Computers, Materials & Continua CMC, vol.65, no.1, pp.817-833, 2020

Automated Chinese Essay Scoring Based on Deep Learning

Shuai Yuan[1], Tingting He[2, 3, *], Huan Huang[4], Rui Hou[5] and Meng Wang[6]

Abstract: Writing is an important part of language learning and is considered the best approach to demonstrate the comprehensive language skills of students. Manually grading student essays is a time-consuming task; however, it is necessary. An automated essay scoring system can not only greatly improve the efficiency of essay scoring, but also provide more objective score. Therefore, many researchers have been exploring automated essay scoring techniques and tools. However, the technique of scoring Chinese essays is still limited, and its accuracy needs to be enhanced further. To improve the accuracy of the scoring model for a Chinese essay, we propose an automated scoring approach based on a deep learning model and validate its effect by conducting two comparison experiments. The experimental results indicate that the accuracy of the proposed model is significantly higher than that of multiple linear regression (MLR), which was commonly used in the past. The three accuracy rates of the proposed model are comparable to those of the novice teacher. The root mean square error (RMSE) of the proposed model is slightly lower than that of the novice teacher, and the correlation coefficient of the proposed model is also significantly higher than that of the novice teacher. Besides, when the predicted scores are not very low or very high, the two predicted models are as good as a novice teacher. However, when the predicted score is very high or very low, the results should be treated with caution.

Keywords: Automated essay scoring, deep learning, convolutional neural network, Chinese essay.

1 Introduction

Writing is important for language learning and testing, and it is also good to practice and demonstrate the language skills and knowledge of a student. Thus, in many high-stakes

[1] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, 430079, China.

[2] School of Computer, Central China Normal University, Wuhan, 430079, China.

[3] Information Retrieval and Knowledge Management Research Laboratory, Central China Normal University, Wuhan, 430079, China.

[4] School of Education, South-Central University for Nationalities, Wuhan, 430074, China.

[5] College of Computer Science, South-Central University for Nationalities, Wuhan, 430074, China.

[6] Information School, University of Washington, Seattle, WA98105, USA.

[*] Corresponding Author: Tingting He. Email: tthe@mail.ccnu.edu.cn.

Received: 06 March 2020; Accepted: 02 June 2020.

language assessments such as the Test of English as a Foreign Language (TOEFL), Graduate Record Examination (GRE), and Test of English for International Communication (TOEIC), writing is always one of the most important parts. However, evaluating these essays accurately and efficiently is not easy. One challenge is that essay ratings highly vary between humans as different human graders may attend to different features and hold different standards [McNamara, Crossley, Roscoe et al. (2015)]. A solution to this variability across raters has been to train expert raters to use scoring rubrics [McNamara, Crossley, Roscoe et al. (2015)]. While the reliability of human scores using scoring rubrics is considerably high, the sheer number of essays for these high-stakes assessments makes it cost-ineffective to have human raters exclusively score these assessments [Shermis (2014)]. Besides a large number of essays for these high-stakes assessments, more essays for some low-stakes assessment need to be scored by teachers. To facilitate the grading of these essays and alleviate the burden of teachers, many researchers began to explore the automated essay scoring (AES) techniques and systems.

Automated essay scoring is used to evaluating the quality of written essays via computer programs and provide a single score, a detailed evaluation of essay features, or both [Page (1966)]. Since Page [Page (1966)] and his colleagues proposed the first AES system, the field has been developing over 50 years. In the last decades, many good AES systems have been developed, e.g., Project Essay Grader (PEG), Intelligent Essay Assessor (IEA), Electronic Essay Rater (E-rater), and some others which have been used in some high-stakes standardized assessments. Compared with human scoring, automated essay scoring is more efficient than and as reliable as that obtained via human raters, and therefore, it is important for high-stakes standardized assessments. Except using the AES systems to grade essays for some high-stakes and low-stakes assessments, AES systems have been adopted to enhance writing instructions [Wilson and Czik (2016)]. In this case, its objective goes beyond solely providing an accurate score [McNamara, Crossley, Roscoe et al. (2015)]. The AES system provides students with immediate automated feedback in the form of essay ratings and individualized suggestions for improvement when revising [Wilson and Czik (2016)]. Although many good AES systems have been developed and applied in practice, most of them are for English essays. For Chinese essays, to the best of our knowledge, there does not exist any practical AES system. This is attributed to the limitation of Chinese information processing techniques, and because some advanced syntax and semantic features are difficult to extract, which can affect the accuracy of the predicted scores.

In recent years, deep neural networks have been used in many areas and obtained remarkable performance, especially in computer vision, speech recognition, and natural language processing as it can extract some advanced features automatically [Xu, Zhang, Xin et al. (2019)]. Accordingly, we try to apply the deep learning techniques in automated essay scoring, and propose a novel automated scoring approach for Chinese essays. The main contributions of this research include 1) a summary of the commonly used features for the automated scoring of Chinese essays; 2) a novel convolutional neural network (CNN) architecture for automated scoring of Chinese essays; and 3) two Chinese essay datasets, with essays written by native Chinese speakers and those obtained from standardized examinations. We compared our essay scoring approach with MLR based approach and the novice teacher on these datasets, and we showed that the

accuracy of our approach is significantly higher than the commonly used method and that of the novice teacher.

2 Related work

2.1 Automated scoring techniques for English essays

For English essays, there are many commercial AES systems such as PEG, IEA, E-rater, IntelliMetric, and Bayesian Essay Test Scoring System (BETSY), some of which have been used in high-stakes assessments. PEG is the first automated essay scoring system developed by Page in 1966 [Page (1966)]. The system uses some surface measures to approximate the intrinsic features of essays, and then, it uses MLR with these surface features to build a scoring model [Dikli (2006)]. IEA, produced by the Pearson Knowledge Analysis Technologies in the late 1990s, uses latent semantic analysis (LSA) techniques to predict the essay score [Foltz, Streeter, Lochbaum et al. (2013)]. Unlike PEG, IEA focuses on content-related features, and not on form-related features. E-rater, developed by the American Educational Testing Service in the late 1990s, extracts more complex features with natural language processing technology [Attali and Burstein (2006)]. The features not only include form-related ones, but also content-related ones. Further, it uses MLR to build a prediction model [Burstein, Tetreault and Madnani (2013)]. IntelliMetric is an AES system based on artificial intelligence, which integrates the domain knowledge of marking experts and is called a "learning machine that can internalize the collective wisdom of expert graders" [Burstein, Kukich, Wolff et al. (1998); Schultz (2013)]. BETSY was developed by Lawrence M. Rudner; it uses a Bayesian statistical model to score essays from the perspective of text classification [Rudner and Liang (2002)]. The system uses a large set of essay features, which include a large amount of content-related and form-related features [Rudner and Liang (2002)].

Over time, AES systems have slowly become embedded within automated writing evaluation (AWE) systems that assign scores along with feedback on errors [Roscoe and McNamara (2013)]. Examples include the PEG Writing, Criterion, MY Access!, and Writing Roadmap [Dikli (2006)]. The feedback provided by these systems are helpful for students to improve errors on mechanics, grammar, and spelling. However, these feedbacks have a negligible effect on improving essay performance. The most effective interventions for writing instruction are to explicitly and systematically teach students how to use strategies for planning, drafting, editing, and summarizing [Graham and Perin (2007)]. According to these, some researchers recently developed intelligent tutors for writing, which emphasize on writing strategy instructions and providing feedback that addresses deeper aspects of the essay. For example, Glosser is an automated feedback system that provides contextualized feedback to students about their professional texts [Calvo and Ellis (2010)]. The feedback includes four aspects: structure, coherence, topics, and keywords. Escribo is a computer-based scaffolding environment to facilitate student's development of expertise in academic writing [Proske, Narciss and McNamara (2012)]. In Escribo, students receive online support for prewriting, drafting, and revising processes, along with feedback about their choices at each stage [Roscoe and McNamara (2013)]. Writing-Pal is an intelligent tutor system that assists students to revise and improve their essays [Roscoe and McNamara (2013)]. It focuses on writing strategy

instruction and formative feedback and provides no specific error feedback on style, mechanics, spelling, or grammar. Besides, Lachner et al. [Lachner, Burkhart and Nücklesa (2017)] developed a visualization tool that visualizes cohesion deficits of student's explanations in a concept map.

2.2 Automated scoring techniques for Chinese essays

Compared to studies on the automated scoring of English essays, studies on automated scoring of Chinese essays started late [Liang and Wen (2007)], with limited practical systems. The studies on automated scoring of Chinese essays can be divided into two categories based on the evaluated essays: the first one evaluates Chinese essays written by native Chinese speakers; the second one evaluates those written by the non-native Chinese speakers.

For the first category, Cao et al. [Cao and Yang (2007)] explored the automated scoring of essays from a unified examination of a senior high school. They first used LSA to assess the content score of an essay, and then used the MLR to assess the final score of the essay. The correlation coefficient between the predicted scores and human scores was 0.55. Peng et al. [Peng, Ke, Zhao et al. (2012)] proposed three enhanced word scoring methods, and further used these methods to score essays automatically. Wang et al. [Wang, Li, He et al. (2016)] proposed an automated essay scoring method based on text semantic dispersion. They used deep learning techniques to extract the semantic dispersion features, and then integrated them into the multiple regression model. Experimental results suggested that dispersion features can significantly improve the accuracy of the predicting model. Fu et al. [Fu, Wang, Wang et al. (2018)] explored the techniques of elegant sentence recognition in Chinese essays, and further applied the extracted elegant sentence features into the AES task. The experimental results showed that the elegant sentence features reduced the large-margin predictive error. Zhong et al. [Zhong and Zhang (2019)] studied the extraction of linguistic intuition features, and explored the effects of these features on essay score prediction. The experimental results are promising.

For the second category, Li [Li (2006)] explored the automated scoring of essays which come from the national three-level Minzu Hanyu Kaoshi (MHK) test. He extracted 45 shallow features from essays, and then used MLR to build scoring models. Finally, he got a prompt-specified scoring model, and the correlation coefficient between the predicted scores and human scores was 0.566. Cai et al. [Cai, Peng, and Zhao (2011)] also explored the automated scoring of essays coming from the MHK test. As the relevance of shallow features with essay scores was not very high, they used natural language processing and information retrieval techniques to extract two advanced features. Further, they creatively proposed a triple-segmented regression for prediction modeling and found that it is more accurate than the common MRL. Unlike the above-mentioned studies, Huang et al. [Huang, Xie and Xun (2014)] explored the automated scoring of essays coming from the Hanyu Shuiping Kaoshi (HSK) test. Based on the writing assessment in an HSK test, they proposed 107 shallow features and discovered 19 features that have strong correlation with the score of the essay. Using these features, they employed MRL to build a scoring model, and found that its results are more accurate than the baseline.

2.3 Summary of related works

Analyzing all related works comprehensively, we can see that there have been many practical AES systems and AWE systems for English essays. Some AES systems have been used in some high-stakes or low-stakes assessments, and some AWE systems have been widely used in writing instruction. Currently, researchers are focusing on the automated generation of formative feedback that addresses deep aspects of the essay. However, for Chinese essays, there is no practical AES system yet. The focus of researchers is on the extraction of advanced features to improve scoring accuracy.

Most AES systems followed a typical methodology. First, a set of target essays are collected and scored by expert teachers. Second, a set of features are extracted from the essays by statistical methods and natural language processing techniques. Finally, a computational algorithm is used to train a prediction model using the extracted features. Based on this workflow, the features and computational algorithm are key components of AES system. For feature extraction, many advanced features have been extracted from English essays with the development of natural language processing techniques. However, for Chinese essays, we can only extract some surface features because of the limitation of natural language processing techniques. For computational algorithm, most practical AES systems used MRL to train a scoring model, and only few ones used the Bayesian method. Recently, many researchers have explored automated scoring techniques based on deep learning for English essays and got many promising results [Dong and Zhang (2016); Taghipour and Ng (2016); Dong, Zhang and Yang (2017); Jin, He, Hui et al. (2018)]. For Chinese essays, some researchers also explored deep learning to extract some advanced features, such as elegant sentence, linguistic intuition, and so on [Fu, Wang, Wang et al. (2018); Zhong and Zhang (2019)].

3 Automated essay scoring framework based on deep learning

Based on a literature review, we know that it is difficult to extract deep linguistic and semantic features such as enrichment, fluency, and rhetoric from Chinese essays, because the grammatical structure is more complex in Chinese. To discover these advanced features automatically and improve scoring accuracy, we propose using deep learning techniques to improve automated essay scoring. The automated essay scoring framework based on deep learning is shown in Fig. 1.

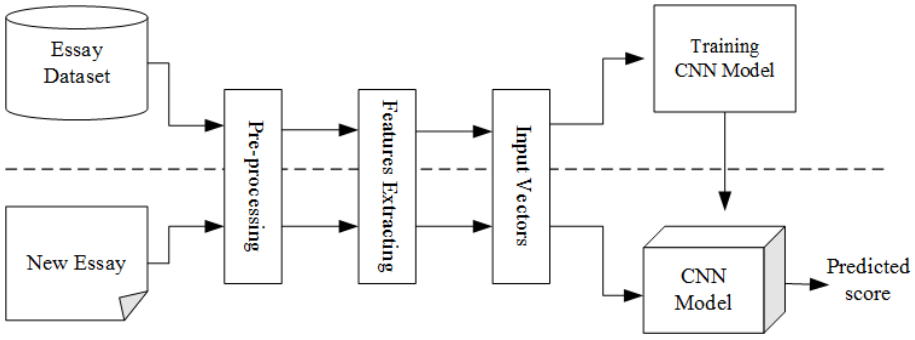

Figure 1: Automated essay scoring framework based on deep learning

Fig. 1 shows that the framework includes two parts: prediction modeling, which uses a standard essay dataset to train a CNN model for essay scoring, and essay scoring, which applies the trained CNN model to some new essays and returns the predicted scores. Before CNN model training and application, we must extract some features from essays and transform each essay to a vector.

3.1 Selected features

Features are very important for prediction modeling. However, it is also a difficult task especially for unstructured data such as texts. Over the last 50 years, researchers have explored many different features for automated essay scoring from multiple perspectives. Recently, Zupanc et al. [Zupanc and Bosnic (2017)] systematically compared the AES systems from three aspects: type of attributes, methodology, and prediction model. They divided the features used by the state-of-art systems into three groups: style, content, and semantic features [Zupanc and Bosnic (2017)]. Style features focus on linguistic characteristics such as lexical sophistication, grammar, and mechanics. Content features describe the semantic similarity between the source essay and the already graded essays. Semantic features are based on verifying the correctness of contextual meaning. In their study, they summarized 72 linguistic and content features that had been used by previous studies and proposed 29 semantic features [Zupanc and Bosnic (2017)]. These features are import for future research.

For automated Chinese essay scoring, Cai et al. [Cai, Peng and Zhao (2011)] studied the effect of the level of linguistic difficulty and degree of content agreement on the MHK essay scoring. Huang et al. [Huang, Xie and Xun (2014)] extracted 107 features from HSK essays and explored their effects on automated scoring. These features describe Chinese character usage, wording usage, grammatical errors, paragraph expression, and degree of elegant-formality. More recently, Wang et al. [Wang, Li, He et al. (2016)] proposed two representation methods of text semantic dispersion, and further explored their effects on automated Chinese essay scoring. Fu et al. [Fu, Wang, Wang et al. (2018)] proposed a method to recognize elegant sentences and explored its effect on automated essay scoring. Zhong et al. [Zhong and Zhang (2019)] extracted some language intuition features and found they are useful for essay score predicting. From the above-mentioned studies, it is clear that features used in automated Chinese essay scoring also include three aspects: linguistics, content, and semantics. However, many advanced features are still difficult to extract automatically.

As our focus is not feature selection and extraction, we only selected 35 commonly used features for our study. The features are listed in Tab. 1; these features are divided into two groups: linguistic and semantic features.

Table 1: Linguistic and semantic features

Type of features	Features
Linguistic	1. Number of words
	2. Number of POS tags
	3. Frequency of nouns in 1000 words
	4. Frequency of verbs in 1000 words
	5. Frequency of adjectives in 1000 words

	6. Frequency of adverbs in 1000 words
	7. Frequency of pronouns in 1000 words
	8. Frequency of prepositions in 1000 words
	9. Frequency of interrogatives in 1000 words
	10. Frequency of numerals in 1000 words
	11. Frequency of conjunctions in 1000 words
	12. Frequency of particles in 1000 words
	13. Frequency of function words in 1000 words
	14. Frequency of localizers in 1000 words
	15. Frequency of measure words in 1000 words
	16. Frequency of Onomatopoeias in 1000 words
	17. Age of acquisition of words
	18. First person singular pronoun incidence
	19. First person plural pronoun incidence
	20. Second person singular pronoun incidence
	21. Second person plural pronoun incidence
	22. Third person singular pronoun incidence
	23. Third person plural pronoun incidence
	24. Number of sentences
	25. Average of sentence length--Character
	26. Standard deviation of sentence length--Character
	27. Average of sentence length--Word
	28. Standard deviation of sentence length--Word
	29. Number of paragraphs
	30. Average of paragraph length
	31. Standard deviation of paragraph length
Semantic	32. Average of LSA overlap between the adjacent sentences
	33. Standard deviation of LSA overlap between the adjacent sentences
	34. Average of LSA overlap between the sentence and its preceding sentences
	35. Standard deviation of LSA overlap between the sentence and its preceding sentences

3.2 CNN architecture

After each essay is transformed into a vector with the feature extraction techniques, we need a machine learning algorithm to train a scoring model. Over the last 50 years, researchers have explored many traditional machine learning algorithms, such as linear regression, logistic regression, random forest, support vector machine, and so on. As is known, these traditional machine learning algorithms rely heavily on the features selected. However, for Chinese essays, we can only extract simple features because of the limited natural language processing techniques, and this seems to be the case regardless of the score of the essay [Cai, Peng and Zhao (2011); Huang, Xie and Xun (2014)]. To resolve this problem, we propose using deep learning to train a scoring model for Chinese essays,

because deep learning can construct advanced and abstract features automatically through the combination of simple features.

Originally invented for computer vision, CNN models have subsequently been shown to be effective for many natural language processing tasks [Kim (2014)]. These models can recognize advanced and abstract features by applying convolution operation to local features. Therefore, we propose a CNN model to recognize the advanced and abstract features of Chinese essays and predict the essay scores more precisely. The proposed CNN architecture, as shown in Fig. 2, is a slight variant of the LeNet-5 [Lecun, Bottou, Bengio et al. (1998)].

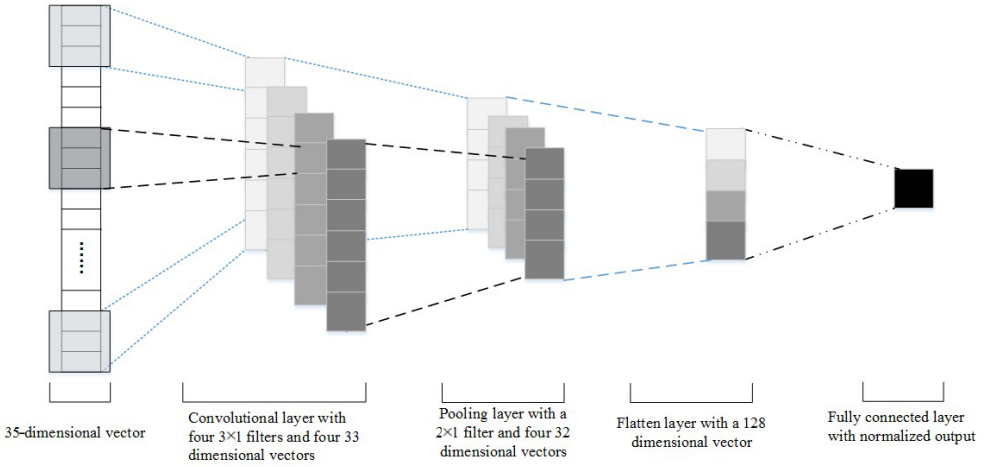

Figure 2: Model architecture for an example essay

The proposed CNN model includes five layers: input, convolutional, pooling, flatten and fully connected layers. The input layer is a 35-dimensional vector, and each dimension represents a feature of an essay. As the input layer is a vector and not a matrix, we use a one-dimensional convolutional operation on the input vector. A convolutional operation is applied to each window, and it produces a new feature map. After the convolutional operation, we use the Rectified Linear Unit (ReLU) function to introduce nonlinearity to the model. As we use four 3×1 convolutional filters, we obtain four new feature maps; each feature map is a 33-dimensional vector. We then apply the max-over-time pooling operation over each feature map with a 2×1 filter and obtain four 32-dimensional vectors. Next, we flatten four vectors and form the penultimate layer and pass it to the fully connected layer. As essay grading is a regression problem, the fully connected layer only has one node and it does not use any active function. Finally, we choose the *mean squared error* as the loss function. However, to prevent the result of the fully connected layer from exceeding the maximum score, we apply the following regularization function to the result [Wang, Li, He et al. (2016)].

$$\text{Final}Score = FullScore \times sigmoid\left(\frac{pre\_score - \overline{score}}{s/\sqrt{n}}\right) \tag{1}$$

where *FullScore* represents the maximum score of an essay, *pre_score* represents the output result of fully connected layer, $\overline{score}$ represents the average score of samples, *s* represents the standard deviation of the scores of samples, *n* represents the number of samples, and *sigmoid* represent the *sigmoid* function.

4 Experiment and result analysis

To validate the above essay scoring approach, three comparative experiments were conducted on two Chinese essay datasets that we developed.

4.1 Essay dataset

To train and validate the essay scoring models, we first need a standard essay dataset. Researchers have already developed some essay datasets for automated essay scoring. For example, in a demonstration of existing and emerging automated scoring systems for essays sponsored by the Hewlett Foundation, 22029 student essays were collected for eight different prompts representing six states [Shermis (2014)]. However, most of these essay datasets are for the English language. Only few datasets are for the Chinese language [Huang, Xie and Xun (2014)], and they are not public datasets. Further, the essays of most of these datasets are written by non-native Chinese speakers or they come from non-standard assessments. Therefore, we developed two Chinese essay datasets for our research.

The first dataset includes 100 essays for one prompt, and all essays come from a final exam conducted in the fall semester of 2017. The students are from a middle school in China (grade level 9). As the essays are from a large-scale standard test, each essay was pre-scored by two different experienced Chinese language teachers. The final score of each essay is the average of two scores. The second dataset includes 106 essays written by students in another middle school. The students are also from grade level 9, and these essays are also for one prompt. However, these essays are from a mid-term examination in the fall semester of 2017 and are scored by only one expert teacher. The total points of both prompts are 50 points. The distributions of the final scores are shown in Fig. 3.

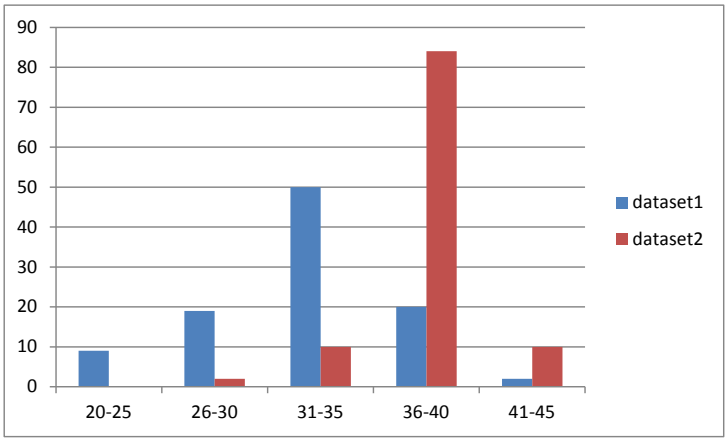

Figure 3: Distribution of the final scores and distribution of the length of the essays

Unlike the datasets used in previous research studies, the developed datasets have two important features: First, the essays come from the mediate or final examination of grade 9. These examinations are important for students, and therefore, all teachers graded the essays using the same rubric; this helps ensure the reliability of the scores. Second, the essays were written by native Chinese speakers. This makes the scoring model that we trained more suitable for most high-stakes assessments in China, such as that of the National College Entrance Examination.

4.2 Evaluation criteria

To evaluate the accuracy of the essay scoring model, researchers generally used three evaluation criteria: correlation coefficient, root mean square error (RMSE), and accuracy. In the experiment, we use these indicators to compare the accuracy of the proposed essay scoring model to that of the baseline.

Root mean square error is commonly used to measure the difference of two groups of numerical values, and it is widely used to test the accuracy of essay scoring model. The calculating formula is shown as follow:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \tag{2}$$

where n represents the number of essays, x_i represents the predicted score of an essay, and y_i represents the score of the essay given by the teachers.

Pearson's correlation coefficient is another commonly used indicator to test the accuracy of the essay scoring model. The Pearson's correlation coefficient can be calculated as following Eq. (3).

$$\mathrm{R} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

where n, x_i, and y_i have the same meaning in Eq. (2), and $\bar{x}$ is the mean of predicted scores, and $\bar{\mathrm{y}}$ is the mean of scores given by teachers.

Besides the root mean square error (RMSE) and correlation coefficient, some researchers proposed using the accuracy rate to test the accuracy of the scoring model [Huang, Xie and Xun (2014)]. The accuracy rate of an essay scoring model can be calculated as following Eq. (4).

$$\mathrm{A}(p) = \frac{n_p}{N} \times 100\% \tag{4}$$

where n represents the number of essays, n_p is the number of essays whose score is more or less p points compared to the scores given by teachers.

4.3 Implementation and results

Based on the developed datasets, we conducted three comparative experiments to validate the proposed automated scoring approach.

4.3.1 Comparison of the CNN-based model and the MLR-based model

In the first experiment, we compared the essay scoring model based on CNN (CNN-based Model) with the essay scoring model based on MLR (MLR-based Model), which was used by most previous studies. The experiment process included three steps. First, we used the Chinese version of *Coh-Metrix*[1] to extract the 35 selected features of each essay, and we transformed each essay into a 35-dimensional vector. Second, we used the *skikit-learn*[2] framework to implement the MLR-based scoring model. Third, we used the *Keras*[3] and *Tensorflow*[4] frameworks to implement the CNN-based scoring model. Finally, we used the transformed vectors and the scores given by teachers as input to train the two scoring models and validate their effects.

When training the MLR-based scoring model, we did not use the feature selection method to delete any feature. However, we eliminated collinear features to guarantee stability and efficiency of the scoring model. When training the CNN-based scoring model, we chose Root Mean Square Prop (*RMSProp)* as an optimizer and maintained the default parameters. Besides, we set the maximum number of iterations for the scoring model training to 250, and the *mean square error* will be evaluated every time on the testing data. As the sample sizes are small, we adopted five-fold cross-validation in this experiment to avoid sample selection bias. Thus, each dataset was randomly divided into 5 groups, and each group was used as testing data, while the other 4 groups were used as training data. That is, every instance has a predicted value. The experimental results on dataset 1 and dataset 2 are shown respectively in Tabs. 2 and 3:

Table 2: Comparison of the MLR-based model and the CNN-based model on dataset 1

Model	A (3)	A (5)	A (7)	RMSE	R
MLR-based model	45%	65%	77%	6.02	0.28
CNN-based model	53%	75%	87%	4.73	0.37

Table 3: Comparison of the MLR-based model and the CNN-based model on dataset 2

Model	A (3)	A (5)	A (7)	RMSE	R
MLR-based model	77.36%	95.28%	97.17%	2.73	0.29
CNN-based model	59.43%	86.79%	94.34%	3.98	0.45

From Tab. 2, we can see that the three accuracy rates of the CNN-based scoring model are significantly higher than that of the MLR-based scoring model. The correlation coefficient of the CNN-based scoring model is much higher than the MLR-based scoring model. Besides, the root mean square error of the CNN-based scoring model is slightly smaller than the MLR-based scoring model. These results indicate that the performance

[1] http://www.memphis.edu/iis/projects/coh-metrix.php.

[2] http://www.rapidminer.com.

[3] https://keras.io/.

[4] https://tensorflow.google.cn.

of the CNN-based scoring model is better than the MLR-based scoring model for dataset 1. However, Tab. 3 indicates that the three accuracy rates of the CNN-based scoring model are lower than that of the MLR-based scoring model. The RMSE of the CNN-based scoring model is also higher than the MLR-based scoring model, which suggest that the accuracy of the CNN-based scoring model is better than that of the MLR-based scoring model on dataset 2. Nevertheless, the correlation coefficient of the CNN-based scoring model is significantly higher than the MLR-based scoring model, which means the scores predicted by the CNN-based scoring model are more consistent with the human scores.

4.3.2 Comparison of the CNN-based model and a novice teacher

From Tabs. 2 and 3, we can also see that the agreements between the predicted scores and the human scores are poor. As essay ratings are very subjective and depend on the assessor (especially for novice teachers), we checked if the predicted result was comparable to the result provided by a novice teacher. In the second experiment, we compared the scores predicted by the CNN-based scoring model with the scores given by another novice teacher. After we finished the first experiment, we asked another novice teacher to score each essay based on the same rubric. Then, we respectively calculated their accuracy rates, root mean squared errors, and correlation coefficients. The experimental results on datasets 1 and 2 are listed in Tabs. 4 and 5 respectively.

Table 4: Comparison of the CNN-based model and the novice teacher on dataset 1

Model	A (3)	A (5)	A (7)	RMSE	R
Novice teacher	55%	75%	86%	5.11	0.24
CNN-based model	53%	75%	87%	4.73	0.37

Table 5: Comparison of the CNN-based model and the novice teacher on dataset 2

Model	A (3)	A (5)	A (7)	RMSE	R
Novice teacher	50.94%	81.13%	96.23%	4.18	0.25
CNN-based model	59.43%	86.79%	94.34%	3.98	0.45

Tab. 4 indicates that the three accuracy rates of the CNN-based scoring model are comparable to those of the novice teacher. The RMSE of the CNN-based scoring model is slightly lower than that of the novice teacher. The correlation coefficient of the CNN-based scoring model is also significantly higher than that of the novice teacher. From Tab. 5, we can see that the similar result. These results suggest that the results of the CNN-based scoring model are comparable to those of the novice teacher.

4.3.3 Comparison of the CNN-based model, the MLR-based model and the novice teacher on samples in different ranges of scores

To further validate the practicality of the scoring model based on CNN, we conducted the third experiment. In this experiment, we further compared the accuracy rates of the MLR-based scoring model, CNN-based scoring model, and those of the novice teacher on samples in different ranges of scores. As all accuracy rates are very low when the error is

less than 3, and the accuracy rates are nearly equal when the error is less than 7; therefore, we only compared accuracy rates when the error is less than 5. Besides, as the number of predicted scores in the range of 20-25 is low, we did not consider the accuracy rates in this range in dataset 2. The experimental results on datasets 1 and 2 are shown in Figs. 4 and 5, respectively.

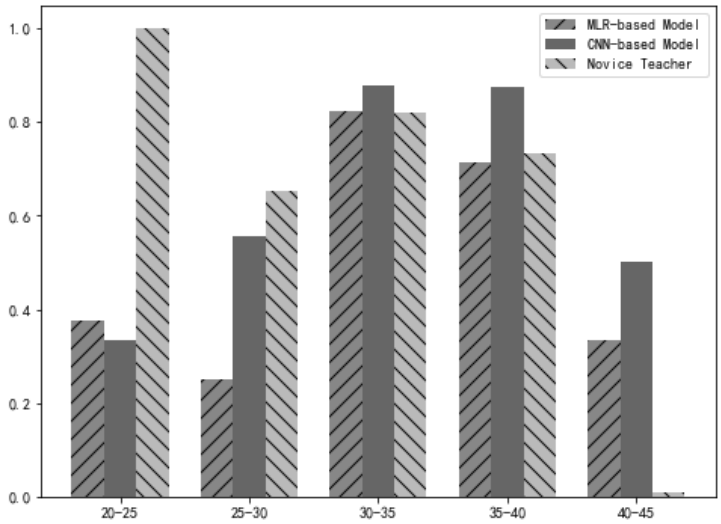

Figure 4: Accuracy rates for a different range of scores on datasets 1

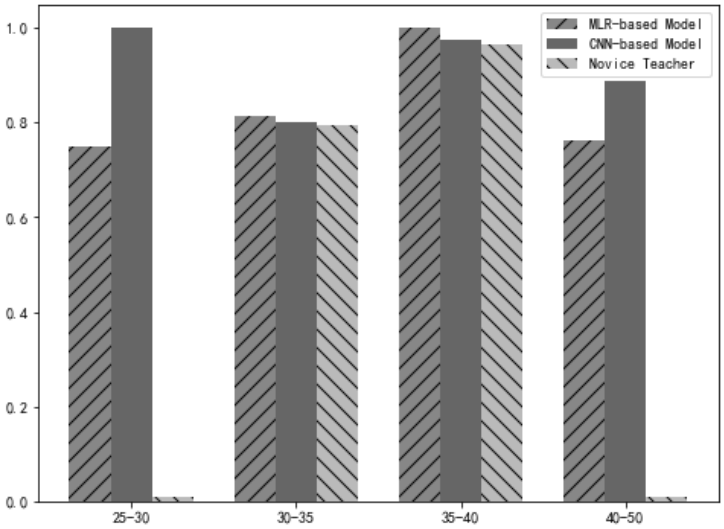

Figure 5: Accuracy rates for a different range of scores on datasets 2

According to Figs. 4 and 5, the accuracy rates of the MLR-based scoring model, CNN-based scoring model, and the novice teacher are very high in range of 30-40. In this range, the CNN-based scoring model outperforms the MLR-based scoring model and the novice teacher in dataset 1, and the results are comparable in dataset 2. However, when the

predicted score is very high or very low, the accuracy rates of the CNN-based scoring model and the MLR-based scoring model are not sufficiently high, and the accuracy rates of the novice teacher are also very bad. These suggest that the predicted score is more reliable if the predicted score is moderate.

5 Conclusion and discussion

In this paper, we proposed a new approach based on deep learning technology to score Chinese essays automatically. To validate the effectiveness of the proposed approach, we compared the results with a commonly used predictive method, multiple linear regression. According to the experimental results, we found that the accuracy of the essay scoring model based on deep learning is significantly higher than that of the model based on multiple linear regression. There result is attributed to two reasons: the CNN-based model can represent the complex nonlinear relation between the features and the score, and the CNN-based model captured some abstract features from combination of the basic features, which is significant for essay scoring.

To validate the practicality of the proposed approach, we further compared its results with those of a novice teacher. The experimental results indicated that the accuracy of the CNN-based model outperforms that of the novice teacher. Further, we also compared the accuracy rate of the MLR-based model, CNN-based model, and of the novice teacher on samples under different ranges of score. We found the CNN-based model and the MLR-based model are as good as the novice teacher when the predicted scores are moderate. This implies that the predicted score of the proposed approach has significant referential meaning when the predicted score is in middle range. However, when the predicted score is very high or very low, the results should be treated with caution. Based on the experimental results, we can also say that the automatic scoring for Chinese essays is more challenging than that for English essays.

In the future study, we plan to overcome the following shortcomings of this study: First, the selected features need to be further expanded as we only consider 35 features in this study; second, the sizes of the essay datasets are not large enough, which need to be further supplemented; and third, the existing convolutional neural network model needs to be further improved.

Funding Statement: This work is supported by the National Science Foundation of China (No. 61532008; No. 61572223), the National Key Research and Development Program of China (No. 2017YFC0909502), and the Ministry of Education of Humanities and Social Science project (No. 20YJCZH046).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Attali, Y.; Burstein, J. (2006): Automated essay scoring with e-rater V.2.0. *Journal of Technology, Learning and Assessment*, vol. 4, no. 3, pp. 1-21.

Burstein, J.; Kukich, K.; Wolff, S.; Lu, C.; Chodorow, M. (1998): Enriching automated essay scoring using discourse marking. *Proceedings of the Workshop on Discourse Relations and Discourse Marking, Annual Meeting of the Association of Computational Linguistics*, Montreal, Canada, pp. 15-21.

Burstein, J.; Tetreault, J.; Madnani, N. (2013): The E-rater: An Automated Essay Scoring System. In: Shermis, M. D.; Burstein, J. C., *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, pp. 55-67. Routledge, New York.

Cai, L.; Peng, X. Y.; Zhao, J. (2011): Research on assisted scoring system for Chinese proficiency test for minorities (in Chinese). *Journal of Chinese Information Processing*, vol. 25, no. 5, pp. 120-126.

Calvo, R. A.; Ellis, R. A. (2010): Students' conceptions of tutor and automated feedback in professional writing. *Journal of Engineering Education*, vol. 99, no. 4, pp. 427-438.

Cao, Y. W.; Yang, C. (2007): Automated Chinese essay scoring with latent semantic analysis (in Chinese). *Examinations Research*, vol. 3, no. 1, pp. 63-71.

Dikli, S. (2006): An overview of automated scoring of essays. *Journal of Technology Learning & Assessment*, vol. 5, no. 1, pp. 1-36.

Dong, F.; Zhang, Y. (2016): Automatic features for essay scoring-an empirical study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1072-1077.

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

Foltz, P. W.; Streeter, L. A.; Lochbaum, K. E.; Landauer, T. K. (2013): Implementation and applications of the intelligent essay assessor. In: Shermis, M. D.; Burstein, J. C., *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, Routledge, New York, pp. 68-88.

Fu, R. J.; Wang, D.; Wang, S. J.; Hu, G. P.; Liu, T. (2018): Elegant sentence recognition for automated essay scoring (in Chinese). *Journal of Chinese Information Processing*, vol. 32, no. 6, pp. 88-97.

Graham, S.; Perin, D. (2007): A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, vol. 99, no. 3, pp. 445-476.

Huang, Z. E.; Xie, J. L.; Xun, E. D. (2014): Study of feature selection in HSK automated essay scoring (in Chinese). *Computer Engineering and Applications*, vol. 50, no. 6, pp. 118-122.

Jin, C. C.; He, B.; Hui, K.; Sun, L. (2018): TDNN: a two-stage deep neural network for prompt-independent automated essay scoring. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 1088-1097.

Kim, Y. (2014): Convolutional neural networks for sentence classification. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1746-1751.

Lachner, A.; Burkhart, C.; Nücklesa, M. (2017): Formative computer-based feedback in the university classroom: specific concept maps scaffold students' writing. *Computers in Human Behavior*, vol. 72, pp. 459-469.

Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. (1998): Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324.

Li, Y. N. (2006): *Automated Essay Scoring for Testing Chinese As Second Language* (in Chinese). Beijing Language and Culture University, China.

Liang, M. C.; Wen, Q. F. (2007): A critical review and implications of some automated essay scoring systems (in Chinese). *Technology Enhanced Foreign Language Education*, no. 5, pp. 18-24.

McNamara, D. S.; Crossley, S. A.; Roscoe, R. D.; Allen, L. K.; Dai, J. M. (2015): A hierarchical classification approach to automated essay scoring. *Assessing Writing*, vol. 23, pp. 35-59.

Page, E. B. (1966): The imminence of grading essays by computer, *Phi Delta Kappan*, vol. 47, no. 5, pp. 238-243.

Peng, X. Y.; Ke, D. F.; Zhao, Z.; Chen, Z. B.; Xu, B. (2012): Automated Chinese essay scoring based on word scores (in Chinese). *Journal of Chinese Information Processing*, vol. 26, no. 2, pp. 102-108.

Proske, A.; Narciss, S.; McNamara, D. S. (2012): Computer-based scaffolding to facilitate students' development of expertise in academic writing. *Journal of Research in Reading*, vol. 35, no. 2, pp. 136-152.

Roscoe, R. D.; McNamara, D. S. (2013): Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, vol. 105, no. 4, pp. 1010-1025.

Rudner, L. M.; Liang, T. (2002): Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, vol. 1, no. 2, pp. 3-21.

Schultz, M. T. (2013): The IntelliMetric automated essay scoring engine-a review and an application to Chinese essay scoring. In: Shermis, M. D.; Burstein, J. C., *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, pp. 55-67. Routledge, New York.

Shermis, M. D. (2014): State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, vol. 20, pp. 53-76.

Taghipour, K.; Ng, H. T. (2016): A neural approach to automated essay scoring. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1882-1891.

Wang, Y. H.; Li, Z. J.; He, Y. Y.; Chao, W. H.; Zhou, J. S. (2016): Research on key technology of automatic essay scoring based on text semantic dispersion (in Chinese). *Journal of Chinese Information Processing*, vol. 30, no. 6, pp. 173-181.

Wilson, J.; Czik, A. (2016): Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, vol. 100, pp. 94-109.

Xu, F.; Zhang, X. F.; Xin, Z. H.; Yang, A. L. (2019): Investigation on the Chinese text sentiment analysis based on convolutional neural networks in deep learning. *Computers, Materials & Continua*, vol. 58, no. 3, pp. 697-709.

Zhong, Q. D.; Zhang, J. X. (2019): Chinese composition scoring algorithm embedded with language deep perception (in Chinese). *Computer Engineering and Applications* (in Press).

Zupanc, K.; Bosnic, Z. (2017): Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*, no. 120, pp. 118-132.