AUTOMATED CHINESE ESSAY SCORING USING PRE-TRAINED LANGUAGE MODELS

Lulu Dong[1, 2], Lin Li[1, 2], HongChao Ma[3] and YeLing Liang[1, 2]

[1]The State Key Laboratory of Tibetan Intelligent Information
Processing and Application, Qinghai, Xi Ning, China
[2]Department of Computer Science, Qinghai Normal University, Xi Ning, China
[3]Beijing Language and Culture University, Beijing, China

ABSTRACT

Automated Essay Scoring (AES) aims to assign a proper score to an essay written by a given prompt, which is a significant application of Natural Language Processing (NLP) in the education area. In this work, we focus on solving the Chinese AES problem by Pre-trained Language Models (PLMs) including state-of-the-art PLMs BERT and ERNIE. A Chinese essay dataset has been built up in this work, by which we conduct extensive AES experiments. Our PLMs-based AES models acquire 68.70% in Quadratic Weighted Kappa (QWK), which out-perform classic feature-based linear regression AES model. The results show that our methods effectively alleviate the dependence on manual features and improve the portability of AES models. Furthermore, we acquire well-performed AES models with a limited scale of the dataset, which solves the lack of datasets in Chinese AES.

KEYWORDS

Chinese Automated Essay Scoring, Neural Network, Pre-trained Language Model, Quadratic Weighted Kappa.

1. INTRODUCTION

Writing is a measure of language learners meta-cognitive and linguistic abilities, thus Chinese writing draws increasing attention from learners. With the boom in learning Chinese all over the world, Chinese essay scoring becomes a challenge for both Chinese teaching and testing. It is not only because scoring essays is a time and labor-consuming task, but also different human raters have divergence on the same essays. AES is an effective and efficient substitution for human raters by assigning a holistic score to an essay. AES is a reasonable approach to alleviate the conflict between the increasing number of Chinese essays and the lack of human raters is helpful to reduce subjectivity in human scoring [1]. Classic feature-based AES systems have succeed for English AES like PEG [2], IEA [3], E-rater [4, 5], and BETSY [6]. The performance of a classic AES system is largely determined by its feature set, however, building a high-quality set is a time-consuming and laborious task. Furthermore, it is also a challenge even for experts to take all key scoring aspects into consideration. To reduce the dependency of AES systems on manually building feature sets, Neural Networks such as convolution and recursive neural networks have been introduced into the AES task [7]. The Neural Network-based AES approach avoids complex feature engineering but it is a corpus-greedy method. That is, a large scale of essay corpus is the prerequisite for acquiring a well-performed AES system.

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 01-12, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111901

Language models succeed in many Natural Language processes communities because of their strong capability in language representation. The state-of-the-art PLMs [8] own powerful architecture and are trained on huge scale corpus by different pre-trained tasks. PLMs are successfully applied to complete AES tasks by researchers such as XLNET and BERT. Compared with feature-based and Neural Network-based AES methods, PLMs-based AES systems show a better agreement with human raters trained by the same scale of training corpora [9]. Currently, challenges for Chinese AES can be attributed to the lack of a powerful model and available large-scale corpus. To alleviate the difficulties, we propose to apply fine-tuned PLMs into the Chinese AES task in this work.

The rest of this paper is organized as follows. Section 2 provides an overview of related work in the literature. Section 3 gives a clear definition for the AES task and a detailed explanation for the evaluation metric used in this work. We provide the details of our approach in Section 4, present and analyze results in Section 5. Finally, we draw our conclusion in Section 6.

2. RELATED WORK

Studies in English AES. The performance of sequence models like Long Short-Term Memory (LSTM) exceeds previous feature-based methods on AES tasks by its powerful capability of feature engineering and complex pattern encoding. Alikaniotis et al [10] applied Bidirectional LSTM (BiLSTM), based on which Fei and Yue [11] added convolution layer and pre-trained word embedding into their work. Attention mechanism has been used to AES models to alleviate gradient vanishing and long-distance dependency problem of sequence models. Fei et al [12] adopted attention into the AES task, by which the model can capture significant context and word-level information. With the success of the pre-trained language model in many NLP tasks, Rodriguez et al [13] proposed an AES model based on BERT [14] and XLNET [15]. In general, Neural Networks achieve better agreement with human raters in the English AES task.

Studies in Chinese AES. A few AES systems have been proposed for different Chinese tests like HSK (Hanyu Shuiping Kaoshi) [16], MHK (Chinese Proficiency Test for Minorities in China) [17], and high school essays test [18]. These systems assign a holistic score with linear regression models based on various linguistic features. Kakkonen et al [18,19] proposed a linear regression model whose features are represented by Latent Semantic Analysis (LSA). Previous AES work in Chinese shows a medium correlation between predicted scores and manual scores, which can not fully satisfy the practical application. Powerful modeling approaches have been used to improve the performance of Chinese AES like Supported Vector Machine (SVM) and Back Propagation (BP) neural network. For instance, Ma et al [20] made a comparison study in Chinese AES between SVM and BP neural network. And their results show that BP neural network achieves a higher correlation with the human raters than SVM. Fu [21] et al proposed a hybrid AES model that combined Recurrent Neural Network (RNN) with BiLSTM.

Pre-trained Language Models. PLMs have made great achievements in a variety of NLP tasks, for instance, BERT(Bidirectional Encoder Representation from Transformers) performs better than native human speakers in some GLUE benchmark tasks. PLMs attract increasing interest from organizations and researchers, thus many PLMs are proposed like BERT [14], ERNIE [22], and GTP [8]. BERT is one of the most successful PLMs, which is built on Transformer architecture and adopts a bidirectional mechanism. RoBERTa [23], a variant of BERT, is trained by more training data and training batches than BERT. Ernie is trained by different pre-trained tasks and training strategies that are helpful to add word and phrase-level information into the model. In this work, we focus on three PLMs based on Transformer architecture, thus we use BERT, ERNIE, and RoBERTa to explore the automatic score of Chinese essays.

3. AUTOMATED ESSAY SCORING FOR CHINESE

In this section, we provide a clear definition of the AES task, introduce the dataset we employed, and elucidate the evaluation metric used for assessing the performance of our AES models proposed in this work.

3.1. Task Description

An essay can be evaluated from different aspects including relevance to prompt, the correctness of grammar, and usage of lexical. In this work, we focus on holistic essay scoring, that is, assign an essay an overall score according to its general quality. Thus, the AES model proposed in this work takes a raw essay as input and provides a holistic score as output. We formalize Chinese AES as a classification task, which adopts a 12-class classification. Each class represents a range of 5 points assigned by human raters from 40 to 95.

Our classification models aim to deduce accurate score labels by learning implicit and explicit features of essays. What shall be noted is that scores assigned by human raters are treated as the golden standard, thus our models concentrate on maximizing the agreement between predicted scores and human raters.

3.2. Chinese Essay Corpora

Large-scale Chinese essay corpora with corresponding scores are essential for Chinese AES study. The corpora used in this work contains 7,141 Chinese essays, all of which were written by Chinese as a second language learner. Statistics of our corpora are shown in Table 1, which shows that the size of our dataset is only a third of ASAP dataset. We used 60%, 20%, and 20% essays of each prompt as training, validation, and test dataset, respectively. The corresponding scores are treated as labels adopted in this work, which is assigned by experts in a fair and strict scoring process. Specifically, each essay is graded by two independent human raters and a third senior rater will be introduced when the first two raters cannot reach an agreement in terms of the score.

Table 1. Statistic of our Chinese Essays Dataset

Prompt	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sample	861	703	825	198	325	739	1330	518	687	955
Rating range	40-95	40-95	40-95	65-95	40-95	40-95	40-95	40-95	50-95	40-95

3.3. Evaluation metrics

AES systems used to be evaluated by a variety of statistical measures like Pearson correlation coefficient and Kappa [24]. Since the ASAP competition adopts Quadratic Weighted Kappa (QWK) as an evaluation metric, QWK becomes a widely used approach for AES task by several work [24]. We also take QWK as our evaluation method not only because it is a popular method to that ensure our result is comparable with other AES work, but also because its principle is more reasonable for evaluating AES models than other metrics.

QWK is capable of capturing disagreement between two raters by fusion matrix. The range of QWK is from 0 to 1, and K = 1 if the two raters achieve complete agreement with each other. The brief procedure of computing QWK of two groups of essay scores is defined as follows. Firstly, the weight matrix W is built up as defined by formula 1, where i and j are the scores given by a human rater and an AES system respectively, N is the number of essay grades.

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2} \qquad (1)$$

Then, we construct the matrix Q and the prediction matrix E, where $Q_{i,j}$ refers to the number of times an essay is graded as i by human raters and as j by the automatic score method. E is the outer product of the manual score vector and the AES score vector. Finally, the QWK score is computed as the formula 2:

$$K = 1 - \frac{\sum_{i,j} W_{i,j} Q_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \qquad (2)$$

4. AES MODELS FOR CHINESE

Due to the size limitations of our dataset, only can we use a limited scale of samples to training a model. Thus we employ the strong ability of PLMs in encoding linguistic information and learning about the complex relationship between essays and corresponding scores and achieve a better performance in the AES task. Firstly we build up a regression-based AES model as our baseline model, which is an explainable and stable benchmark for this work. And several PLMs have been proposed with the success of PLMs in many NLP areas like Machine Translation, etc. Then we explore how to apply three powerful Chinese PLMs including BERT, ERNIE, and RoBERTa into our task.

4.1. Baseline Model

The regression-based AES method shows relatively stable and simple performance in both English and Chinese [25]. In the ASAP competition, A regression-based AES model won third place in the ASAP competition [26], following which we build up our benchmark for this work with multi-level latent linguistic features. We comprehensively consider features that contribute to essay scoring, by which we acquired an integrated feature set for Chinese AES. In this section, we will introduce the feature set we built up for Chinese AES, and describe the construction and implementation of the linear regression model.

Features Set A regression-based AES model needs a high-quality feature set. However, feature selection depends on manual selection task is full of challenges due to it is hard for people to comprehensively consider all key information for the AES task. So we construct a latent semantic feature set consisting of three aspects of linguistic features including characters, words, and sentences as shown in appendix Table A1.

Character-level features According to the difficulty, Chinese characters are divided into four grades by HSK [27]. Intuitively, the level of character usage is an effective measure to evaluate writing skills. That is, an essay consisting of more various and high-level characters will be rated a higher score. Based on this assumption, we proposed 11 character features.

Word-level features The words choosing and applying largely determine the level of an essay, so we extend the feature set with eight different features about word usage like the number of tokens, misused words, etc. Through these features, the baseline model is capable of evaluating a writer's level of Chinese basic vocabulary, complex words, and phrases.

Sentence-level features We also consider sentence-level features except for character and word features, that including the ratio of the number of clauses to the total number of sentences, the average sentence length, and the total count of sentence errors. The sentence features ensure the baseline model can make a reasonable prediction of a writer's ability in complex sentence patterns and grammar knowledge.

Regression-based Model A linear regression model is a statistical approach which can make reasonable prediction by learning linear relationship between independent variables and dependent variables [28]. Multiple linear regression was constructed as the baseline model according to the function $Y = aX + b,$ where X refers to the multidimensional features as input, and Y is the score predicted by the linear regression model. The linear regression model Y = aX+b can be realized by the full connection layer. Through a full connection layer, the input X and output Y are connected and the parameters a and offset term b are allocated.

The mean square error (MSE) is taken as the loss function, the Root Mean Square Prop (RMSProp) is the optimization algorithm, and the extracted features are used as the linear regression model of dataset training. In the construction of the linear regression model, the normalization technology is used to normalize the extracted feature data. The data in different ranges are in the same distribution range, which can make the optimization algorithm have a faster convergence speed.

4.2. Our PLMs-based AES Models

Neural Networks require large-scale training data for learning complex patterns between essays and corresponding scores. To avoid the over-fitting problem [7], we explore how to effectively apply PLMs into the Chinese AES.

BERT is the abbreviation for Bidirectional Encoder Representations from Transformers [14], which is one of the most successful PLMs and outperforms human participants in many tasks of GLUE. BERT’s success can be attributed to its effective language modeling approach and the bidirectional multi-layer transformer architecture. As a deep bidirectional transformer model, which generates a feature vector for each element (like a word) of the input sequence with consideration of its preceding and succeeding context [29]. In this paper, we use the BERT-base Chinese to our task, which includes an embedding layer, 12 encoder layers, and a pooling layer. The parameters are up to 110M.

BERT consists of the Encoder of Transformer model and the Transformer learns the representation of linguistic units in context by self-attention and full connection layer. When encoding a linguistic unit of an input sequence, the self-attention mechanism of BERT determines assigning how much attention to each unit. These three vectors are the result of multiplying an embedding vector by a matrix W, which is randomly initialized. The self-attention is defined as for formula 3.

$$\text{Self-}attention = Soft\max(\frac{QK^T}{\sqrt{d_k}})V \qquad (3)$$

Self-attention calculates three new vectors, which are called a query, key, and value (Q, K, and V) respectively, namely $Q = W^{Q} X^{T}$, $K = W^{K} X^{T}$, $V = W^{V} X^{T}$. Calculate the score of self-attention, which determines how much attention we pay to the rest of the input sentence when we encode a word in a certain position. The calculation method of this fractional value is to do point multiplication between query and key and divide the result of point multiplication by a constant $\sqrt{d_k}$. Then a softmax function is used to get the weight, and the result is the correlation of each word to the word in the current position. Multiply the value from value and softmax, and the result is the value of self-attachment in the current node.

BERT is trained by two different word level and sentence level tasks. One task is MLM (Masked Language Modeling) which learns language distribution by recovering several randomly masked words in a sentence. The other is NSP (next sense prediction) that main purpose is to predict the sequence of two sentences. Except for employing knowledge learning in the pre-training stage, BERT shall be properly fine-tuned for a specific downstream task. In this paper, we fine-tune the parameters of middle layers to adjust the embedding of the input sequence and parameters of the prediction layer to improve BERT's performance in our task (more details about fine-tuning are in Section 5).

RoBERTa Researchers of Facebook and the University of Washington carefully investigate the effects of hyper-parameters and scale of training corpus on the performance of BERT. And the result shows that the training of BERT is insufficient, thus they propose RoBERTa [23] improve BERT from the following aspects: (1) extending the scale of training corpus into 160G; (2) increasing the number of parameters into eight thousand; (3) extending training processing by using 500 thousand training epoch.

RoBERTa proposes a full-sentences mechanism, which refers to the length of the input sequence that has been extended from two sentences to a fixed-length context (i.e. paragraphs or articles). And the static MLM of BERT has been replaced with the dynamic MLM, that is, the inputs are masked just before used to the training. By these strategies, RoBERTa outperforms in many NLP tasks. In this paper we use RoBERTa-Base-Chinese to complete AES.

ERNIE employs the architecture of BERT, whose effectiveness has been proved by many studies. The primary idea of BERT is that a powerful representation of language can be effectively learned by simple pre-training tasks and a huge scale of the corpus. ERNIE proposes that the language representation can be further improved by more informative pre-training tasks. Thus, ERNIE adopts three different level mask units (words, phrases, and name entities) in MLM to acquire more semantic information [22].

Based on this assumption, ERNIE uses DLM (Dialogue Language Model) to model query response dialogue structure, takes dialogue pair as input, introduces dialogue embedding to identify the role of dialogue, and uses dialogue response loss to learn the implicit relationship of dialogue, to further improve its semantic representation ability. ERNIE can potentially learn knowledge dependency and longer semantic dependency by unifying the mask to make the model more generalized. Experiments show that ERNIE achieves good results in some Chinese NLP tasks, which is related to its use of forum data for dialogue modeling, which the training corpus for ERNIE is multi-source like encyclopedia article and dialogue, it ensures ERNIE learn various language information distributed in different genre of the corpus.

4.3. Fine-tuning Strategy

When applying PLMs to downstream tasks, many difficulties need to be solved like catastrophic forgetting, which makes the model quickly forget what it learned before. In the work, we use three basic versions of the pre-training language model and try to use a variety of fine-tuning strategies to obtain better experimental results. In the process of fine-tuning, we mainly focus on the influence of sequence length, learning rate, and batch-size on the experimental results. In addition, the final optimization strategy and operation configuration parameters are shown in Table 2.

Table 2. Parameter configuration

Parameter	Value	Parameter	Value	Parameter	Value
optimizer	AdamWeightDecayStrategy	lr-scheduler	Linear decay	learning-rate	2e-5
Weight-decay	0.01	Max-seq-len	510	batch-size	32
warmup-proportion	0.1				

5. Experiments and Results

5.1. Experiments

In this work, we run intensive experiments to investigate the way how to solve the Chinese AES problems by different PLMs including BERT, ERNIE, and RoBERTa. To compare and analyze the performance of different PLMs, we utilized the same dataset in all experiments. Google Colab platform was used to execute our baseline model, and the Baidu PaddlePaddle platform was employed to training and testing our PLMs models.

5.2. Results and Analysis

We tentatively make use of CNN and RNN in Chinese AES, and our results show worse performance than our baseline system. We think the bad performance of CNN and RNN is because they cannot learn about the complex relationship between essays and scores with a limited size of corpora. Our results suggest that it is not practical to training a Neural Network-based AES model from scratch.

Table 3. QWK Results (%)

Model	1	2	3	4	5	6	7	8	9	10	Avg
WSP-T-FT[30]	86.30	58.60	49.50	56.70	-	-	-	-	-	-	62.8
Baseline	57.83	52.40	48.25	62.27	55.32	61.88	56.39	59.19	60.92	59.58	57.40
BERT	77.15	74.83	68.43	73.08	62.24	75.21	56.09	64.51	50.97	72.10	67.46
ERNIE	76.00	79.96	69.62	65.10	62.24	76.36	58.92	61.72	48.68	76.54	67.51
RoBERTa	80.31	80.17	71.29	66.37	55.62	75.38	60.28	63.92	55.96	77.70	68.70

QWK results are shown in Table 3 and we also evaluate our models by Pearson correlation (as shown in Table 4) to compare with previous work. QWK is widely adopted for evaluating AES, while the Pearson coefficient could reflect ranking consistency. The WSP-T-Finetune model was Song [30] adapts multi-stage Pre-training strategy cooperate on the attentional recurrent convolutional neural network with the essay written by a Chinese student. The result shows that the pre-training-based approach is effective for AES. The average QWK of our baseline model is

57.40% and all the three PLMs-based AES models outperform our baseline system in terms of QWK. RoBERTa achieves the best QWK comparing with BERT and ERNIE, whose improvement of QWK reaches 11.30%. The different PLMs-based AES models show very similar performance in the AES task. Our results suggest that fine-tuned PLMs-based AES model is a practical way for Chinese AES with the limitation of the scale of the corpus. In addition, our PLMs-based models also obviously performed better than the work of [30] in the Chinese AES task.

Table 4. Experiment result Pearson correlation (%)

Model	1	2	3	4	5	6	7	8	9	10	Avg
WSP-T-FT[30]	87.70	62.90	53.40	60.60	-	-	-	-	-	-	66.20
Baseline	62.37	58.07	55.54	60.55	57.93	66.64	56.09	61.37	63.80	63.84	60.62
BERT	82.07	76.30	70.02	74.83	62.48	78.53	59.04	71.91	56.12	75.74	70.70
ERNIE	81.40	80.57	70.56	66.04	63.01	79.39	60.68	67.23	50.22	79.61	69.87
RoBERTa	83.40	81.05	72.71	66.37	56.55	77.14	61.26	70.91	57.64	79.84	70.69

A post-hoc analysis has been done for investigating the different performances of our models on different prompts. As shown in Figure 1, RoBERTa outperforms other models on most prompts (8 of 10) except for prompts 4 and 5. We think the best performance of RoBERTa in this work can be attributed to its long contextual training strategy, that is, RoBERTa learns a better language representation by using long context information. RoBERTa only falls behind other models on 2 prompts, we think this is caused by the very small sample size (Table 1). This result reveals the key effect of training set size on RoBERTa, that is, RoBERTa performs better than other PLMs models with the proper amount of training samples in the AES task. This assumption also can be proved by the obvious QWK improvement of RoBERTa on prompts 1, 3, 7, and 10 with a larger number of samples.

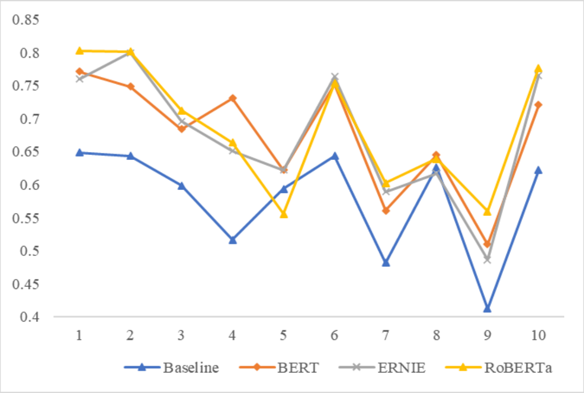

Figure 1. Results QWK under different prompts

6. Conclusion

In this paper, we build up a strong baseline system for the Chinese AES task by a linear regression model. Furthermore, we investigate how to apply PLMs into our AES task including BERT, RoBERTa, and ERNIE. By running intensive experiments on Chinese AES, we find that PLMs-based significantly outperform our baseline system. The results show that RoBERTa achieves 68.70% in QWK, which is 11.30% higher than the baseline. The designing and performance of Chinese AES models are still limited to the size of our corpus, thus larger Chinese essay corpora will be helpful for further study in Chinese AES.

7. FUTURE WORK

In the future we will further analyze the internal structure of PLMs, a more reasonable fine-tuning strategy is adopted to further improve the effectiveness of the automatic scoring model. And we are interested in probe what features or traits are captured by PLMs for scoring. Furthermore, we will explore and make public the larger pre-training Chinese dataset with supervised labels or self-supervised learning strategies.

ACKNOWLEDGMENTS

The authors of this paper received support under Grant CTI2020B05 from Hankao International Education Technology (Beijing) Co., Ltd, which is gratefully acknowledged.

REFERENCES

[1] Balfour S P. Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™[J]. Research & Practice in Assessment, 2013, 8: 40-48.

[2] Page E B. Grading essays by computer: Progress report[C]//Proceedings of the invitational Conference on Testing Problems. 1967.

[3] Foltz P W, Laham D, Landauer T K. The intelligent essay assessor: Applications to educational technology[J]. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1999, 1(2): 939-944.

[4] Attali Y, Burstein J. Automated essay scoring with e-rater® V. 2[J]. The Journal of Technology, Learning and Assessment, 2006, 4(3).

[5] Burstein J. The E-rater® scoring engine: Automated essay scoring with natural language processing[J]. 2003.

[6] Rudner L M, Liang T. Automated essay scoring using Bayes5 theorem[J]. The Journal of Technology, Learning and Assessment, 2002, 1(2).

[7] Taghipour K, Ng H T. A neural approach to automated essay scoring[C]//Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 1882-1891.

[8] Qiu X, Sun T, Xu Y, et al. Pre-trained models for natural language processing: A survey[J]. Science China Technological Sciences, 2020: 1-26.

[9] Mayfield E, Black A W. Should You Fine-Tune BERT for Automated Essay Scoring?[C]//Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. 2020: 151-162.

[10] Alikaniotis D, Yannakoudakis H, Rei M. Automatic text scoring using neural networks[J]. arXiv preprint arXiv:1606.04289, 2016.

[11] Dong F, Zhang Y. Automatic features for essay scoring-an empirical study[C]//Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 1072-1077.

[12] Dong F, Zhang Y, Yang J. Attention-based recurrent convolutional neural network for automatic essay scoring[C]//Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). 2017: 153-162.

[13] Rodriguez P U, Jafari A, Ormerod C M. Language models and automated essay scoring[J]. arXiv preprint arXiv:1909.09482, 2019.

[14] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[15] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. arXiv preprint arXiv:1906.08237, 2019.

[16] Liang Maocheng and Wen Quifang. Review and Enlightenment of foreign automatic scoring system for composition[J].Media in Foreign Language Instruction, 2007, 5: 18-24.

[17] Li Yanan. Automated Essay Scoring For Testing Chinese As A Second Language[D]. Beijing Language and Culture University.

[18] Cao Y, Yang C. Automated Chinese essay scoring with latent semantic analysis[J]. Examinations Research, 2007, 3(1): 63-71.

[19] Kakkonen T, Myller N, Timonen J, et al. Automatic essay grading with probabilistic latent semantic analysis[C]//Proceedings of the second workshop on Building Educational Applications Using NLP. 2005: 29-36.

[20] Ma Hongchao and Guo Li and Peng Hengli. Comparison of Automatic Scoring Effect of Writing Based on SVM and BP Neural Network[J]. Examination research, 2019, (5).

[21] Fu R, Wang D, Wang S, et al. Elegart sentence recognition for automated eassay scoring[J]. J. Chin. Inf. Process, 2018, 32(6): 88-97.

[22] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.

[23] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[24] Yannakoudakis H, Cummins R. Evaluating the performance of automated text scoring systems[C]//Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications. 2015: 213-223.

[25] Dikli S. An overview of automated scoring of essays[J]. The Journal of Technology, Learning and Assessment, 2006, 5(1).

[26] Thyagarajan A, Bhomick P K. Regression based Automated Essay Scoring[J]. http://saisrivatsa.com/Files/aes.pdf

[27] Office of China National Committee for Chinese Proficiency Test. The Grammar Section in A Grade Syllabus for HSK[M]. Higher Education Press, 1996.

[28] Phandi P, Chai K M A, Ng H T. Flexible domain adaptation for automated essay scoring using correlated linear regression[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 431-439.

[29] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.

[30] Song W, Zhang K, Fu R, et al. Multi-Stage Pre-training for Automated Chinese Essay Scoring[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 6723-6733.

APPENDIX

Table A1. Feature Set used in Regression AES Model

Level	Features
character	Total character count Square root of total character count Forth root of total character count Count of unique characters The ratio of number of unique characters to the total number of words in the essay Count of Grade A characters Count of Grade B characters Count of Grade C characters Count of Grade D characters Count of characters above Grade A to D Count of misuse characters
word	Total word count Unique word count Count of Grade A words Count of Grade B words Count of Grade C words Count of Grade D words Count of misuse words Average count of words in sentences
sentence	Total sentence count Total clause count The ratio of number of clauses to the total number of sentences The average sentence length The total count of sentence errors

AUTHORS

Lulu Dong (1996-)
Postgraduate student
HuiNing, Gansu, China;
College of Computer of Qinghai Normal University;
Research field: Natural Language Processing

Lin Li (1980-),
Associate Professor
College of Computer of Qinghai Normal University;
Research field: Natural Language Processing; Natural Language Generation

HongChao Ma (1979-)
PhD
HuaiNan, Anhui, China;
College of Intensive Studies
Beijing Language and Culture
University;
Research field: Language
Testing and Teaching
Evaluation.

YeLing Liang (1997-)
Postgraduate student
PingYao, Shanxi, China;
College of Computer of Qinghai Normal University;
Research field: Natural Language Processing