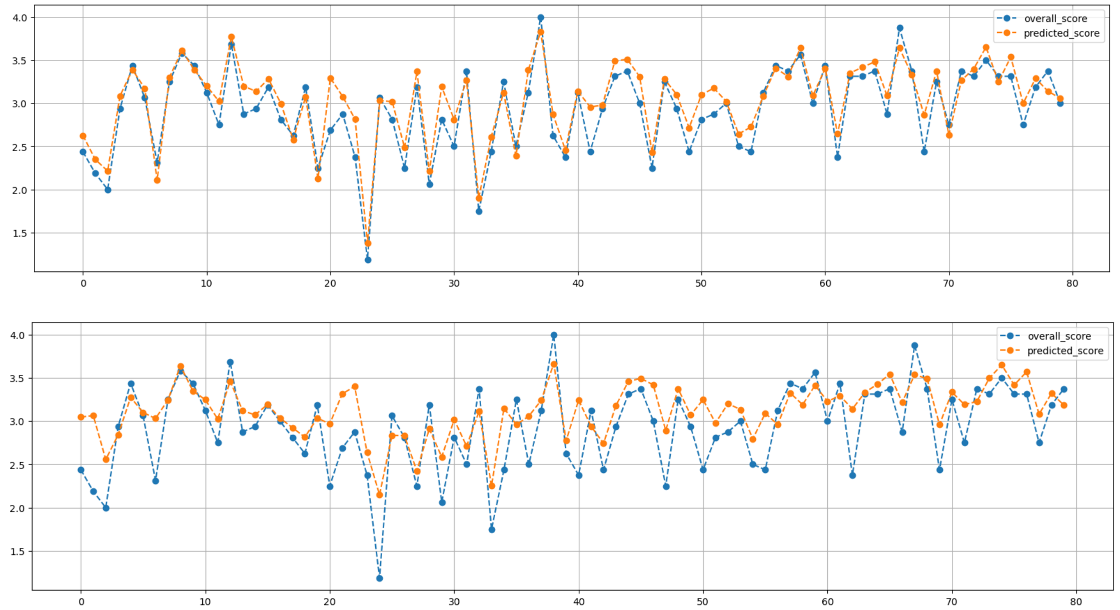

Figure 2：Case study: predicted scores generated by m-BERT-DOC-TOK (top) and m-BERT-SEN (bottom) and ground truth scores on tested poems, which are marked in different colours.

In terms of the performance of encoding poetry from two single scale: poetry and sentence scale, we find that m-BERT-DOC is superior to m-BERT-SEN. This indicates that document scale global features could reflect the level of a poem more accurately than features integrated from separate sentences, though LSTM is equipped with competitive ability to integrate context information. This finding is also a prompt that incorporating global representation from document scale could enhance the prediction performance.

4.2 Effect of R-Drop

We also test the effectiveness of R-Drop strategy. To do this, we drop the bidirectional KL-Divergence loss in the training process (denoted as w/o R-Drop). Due to the limitation of computational resource, we only perform experiment on our two proposed models (model 2 and 3). Result shows that the performance of our models from both multi and sentence scales are improved. In our practical training process, the loss function converged quickly without R-Drop included, but the model overfitting occurred quickly as well with the number of training epochs increases. After including R-Drop, although the convergence of training became slower and the training time of each epoch became longer, it brought about the effect of regularization improvement and mitigate the problem of overfitting, thus the final optimum is much better with superior performance.

4.3 Effect of Attention Mechanism and Pooling Operations

Then we investigate the effect of attention mechanism to the performance of our proposed model. We compare the performance of our models that including attention pooling (model 2 and 3), replacing attention pooling by max pooling (w/o attention pooling w max pooling in model 2 and 3), and no pooling operation (model 1). It is found that the ability to distinguish the contribution of different parts inside a poem to the final score and integrate the linguistic knowledge of poem empowered by attention mechanism could help both the m-BERT-DOC-TOK and m-BERT-SEN model obtain more accurate results than others. Besides, the effect of max pooling is inferior to attention pooling, indicating the importance of the contribution distinguishment and information integration among words.

4.4 Comparison of Multilingual Models

We further explored the effect of poetry representation based on different base LM from document and token scale, with R-Drop strategy and multiple loss when fine-tuning the models. We replace the multilingual BERT part in model 2, the best model we propose, with XLM-Roberta (w/o BERT w XLM-Roberta), while failed to achieve superior performance. According to Conneau et al.[27], XLM-Roberta outperforms multilingual BERT in three classical cross-lingual NLP tasks. A possible explanation for such contradictory results may be that this kind of improvement may mainly owing to more investment in pre-training time and larger training batch size for leverage the learning ability of the model, rather than the optimization of pre-training tasks for consistent improvement, so that its advantage is subject to certain tasks. However, such assumption still requires further experiment on diverse model architecture, more task-specific data and training settings for more evidence.

5. Conclusion

In our research, we investigated the performance of applying BERT-based pre-trained language model with multi-scale poetry representation and adopt an ensembled method by including neural network methods for poetry scoring. As our training data is composed by poems from two languages, we apply the multilingual BERT as backbone model for obtaining multi-scale poetry representation. Moreover, we employed multiple losses and incorporated R-Drop strategy when fine-tuning the model, which lead to competitive performance.

Through experiments, we find that: poetry scoring accuracy could benefit from integrating poetry and word-scale representation with superior prediction accuracy to single scale representation from document or sentence; scores predicted by encoding poetry from document scale could be more consistent with the true level of poetry writing; the attention mechanism and pooling operation plays an important role in integrating important information for accurate prediction; R-Drop is indispensable for leveraging the training performance and learning ability of models. The results of those experiments further support the considerable performance of our proposed model. One finding that requires further experiments is that whether XLM-Roberta could outperform multilingual BERT in poetry scoring with more scored data and thorough training.

One weakness of this study is that it could only assign overall score to poems and lacks the evaluation from sub-criteria. Exploring and optimizing fine-tuning strategy, like intermediate task training and adaptive fine-tuning, is a fruitful area for future research of our poetry scoring task.

Acknowledgement

We thank the University of Southampton for provide funding that was spent for the

NGO organizing poetry writing workshop, collecting and scoring poems.

References

[1] Kaveh Taghipour and Hwee Tou Ng. (2016) A neural approach to automated essay scoring. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891.

[2] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. (2016) Automatic text scoring using neural networks. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* pages 715–725.

[3] Fei Dong and Yue Zhang (2016) Automatic features for essay scoring–an empirical study. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077.

[4] Fei Dong, Yue Zhang, and Jie Yang. (2017) Attention-based recurrent convolutional neural network for automatic essay scoring. *In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

[6] Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., Chen, W., ... & Liu, T. Y. (2021) R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34, 10890-10905.

[7] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. (2011) A new dataset and method for automatically grading esol texts. *In HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technolo gies*, pages 180–189.

[8] Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. (2015) Flexible domain adaptation for automated essay scoring using correlated linear regression. *In Proceedings of EMNLP 2015*, pages 431–439.

[9] Ma˘da˘lina Cozma, Andrei M. Butnaru, and Radu Tudor Ionescu. (2018) Automated essay scoring with string kernels and word embeddings. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

[10] F. Nadeem, H. Nguyen, Y. Liu, and M. Ostendorf, Automated essay scoring with discourse-aware neural models, *in Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications.* Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 484–493.

[11] Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. (2021) Automated cross-prompt scoring of essay traits. *In Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13745– 13753.

[12] Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. (2019) Language models and automated essay scoring. *arXiv: Computation and Language*.

[13] Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., & Le, Q.V. (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding. *NeurIPS.*

[14] Song, W., Zhang, K., Fu, R., Liu, L., Liu, T., & Cheng, M. (2020). Multi-stage pre-training

for automated Chinese essay scoring. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 6723-6733.

[15] Chang, L. H., Rastas, I., Pyysalo, S., & Ginter, F. (2021) Deep learning for sentence clustering in essay grading support. *arXiv preprint arXiv*:2104.11556.

[16] Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Many Hands Make Light Work: Using Essay Traits to Automatically Score Essays. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1485–1495, Seattle, United States. Association for Computational Linguistics.

[17] Wang, Y., Wang, C., Li, R., & Lin, H. (2022) On the Use of BERT for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation. *arXiv preprint arXiv:2205.03835.*

[18] Steimel, K., & Riordan, B. (2020) Towards instance-based content scoring with pre-trained transformer models. *In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence* (Vol. 34).

[19] Fernandez, N., Ghosh, A., Liu, N., Wang, Z., Choffin, B., Baraniuk, R., & Lan, A. (2022) Automated Scoring for Reading Comprehension via In-context BERT Tuning. *arXiv preprint arXiv:2205.09864.*

[20] Dong, L., Li, L., Ma, H., & Liang, Y. (2021). Automated Chinese Essay Scoring using Pre-Trained Language Models. In CS & IT Conference Proceedings (Vol. 11, No. 19). CS & IT Conference Proceedings.

[21] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019) Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

[22] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019) Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942.*

[23] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. (2015) Show, attend and tell: Neural image caption generation with visual attention. *ICML.* volume 14, pages 77–81

[24] Minh-Thang Luong, Hieu Pham, and Christopher D Manning (2015) Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025.*

[25] Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li. (2018) Automatic Poetry Generation with Mutual Reinforcement Learning. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3143–3153, Brussels, Belgium.

[26] Chris Routh (2022) The Den of Amateur Writing. https://www.amateurwriting.net/

[27] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019) Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116.*

[28] Diederik P. Kingma and Jimmy Ba. (2015) Adam: A method for stochastic optimization. In 3rd International Conference for Learning Representations.