Explore Deep Learning for Chinese Essay Automated Scoring

Haojin Li¹ and Tao Dai²*

¹Xi’an Gaoxin No.1 High School, Xi’an 710075, Shaanxi, China
²School of Software Engineering, Xi’an Jiaotong University, Xi’an 710049, Shaanxi, China
Email: zzj.ddt@stu.xjtu.edu.cn

Abstract. Automated Essay Scoring (AES) has gained increasing attention in recent years. In this paper, we propose a novel automated Chinese essay scoring model, called BLA (BERT and Bi-LSTM with Attention), by using neural network. The model uses a BERT network to obtain the sentence vectors for an essay, and then uses a Bi-LSTM network with two layers to extract the essay vector. We also consider topic instruction as a long sentence and obtain its vector representation by BERT. The obtained topic instruction vector is then used as attention information to further help obtaining more effective essay vector. Moreover, we also present a new dataset that contains topic instruction and related essays collected from Chinese high school. The detailed experimental results show that our proposed model outperforms other baseline methods.

1. Introduction

Essay writing is a crucial task in Gaokao which is a college entrance examination in China. In a common scenario, students will finish the essay according to a specific hint or essay topic. It takes massive time and effort for teachers to manually grade all essays. With the help of artificial intelligence technology, Automated Essay Scoring (AES) has become possible to put into practical use in order to essentially alleviate the above problem.

There exists a variety of AES approaches in the literature, and most of them are focus on English essay scoring. Page [1] devised one of the earliest AES system the Project Essay Grade (PEG). The key part of PEG is a linear regression with several text features. Since then, many works follow this approach by adding more effect text features [2-6]. The main defect in the above approaches is they all applied traditional machine learning methods. That means they all highly depend on manually selected features. Recently, many researchers are trying to use deep learning to solve the above problem [7-12]. Nevertheless, when comes to Chinese essay scoring, there are still three problems for the above models. Firstly, there is no publicly available Chinese essay benchmarking dataset for the AES task. Therefore, it is hard to validate the usability of existing model for the Chinese essay scoring task. Secondly, Chinese text needs to be segmented into word before feed into the above models. Unlike the English text, there are no obvious delimiters in Chinese text. Therefore, Chinese NLP task need to segregate characters into words first. However, improper Chinese word segmentation will mislead the meaning of the whole sentence and consequently led to poor performance for Chinese NLP tasks. Finally, many existing AES models neglect the information topic instruction. In Chinese essay writing tasks, the writer usually needs to finish the essay according to specific topic instruction. It is reasonable that an excellent essay should not only written well, but also follow and cover the main
content of the given topic instruction. How to automatically and effectively apply the information of topic instruction into the process of essay scoring is an essential problem for the AES task.

To deal with the above problems, we propose a Chinese essay scoring model by using neural network in this paper. The proposed model takes raw Chinese character as inputs and uses Bidirectional Encoder Representations from Transformers (BERT) [13] to generate sentence representation. Comparing with traditional word embedding, BERT can take individual Chinese characters as inputs, which can essentially avoid the problem of word segmentation. Our model then uses two stacked Bi-LSTM to generate the representation for paragraphs and essays. Moreover, we add an average with attention (AWA) layer for Bi-LSTM, which can effectively utilize the attention from topic instruction to generate more reasonable embedding for paragraphs and essays. We also present a new Chinese essay dataset that contains essays and related topic instructions selected from Chinese high school. To the best of our knowledge, this dataset is the first well-organized dataset for the Chinese essay automated scoring task.

2. Related Work

There are existing a variety of automated essay scoring approaches. Most of them are dealing with English essays. The Project Essay Grade (PEG) [1] is almost the first AES approach. PEG extracts several features from the text, and uses linear regression to predict the score for the essay. Lonsdale and Strong-Krause [3] tried score texts by using link Grammar parser. The score is calculated according to average sentence-level scores obtained from the cost vector of the parse, style-based feature, etc., and used regression to predict the score for the essay. Berggren et al. [14] investigated the performance of using regression or classification for Norwegian essay automated scoring in ASK corpus. They found the AES task is best modelled as regression.

The above works are mainly based on traditional regression models. The main defect is that they all heavily rely on manually defined features. Therefore, many researchers apply neural network model to automatically learn features. Tay et al. [10] assumed that the coherence score of an essay can improve the holistic scoring, and proposed SkipFlow model. Their model is based on LSTM and is able to extract neural coherence features that represent relationships between multiple positions of text. Nadeem et al. [11] showed that use a pre-trained neural model can improve the performance of AES.

3. Exploring Deep Learning for Automated Chinese Essay scoring

3.1. Model Overview

In this subsection, we give a brief overview of our proposed BLA model. Given an essay \( e = \{w'_1, w'_2, ..., w'_n\} \) and a topic instruction \( t = \{w'_1, w'_2, ..., w'_n\} \), BLA model is able to assign a proper score for \( e \) by considering not only the content information, but also the relevance between \( e \) and \( t \). As can be seen in figure 1, there are mainly four steps to accomplish the scoring task. These steps are essentially related to four neural network layers. In the following subsections, we will detailly describe these layers.

3.2. Sentence Embedding Layer

The goal of the sentence embedding layer is to obtain the vector embeddings of sentences. We use a BERT network to achieve the above goal. BERT allows us to send two sentences as input. We insert a [CLS] token and a [SEP] token before and after a sentence, respectively. In vanilla BERT, these two tokens are used to aggregate features from one sentence. The \( i \)-th sentence will be embedded into \( v_{i} \). Although BERT is able to generate vectors for all input characters, we only choose the vector of [CLS] to represent the corresponding sentence.
3.3. **Paragraph Embedding Layer**

The architecture of this layer is shown in figure 2. The input $V_p \in \mathbb{R}^{d_v \times N_p}$ is the total sentence representation in paragraph $p_i$. We then use Bi-LSTM to embed these vectors into $Z_p \in \mathbb{R}^{d_z \times N_p}$. For each sentence vector $z^i_p$ in $Z_p$, we consider the attention from topic instruction $v_i$ as semi-supervised information, and generate a weight $\alpha \in \mathbb{R}^{d_z \times N_p}$ using average with the attention layer as follows.

$$
\begin{align*}
    u^i_p &= \text{ReLU}(W_v[z^i_p; v_i]) \\
    \alpha &= \text{softmax}(w_u U_p)
\end{align*}
$$

where $W_v \in \mathbb{R}^{d_z \times (d_v + d_i)}$ and $w_u \in \mathbb{R}^{d_z}$ are parameters of neural network. $[\cdot]$ is concatenation function. $U_p \in \mathbb{R}^{d_z \times N_p}$ denotes all hidden vectors $u^i_p$ in average with the attention layer. ReLU(·) represents the rectifier activation function.

$$
\begin{align*}
    V_p &\xrightarrow{\text{LSTM}} Z_p \\
    Z_p &\xrightarrow{\text{LSTM}} h_p
\end{align*}
$$

Figure 2. The paragraph embedding in BLA.

After getting $\alpha$, we then obtain the paragraph embedding $h_p$ for paragraph $p_i$ by

$$
    h_p = Z_p \alpha
$$

Figure 1. The architecture of BLA model.
It should be noted that the average with attention is to use $\alpha$ to choose which sentences are more correlated with given topic instruction.

3.4. Essay Embedding Layer

Similar with the paragraph embedding layer, we also use Bi-LSTM and AWA layer to obtain the final essay vector. For essay $e_i$, we use Bi-LSTM to embed all paragraph vectors into $Z_e \in \mathbb{R}^{d \times N_e}$. Then we consider the attention from topic instruction $v_i$ as a decision process once again, and generate the final essay vector $h_e$ as the paragraph embedding.

3.5. Generating Final Score and Training

Though essay embedding layer, we obtain the final essay vector $h_e$. We then generate the final score for essay $e_i$ by:

$$\hat{y}_e = \sigma(w_e h_e + b_e) \quad (4)$$

where $\sigma(\cdot)$ represents sigmoid function. $w_e \in \mathbb{R}^{d}$ and $b_e \in \mathbb{R}$ are mapping parameters.

Given the essay set $E$ and each ground truth score $y_i$ for essay $e_i$, we optimize our BLA model by minimizing cross entropy loss as:

$$L = -\sum_{i \in |E|} y_i \log(\hat{y}_i) \quad (5)$$

4. Experiments and Evaluations

4.1. Dataset and Evaluation Metrics

There is no public available Chinese essay dataset. Therefore, we manually constructed a new dataset for the Chinese essay automated scoring task. The dataset has been uploaded in “https://github.com/declan-haojin/AES-Dataset” and can be downloaded publicly. The dataset contains 300 essays and related topic instructions collected from Chinese high school students. To evaluate the performance of our model, we use 5-fold cross-validation on the dataset that described in Section 3 for evaluation. We divide the dataset randomly into 5 folds. One fold is used as the testing data, while the remaining four folds are collected as training data. Following common practice, we employ the following four evaluation metrics: Quadratic weighted Kappa (QWK), Normalized root-mean-squared error (nRMSE), Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SCC).

4.2. Comparison with Other Approaches

In order to validate the scoring performance of our model, we choose several compared methods, including: linear regression (LR), support vector regression (SVR) and Bi-LSTM.

The experimental results of the compared methods are shown in table 1. It can be seen that our BLA outperforms all other methods. For example, comparing with Bi-LSTM which is also the second-best model, BLA promotes the performance by 9.4%, 11.7%, 12.2% and 14.6% relative to QWK, nRMSE, PCC and SCC in our dataset, respectively. It also can be seen that all neural network-based methods (Bi-LSTM and BLA) achieve much better results than traditional machine learning methods (LR and SVR). This indicates that the neural network can extract more reliable features than manually defined features. Due to only utilized several manually defined features, LR generates the worst results. SVR achieves better results than LR. The reason is it uses the RBF kernel to achieve non-linear regression. By applying deep neural network, Bi-LSTM can generate useful vectors for both topic instruction and essay, and consequently generate more accurate scores.
Table 1. Performance comparison between other methods.

| Method | QWK      | nRMSE    | PCC      | SCC      |
|--------|----------|----------|----------|----------|
| LR     | 0.4692   | 0.2145   | 0.4964   | 0.5184   |
| SVR    | 0.5147   | 0.1499   | 0.5427   | 0.5623   |
| Bi-LSTM| 0.5523   | 0.1286   | 0.5717   | 0.5932   |
| BLA    | 0.6042   | 0.1136   | 0.6417   | 0.6795   |

4.3. Parameter Tuning

There are two essential hyperparameters $d_s$ and $d_i$ for BLA. $d_s$ is the dimension of the output of Bi-LSTM. $d_i$ is the vector dimension in AWA layer. We only demonstrate the tuning results on QWK, for other metrics have near results according to our experiments. In order to better demonstrate the influence of these two parameters, we also show the results of Bi-LSTM in figure 3.

![Figure 3](image-url)

(a) (b)

**Figure 3.** The performance impact of $d_s$ and $d_i$.

We first empirically fix $d_s = 350$ to evaluate the effects of varying $d_i$. Figure 3a demonstrates performance variance of BLA when changing $d_i$. It can be seen that the performance ascends as the value of $d_i$ increases. When $d_i > 300$, the performance achieves steady. It also should be noted that when $d_i$ is about less than 110, our BLA generates worse performance than Bi-LSTM. The results indicate that the dimension of $d_s$ should be set carefully.

Next, we analyze the impact of $d_i$ by fixing $d_s = 300$. As can be seen in figure 3b, the performance ascends as the value of $d_i$ increases. When $d_i = 200$, the performance of our model achieves steady. In order to balance the performance and computation speed, we choose 300 and 200 as the optimal values for $d_s$ and $d_i$.

5. Conclusion

In this paper, we propose a novel deep learning model, named BLA, for automated essay scoring task. We also present a new dataset containing topic instruction and related essays collected from Chinese high school. To the best of our knowledge, this dataset is the first dataset for Chinese high school essay evaluation. Our BLA model combines BERT and Bi-LSTM to learn the representation of candidate essays and takes attention information from topic instruction to further improve the essay vector. In the future, we will further improve the model by adding more information, such as the
historical score and background of students. We will also continuously increase the number of essays in our dataset.

References
[1] Page E B 1968 The use of the computer in analyzing student essays International Review of Education 14 (2) 210-225.
[2] Landauer T K, Laham D and Foltz P W 2003 Automated scoring and annotation of essays with the Intelligent Essay Assessor Automated Essay Scoring: A Cross-Disciplinary Perspective pp 87-112.
[3] Lonsdale D and Strong-Krause D 2003 Automated rating of ESL essays Proceedings of the HLT-NAACL Workshop: Building Educational Applications Using Natural Language Processing pp 61-67.
[4] Rudner L M and Liang T 2002 Automated essay scoring using Bayes’ theorem The Journal of Technology, Learning and Assessment 1 (2) 3-21.
[5] Attali Y and Burstein J 2006 Automated essay scoring with e-Rater v.2.0 Journal of Technology, Learning, and Assessment 4 (3) 1-30.
[6] Yannakoudakis H, Briscoe T and Medlock B 2011 A new dataset and method for automatically grading ESOL texts The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies pp 180-189.
[7] Taghipour K and Ng H T 2016 A neural approach to automated essay scoring Proc. of EMNLP pp 1882-1891.
[8] Dong F and Zhang Y 2016 Automatic features for essay scoring-An empirical study Proc. of EMNLP pp 1072-1077.
[9] Jin C, He B, Hui K and Sun L 2018 TDNN: A two-stage deep neural network for prompt independent automated essay scoring Proc. of ACL pp 1088-1097.
[10] Tay Y, Phan M C, Tuan L A and Hui S C 2018 SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring Proc. of AAAI pp 5948-5955.
[11] Farah N, Huy N, Yang L and Mari O 2019 Automated essay scoring with discourse-aware neural models Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications pp 484-493.
[12] Liu J, Xu Y and Zhao L 2019 Automated essay scoring based on two-stage learning arXiv e-prints.
[13] Devlin J, Chang M W, Lee K and Toutanova K 2018 BERT: Pre-training of deep bidirectional transformers for language understanding arXiv e-prints.
[14] Stig B, Taraka R and Lilja Ø 2019 Regression or classification automated essay scoring for norwegian Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications pp 92-102.