Understanding Lexical Features for Chinese Essay Grading

Yifei Guan[1,2], Yi Xie[1,2], Xiaoyue Liu[1], Yuqing Sun[1,3(✉)], and Bin Gong[1,3(✉)]

[1] School of Software, Ministry of Education, Shandong University, Jinan, China
poppy@mail.sdu.edu.cn, heilongjiangxieyi@163.com, suiqiyue@163.com, {sun_yuqing,gb}@sdu.edu.cn
[2] School of Computer Science and Technology, Ministry of Education, Shandong University, Jinan, China
[3] Engineering Research Center of Digital Media Technology, Ministry of Education, Shandong University, Jinan, China

Abstract. Essay grading is an important and difficult task in natural language processing. Most of the existing works focus on grading non-native English essays, such as essays in TOEFL. However, these works are not applicable for Chinese essays due to word segmentation and different syntax features. Considering lexical features are important for essay grading, in this paper, we study the expert evaluation standard and propose an interpretable lexical grading method for essays. We first study different levels of vocabulary provided by experts and introduce a quantitative evaluation framework on lexical features. Based on these standards, we quantify the Chinese essay dataset of 12 education grades in primary and middle schools and propose a set of interpretable features. Then a Bi-LSTM network model is proposed for semantically grading essay, which accepts a sequence of word vectors as input and integrates attention mechanism in terms of lexical richness. We evaluate our method on real datasets and the experimental results show that it outperforms other methods on the task of lexically Chinese essay grading. Besides, our method gives interpretable results, which are helpful for practical applications.

Keywords: Essay grading · LSTM · Lexical richness · Interpretable

1 Introduction

It is an important and difficult task to automatically grade essays in natural language processing. Most existing works focus on non-native English essay grading. For example, E-Rater [1], a rating system developed by ETS, has been applied in major official examinations, such as TOFEL and GMAT since 2001, with an accuracy rate of over 97%. The *juku* [2] is a website that provides services on automatic correction of English essay, on which students can submit their essays and get feedback on corrections. However, the English essay grading methods cannot be applied to Chinese tasks due to the differences between the two languages, such as lexical separator and

Y. Sun et al. (Eds.): ChineseCSCW 2019, CCIS 1042, pp. 645–657, 2019.
https://doi.org/10.1007/978-981-15-1377-0_50

tense. To the best of our knowledge, there is not any publicly available work on Chinese essay grading.

Lexical richness is an important indicator to evaluate a student's linguistic level, which reflects his vocabulary and the ability to use the words. Therefore, it is reasonable to select the lexical richness as features to grade the essays. This paper has the following contribution:

(1) We propose a lexical grading framework that integrates expert evaluation. By studying the different levels of vocabulary, idioms and advanced verbs provided by experts, we analyze the lexical features on the Chinese essays of 12 education grades in primary and middle schools and introduce interpretable metrics on the lexical richness of essay.
(2) We propose the Bi-LSTM network [3, 4] with attention mechanism method to extract the semantic features of essay. The model combines two layers of Bi-LSTMs to generate the semantic vector of an essay that considers both the sentence and text aspects.
(3) We adopt the multilayer perceptron network for essay grading with the attention mechanism on the lexical aspect. Based on the lexical features extracted by the experts, the grading results are interpretable, which are much helpful for practical applications.
(4) The method is verified against real datasets and the experimental results show that it outperforms other methods on the task of lexically grading Chinese essay.

The rest of this paper is organized as follows. Section 2 presents the related work. In Sect. 3, we introduce expert review rules on essays and the data sets, and discuss how to extract the lexical features. In Sect. 4, we present the grading model on Chinese essays. Section 5 evaluates our model on real datasets. We conclude the paper in Sect. 6.

2 Related Work

In this section, we present the influential approaches on essay grading. Existing essay grading models include two categories: traditional machine learning and deep learning.

Classical regression and classification algorithms often use the features extracted by experts in automatic essay grading tasks. Project Essay Grade (PEG) [5, 6] is one of the earliest essay grading systems, using linear regression over vectors of lexical features to predict an essay level. PEG relies on the analysis of the latent semantic features of the essays without understanding the semantic content of the essays, such that it cannot give feedback to students. Intelligent Essay Assessor (IEA) [7] adopts Latent Semantic Analysis (LSA) [8] to calculate the semantic similarity between essays without considering the language expression. The E-rater system [1], developed by the Educational Testing Service, has been deployed in the English language test, such as Test of English as a Foreign Language (TOEFL) and Graduate Record Examination (GRE). The system uses a number of different features, including different aspects of vocabulary and grammar. BETSY [9] is a program, funded by the United States Department of Education, which is based on the probability theory and the statistics on a training

corpus to classify texts. In 2012, the Hewlett Foundation sponsored a competition on Kaggle[1] called the Automated Student Assessment Prize (ASAP) [10], aiming to find efficient automated essay grading methods. The dataset released has been widely used for automatic essay grading tasks [11, 12].

In recent years, motivated by the success of deep learning in different domains, many deep neural networks have been proposed for essay grading. Cozma et al. [13] proposed a method combining word vector and SVM with the string kernel function. Alikaniotis et al. [11] employed an LSTM model to learn features for the essay grading task, which learns score-specific word embeddings (SSWEs) for word representation. Taghipour et al. [14] combined LSTM and CNN for automatic essay grading, which outperforms many methods that require handcrafted features. Dong et al. [12] introduced the attention mechanism on the basis of CNN and RNN, and found that the attention mechanism on keywords and sentences helps to judge the quality of essays. Jin et al. [15] proposed a two-stage neural network model to automatically grade prompt-independent essays, and built three stacked Bi-LSTMs to extract the semantic, part-of-speech and syntactic features of essays. Based on LSTM, a new SKIPFLOW mechanism was proposed by Tay et al. [16], which incorporated semantic and logical information of essays.

However, the English essay grading methods cannot be directly applied to Chinese tasks due to the differences between the two languages, such as lexical separator and tense. Although, Fu et al. [17] analyze the gracefulness of sentences in Chinese essays by the combination of CNN and LSTM, but their model cannot grade a complete essay. To the best of our knowledge, there is not any publicly available work on Chinese essay grading. Moreover, the automatic grading tasks require the interpretable results, especially the deep neural network model. To this end, we propose an interpretable Chinese essay grading model, which gives a reasonable explanation for essay grading.

3 Understanding Expert Rules for Essay Lexical Features

In this section, we first introduce a set of expert essay grading standards. Then we present the experimental dataset and define the lexical features of essays. Based on the essay grading rules, we introduce a quantitative evaluation framework on lexical features.

3.1 Expert Review Rules

Since essay grading is the somewhat subjective task, to have the normalized rules on Chinese essay grading, the Ministry of Education asks experts to set up the *Essay Scoring Standard for National New Curriculum Standards College Entrance Examination* (*the standard* for short). This standard evaluates the essays into four levels according to the expressions and the characteristics, the details are given in Table 1. We can see that the lexical features play an important role in essay grading, such as the lexical richness and the usage of advanced words. It is reasonable to use lexical features for Chinese essay grading.

[1] http://www.kaggle.com/c/asap-aes/.

We adopt the *Outline of Chinese Proficiency Vocabulary and Chinese Characters* [18] (*the outline* for short) to extract the measurable lexical features from essays. The outline was officially released by the *Examination Center of the Office of the National HSK Examination Committee* to grade Chinese words. The Chinese vocabulary are graded into four levels, from advanced to simple: A-level, B-level, C-level and D-level. Specifically, A-level and B-level always contain advanced words, such as "哀悼". The common used words are classified to C-level or D-level, such as "帮助" and "发生". In this paper, we adopt the word levels in the outline as rules to generate the lexical features of Chinese essays.

Table 1. Some rules on essay grading in the *Essay Scoring Standard for National New Curriculum Standards College Entrance Examination*.

	First level	Second level	Third level	Fourth level
Expression	Precise content structure Quite fluency verbs	Complete content structure Fluency verbs	Almost complete content structure Fairly fluency verbs	Confusing content structure Not fluency verbs
Characteristic	Quite rich content Quite literary writing	Rich content Literary writing	Fairly rich content Fairly literary writing	Not rich content Not literary writing

3.2 Dataset

We adopt a Chinese essay dataset from primary and middle schools that are provided by our partner. It contains 59,142 student essays covering from Primary Grade Two (P2 for short) to Senior Grade Three (S3 for short). Table 2 shows the statistics of the dataset on each education grade, including the number of essays, the average essay length and the average number of idioms. We count the number of advanced verbs according to an *Advanced Chinese Verb List* (*the list* for short), which includes 199 advanced verbs such as "觊觎" and "斟酌". The average numbers of advanced verbs are listed on the second row from the bottom in Table 2.

Table 2. Statistics on the dataset.

	Primary school					Junior high school			Senior high school		
Education grade	P2	P3	P4	P5	P6	J1	J2	J3	S1	S2	S3
#essay	4867	11636	13194	12028	10566	2045	1969	1492	494	469	382
Avg. #character	178	252	323	354	379	408	425	578	789	904	867
Avg. #idiom	0.96	2.15	2.96	3.88	4.34	4.71	5.10	5.62	7.08	7.54	8.03
Avg. #advanced verb	0.010	0.019	0.053	0.071	0.092	0.13	0.19	0.23	0.31	0.34	0.41
Essay grade	2	3	4	5	6	7	8	9	10	11	12

Since the education grades reflect the average ability of writing skills of students, we adopt the education grades as the essay grade in the learning process. The higher the education grade that the essay is selected from, the higher the corresponding essay grade.

3.3 Interpretable Lexical Features of Essays

In this section, we propose the interpretable features on lexical richness of essay by understating the statistics on the essay dataset from Chinese primary and middle schools with the help of the word levels extracted from the outline.

Vocabulary is one of the basic elements of essays. An essay is more likely to have a higher grade if it contains many high-level words. To understand the correlations between the usage of words and the essay grade, we calculate the number of words in different word levels against the grades of students. As shown in Fig. 1, there are obviously positive correlations between the student grade and the number of high-level words used in each essay. Therefore, it is reasonable to adopt the metric of lexical richness as an indicator in essay grading.

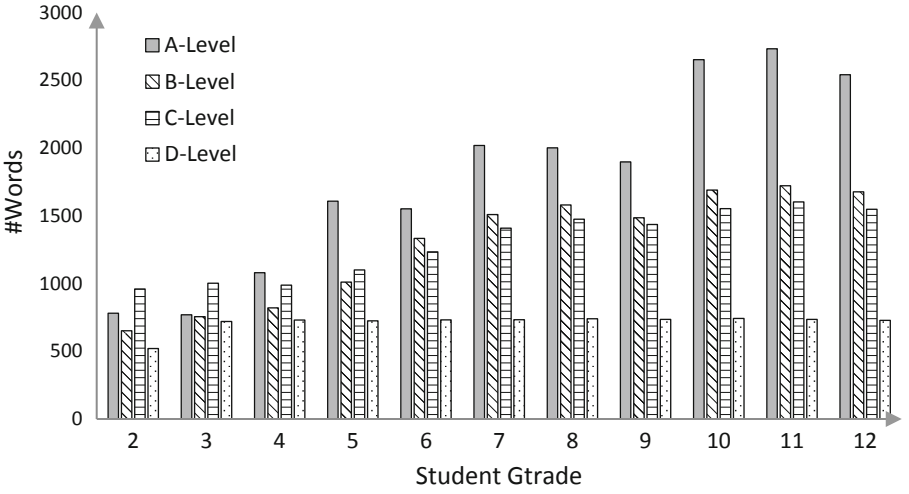

Fig. 1. The correlation between the student grades and the lexical richness.

We also consider other measurable lexical features to represent an essay, including the length of essay, the number of idioms and the number of advanced verbs used in an essay. We also quantify the importance on how much a word contributes for the judgement of the grade of essay by information gain and select 44 words with high information gains against student grades. These words are adopted as the lexical features as well. The interpretable lexical features are summarized in Table 3 for grading Chinese essay. This lexical feature vector for each essay is denoted by E_f, which would be used in the following grading process.

Table 3. Interpretable lexical features of an essay.

1	#A-level word
2	#B-level word
3	#C-level word
4	#advanced verb
5	#character
6	#idiom
7	#high information gain word

4 Chinese Essay Grading Based on Lexical Features

In this section, we discuss how to grade Chinese essays based on the lexical features and the content of essay. There are three parts in our model, as illustrated in a left-right view in Fig. 2.

The left part of data processing is the extraction of lexical features and mapping a document to a sequence of word vectors. The original content of an essay is processed by two modules, the extraction of interpretable lexical features as presented in Sect. 3, denoted by E_f, and to generate the semantic vectors of a document by pre-trained word vectors. The middle part is the essay representation module to encode the essay content as vectors, denoted by E_e, by two Bi-LSTM networks with attention mechanisms. Then these two vectors E_f and E_e are concatenated together as the input of the right part. A multilayer perceptron network is adopted to predict the grades of essays. The details are presented in the following subsections.

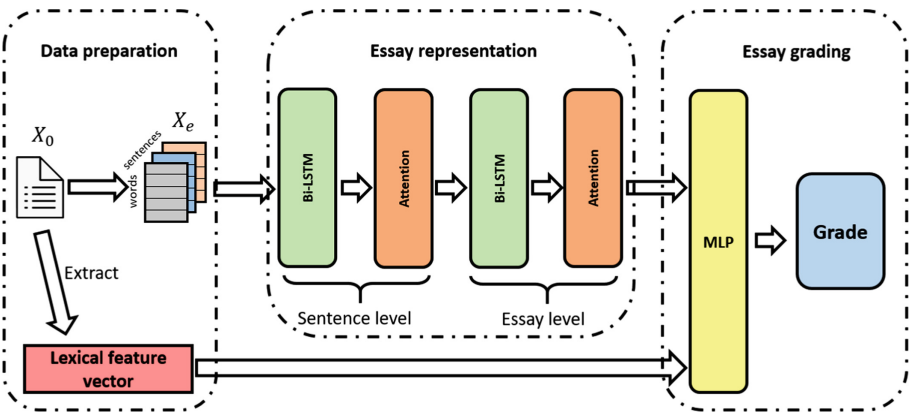

Fig. 2. The interpretable essay grading model.

4.1 Learning the Semantic Representation of Chinese Essay

Given a Chinese essay, the semantic representation is learned by a deep network, denoted by E_e. Let the sequence of sentences $s_1, s_2, \cdots\cdots, s_L$ denotes the contents of an essay, where L is the length of essay, and each sentence s_i contains a sequence of words, represented by $w_1^i, w_2^i, \cdots\cdots, w_{T_i}^i$, where T_i is the length of sentence. The word w_t^i represents the t-th word in the i-th sentence, and is embedded to a word vector x_t^i by *Word2vec* [19] or *Glove* [20]. Then the sequence of word vectors is fed to a Bi-LSTM network, which contains a forward LSTM network reading the sentence s_i from w_1^i to $w_{T_i}^i$, and a backward LSTM network reading the words from $w_{T_i}^i$ to w_1^i:

$$\vec{h}_t^i = \overrightarrow{LSTM}\left(x_t^i\right),\ t \in [1,\ T_i] \tag{1}$$

$$\overleftarrow{h}_t^i = \overleftarrow{LSTM}\left(x_t^i\right),\ t \in [T_i,\ 1] \tag{2}$$

$$h_t^i = \vec{h}_t^i \oplus \overleftarrow{h}_t^i \tag{3}$$

where $\vec{h}_t^i$ and $\overleftarrow{h}_t^i$ represent the hidden states of t-th cell in the forward LSTM and the backward LSTM, respectively. The symbol $\oplus$ denotes the vector concatenation.

To have the semantic representation of a sentence, we adopt the attention mechanism to learn the different contribution α_t^i of word w_t^i in sentence s_i. The h_t^i is fed to a one-layer perception network to extract the hidden state u_t^i. Then the normalized weight α_t^i is learned through a *softmax* function. The context vector u_w is introduced as the combination weights on the outputs of the network, which are randomly initialized and jointly learned during the training process. Finally, the sentence vector s_i is the sum of h_t^i against weights α_t^i:

$$u_t^i = \tanh\left(W_w h_t^i + b_w\right) \tag{4}$$

$$\alpha_t^i = \frac{exp\left(u_t^{i^T} u_w\right)}{\sum_j exp\left(u_j^{i^T} u_w\right)} \tag{5}$$

$$s_i = \sum\nolimits_t \alpha_t^i h_t^i \tag{6}$$

Similarly, we further learn the essay representation vector by a Bi-LSTM based on the sentence vectors. The attention mechanism here is used to analyze the importance of each sentence in an essay. The context vector u_s is randomly initialized and jointly learned during the training process. The semantic representation of essay E_e is learned by the following functions:

$$\vec{h}_i = \overrightarrow{LSTM}(s_i),\ i \in [1,\ L_i] \tag{7}$$

$$\overleftarrow{h}_i = \overleftarrow{LSTM}(s_i),\ i \in [L_i,\ 1] \tag{8}$$

$$h_i = \vec{h}_i \oplus \overleftarrow{h}_i \tag{9}$$

$$u_i = \tanh(W_s h_i + b_s) \tag{10}$$

$$\alpha_i = \frac{exp\left(u_i^{\mathrm{T}} u_s\right)}{\sum_j exp\left(u_j^{\mathrm{T}} u_s\right)} \tag{11}$$

$$E_e = \sum\nolimits_t \alpha_i h_i \tag{12}$$

4.2 Chinese Essay Grading

Considering the lexical features and the contents are both important elements of Chinese essays, we concatenate the lexical feature vector E_f and the semantic representation E_e together, and feed it into a multilayer perceptron. The sigmoid function is adopted as the activation function to predict the grade $\hat{y}$ of essay:

$$\hat{y} = sigmoid\left\{W_c\left[E_e \oplus E_f\right] + b_c\right\} \tag{13}$$

The MSE is adopted as the loss function to measure the variance between the predicted grade and the ground-truth y:

$$\mathcal{L} = mse\,(y,\ \hat{y}) = \frac{1}{n}\sum\nolimits_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{14}$$

5 Experiments

5.1 Experimental Setup

Each essay is segmented into sentences and each sentence is segmented into words. We adopt the 300-dimensional embeddings provided by Beijing Language and Culture University [21] who preform *Word2vec* [19] on the 22.6 G corpus from Wikipedia and other Chinese corpus. Then, we use the word embeddings to initialize the embedding matrix W_e.

In our experiments, the maximum number of words per sentence is limited to 100, and the maximum number of sentences per document to 50. Padding is used to maintain the length of word sequences and sentence. We fix the LSTM hidden state size at 64, and the dimension of both sentence and essay representations obtained by

Bi-LSTM are then 128. The context vectors in the attention layer also have a dimension of 128.

For training, the batch size is 16. We use the ADAM [22] optimizer with learning rate $= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ as parameters. We use 80% of the data for training and 20% for testing.

5.2 Evaluation Metrics

The Quadratic Weighted Kappa (QWK), the Pearson Correlation Coefficient (PCC) and the Spearman Correlation Coefficient (SCC) are adopted as the evaluation metrics in this paper, which are widely applied to measure essay grading models.

The Kappa coefficient is an evaluation metric used for consistency testing or measuring classification accuracy. In this paper, the Kappa is used to measure the consistency between the predicted essay grade and the ground-truth. QWK is improved from Kappa by adding quadratic weights. QWK is calculated as follow:

$$\kappa = 1 - \frac{\sum W_{i,j} O_{i,j}}{\sum W_{i,j} E_{i,j}} \tag{15}$$

$$W_{i,j} = \frac{(i-j)^2}{(R-1)^2} \tag{16}$$

Where $W_{i,j}$ denotes the square weight matrix. j is the predicted essay grade based on our model and i is the ground truth, formally $\hat{y} = j$, $y = i$. R represents the number of essay grades, R = 11. The element $O_{i,j}$ in the observation matrix O denotes the number of essays that satisfy $\hat{y} = j \cap y = i$. The expectation matrix E is calculated from the outer product of the true histogram vector and the predicted histogram vector, and is normalized.

5.3 Comparison Methods

- **SVM.** Support-vector machines are supervised learning models that analyze data used for classification and regression analysis with the lexical features. We use this method as a baseline in the comparison method.
- **2L-LSTM-word2vec** [11]. A two-layer Bi-LSTM model is used to generate a representation vector of the essays, and then the vector is used to obtain the essay grade.
- **CNN-LSTM** [14]. The essay vector is generated by CNN and LSTM, and then the vector is used to obtain the essay grade.
- **CNN-LSTM-ATT** [12]. A CNN layer is employed to encode word sequences into sentences, followed by an LSTM layer to generate the essay representation. An attention mechanism is added to model the influence of each sentence on the final essay representation.
- **TDNN** [15]. This model employs three two-layer Bi-LSTMs to extract the features of the essays in terms of semantics, part-of-speech and syntax, and finally grade the

essays. Since the syntactic tree extracted by this method is not suitable for Chinese, we use the semantic and part-of-speech features in the experiment only.

- **2L-Bi-LSTM-ATT-lexical.** This is our proposed model, using word vectors and lexical features as input. We next compare three variances of our model.
- **2L-Bi-LSTM-ATT.** This model only uses the word vector as the input, which is similar to 2L-Bi-LSTM-ATT-lexical but without using the lexical feature.
- **2L-Bi-GRU-ATT-lexical.** This model replaces the LSTM unit with GRU, using the word vector and combining the interpretable features as input.
- **2L-Bi-GRU-ATT.** Similarly, this model replaces the LSTM unit with GRU and only uses the word vector as the input.

5.4 Results and Analyzes

In this section, different components of our model are compared and analyzed using three correlation metrics. The performance results of each variance on different evaluation metrics is shown in Fig. 3. Then, we compare our model with other state-of-art methods, where the best result for each metric is highlighted in bold in Table 4.

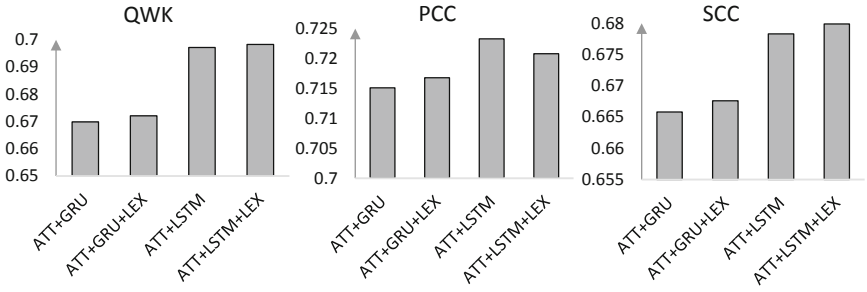

Fig. 3. Variances comparison on different metrics.

By comparing the variances of the methods proposed in this paper, we can see that 2L-Bi-LSTM-ATT-lexical performs better than 2L-Bi-LSTM-ATT in QWK and SCC, indicating that the lexical features are helpful to improve the performance of the model on the essay grading task. Meanwhile, by comparing our model with the method of grading the essays directly using SVM, we find that the performance of using only the lexical features on the essay grading task is not very satisfactory. This indicates that, apart from the lexical features, semantic representations of essays are also essential in essay grading task.

In our experiments, our method with GRU is not as effective as the LSTM method in consistency with ground truth, but the GRU takes less time in training. In the first few epochs, the convergence rate of the GRU method is fast, while in the next few epochs, the convergence rate is slowed down. Since LSTM outperforms GRU, we choose 2L-Bi-LSTM-ATT-lexical instead of 2L-Bi-GRU-ATT-lexical.

Table 4. The QWK, PCC and SCC scores of different models.

Method	QWK	PCC	SCC
2L-Bi-LSTM-ATT-lexical	**0.6977**	**0.7208**	**0.6789**
SVM	0.443	0.506	0.471
2L-LSTM-word2vec	0.6395	0.6615	0.6356
CNN-LSTM	0.6793	0.6803	0.6492
CNN-LSTM-ATT	0.6659	0.6924	0.6658
TDNN	0.6952	0.7191	0.6781

The experimental results show that our model performs better than other methods on QWK, PCC and SCC. The performance results of each model on different evaluation metrics is shown in Table 4. In terms of QWK, 2L-Bi-LSTM-ATT-lexical performs the best among different comparison models. More precisely, 2L-Bi-LSTM-ATT-lexical outperforms 2L-LSTM-word2vec by 10%, demonstrating that the proposed model has a higher consistency with the real essay grading. However, in terms of PCC, 2L-Bi-LSTM-ATT-lexical performs worse than the model without lexical features, but still performs better than other methods. Our model is obviously superior to other comparison models in PCC score except TDNN. Similar to PCC, our model has the best performance in terms of SCC, demonstrating that the proposed model monotonically correlates better with the real essay grading.

At the same time, TDNN has the best performance in comparison models, which is close to our proposed model. However, this model is more complicated and less interpretable and it does not incorporate expert knowledge. Due to the interpretable features extracted by experts, our model is easier to understand and has higher interpretability than the model using only deep neural networks.

6 Conclusion

In this paper, we studied the expert evaluation standard, and proposed an interpretable lexical grading method for essays. Our model accepted a sequence of word vectors as input and integrated attention mechanism in terms of lexical richness. Experimental results show that our model outperforms state-of art models for Chinese essay grading task. Besides, our method gives interpretable results, which are helpful for practical applications.

For future works, we are planning to study the syntactic characteristics and use them together for the essay grading task. One promising solution is to introduce the features on syntactic complexity and elegant sentences of essays. Another important direction is essay grading for the students in the same exam. Since their writing abilities are very close, the essay grading task is more challenging. We will also explore the prompt based Chinese essay grading task and provide useful feedback to authors.

Acknowledgments. This work was supported by the National Key Research and Development Program of China under Grant No. 2018YFC0831401, the National Natural Science Foundation of China under Grant No. 91646119, the Major Project of NSF Shandong Province under Grant No. ZR2018ZB0420, and the Key Research and Development Program of Shandong province under Grant No. 2017GGX10114. The scientific calculations in this paper have been done on the HPC Cloud Platform of Shandong University.

References

1. Attali, Y., Burstein, J.: Automated essay scoring with e-rater® V. 2. J. Technol. Learn. Assess. **4**(3), 1–30 (2006)
2. Juku Correction Website. https://www.pigai.org/
3. Graves, A.: Supervised sequence labelling with recurrent neural networks. Stud. Comput. Intell. **385**, 1–131 (2012)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
5. Page, E.B.: Grading essays by computer: progress report. In: Proceedings of the Invitational Conference on Testing Problems, pp. 87–100 (1967)
6. Daigon, A.: Computer grading of English essays. Engl. J. **55**(1), 46–52 (1966)
7. Foltz, P.W., Laham, D., Landauer, T.K.: The intelligent essay assessor: applications to educational technology. Interact. Multimedia Electron. J. Comput.-Enhanc. Learn. **1**(2), 939–944 (1999)
8. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Process. **25**(2–3), 259–284 (1998)
9. Rudner, L.: Computer grading using Bayesian networks-overview. Wayback Machine (2012)
10. Automated Student Assessment Prize (ASAP). https://www.kaggle.com/c/asap-aes
11. Alikaniotis, D., Yannakoudakis, H., Rei, M.: Automatic text scoring using neural networks. arXiv preprint. arXiv:1606.04289 (2016)
12. Dong, F., Zhang, Y., Yang, J.: Attention-based recurrent convolutional neural network for automatic essay scoring. In: Proceedings of the 21st Conference on Computational Natural Language Learning, pp. 153–162. ACL, Vancouver (2017)
13. Cozma, M., Butnaru, A.M., Ionescu, R.T.: Automated essay scoring with string kernels and word embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 503–509. ACL, Melbourne (2018)
14. Taghipour, K., Ng, H.T.: A neural approach to automated essay scoring. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1882–1891. ACL, Austin (2016)
15. Jin, C., He, B., Hui, K., et al.: TDNN: a two-stage deep neural network for prompt-independent automated essay scoring. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 1088–1097. ACL, Melbourne (2018)
16. Tay, Y., Phan, M.C., Tuan, L.A., et al.: SkipFlow: incorporating neural coherence features for end-to-end automatic text scoring. In: Thirty-Second AAAI Conference on Artificial Intelligence, pp. 5948–5955. AAAI, New Orleans (2018)
17. Ruiji, F., Dong, W., Shijin, W., Guoping, H., Ting, L.: Elegart sentence recognition for automated essay scoring. J. Chin. Inf. Process. **32**(6), 88–97 (2018)

18. Examination Center of the Office of the National HSK Examination Committee: Outline of Chinese Proficiency Vocabulary and Chinese Characters. Economic Science Press, Beijing (2001)
19. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196. IMLS, Beijing (2014)
20. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543. ACL, Doha (2014)
21. Shen, L., Zhe, Z., Renfen, H., Wensi, L., Tao, L., Xiaoyong, D.: Analogical reasoning on Chinese morphological and semantic relations. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 138–143. ACL, Melbourne (2018)
22. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations, Microtome, San Diego (2015)