Understanding Expert Knowledge for Chinese Essay Grading

Xiaoyue Liu[1], Yi Xie[1,2], Tao Yang[1], and Yuqing Sun[1,3](✉)

[1] School of Software, Shandong University, Jinan, China
sun_yuqing@sdu.edu.cn
[2] School of Computer Science and Technology, Shandong University, Qingdao, China
[3] Engineering Research Center of Digital Media Technology, Ministry of Education, Shandong University, Jinan, China

Abstract. Essay grading is an important issue in natural language processing. There are two challenges for Chinese essay grading, namely the subjectivity of expert grading standards and the lack of fine-grained labeled data. In this paper, we propose an automatic Chinese essay grading method based on multi-aspect expert knowledge. We introduce essay grading expert rules to turn the existing standards into indexes, such as 'The Essay Grading Standards for College Entrance Examination' and 'The Chinese Curriculum Standards for Compulsory Education'. Based on the expert rules, we propose different encoders to learn multiple essay features in three aspects, namely the topic consistency, structure rationality and linguistics proficiency. An essay is graded by unifying the three grades in different aspects. Experimental results on two real datasets show the effectiveness of our method. We also analysis the influence of each aspect on the essay grading results. The experiment on the material essay grading dataset shows the practicability of our model in general exam scenarios.

Keywords: Essay grading · Multiple aspects · Expert knowledge

1 Introduction

Essays are the logical organization of texts based on fixed topics and the students' ideas. Compared with manual essay grading, many existing automated essay grading models have the advantages of low cost, high efficiency, systematic and unified grading standards, and freedom from the subjectivity of evaluation experts. Therefore, educational institutions gradually introduce automated essay grading models to replace part or all of the manual grading in some essay examination scenarios.

Chinese essay grading has specific challenges compared to traditional English essay grading tasks. The first challenge is the subjectivity of expert grading standards. Existing expert standards for Chinese essay grading are mainly divided into two types: the first type is the instructional standards in the teaching scenarios, such as 'The Chinese Curriculum Standards for Compulsory Education' [1], which are focus on cultivating the writing abilities of students with different cognitive levels in different education grades;

Y. Sun et al. (Eds.): ChineseCSCW 2021, CCIS 1491, pp. 488–501, 2022.
https://doi.org/10.1007/978-981-19-4546-5_38

another type is the instructional standards for essay evaluation experts in the examination scenarios, such as 'The Essay Grading Standards for College Entrance Examination' [2]. In the above two kinds of standards, teachers are instructed to grade essays by combining the standards with their own subjective judgements, such as 'A low grade essay always be far off the point, a normal grade essay should keep to the point, while a high-grade essay's point should be profound'. However, it difficult for experts to quantify essays in a uniform way according to such standards.

Another challenge is the lack of fine-grained labeled datasets. Existing Chinese essay grading datasets always contain essay texts and overall grades for full marked or excellent essays, lack of essays with various grades in different cognitive levels. Some datasets have the topic or category labels of essays, while few datasets have the grading comments given by experts. It is difficult for automated essay grading models to learn interpretability features from multiple aspects of essays.

To tackle the above challenges, we propose a Chinese essay grading method based on multi-aspect expert knowledge. The contributions of this paper are listed below:

(1) We introduce Chinese essay grading expert rules that integrates existing expert standards. The multi-aspect expert knowledge is proposed to grade the topic consistency, structure rationality and linguistics proficiency.
(2) We propose the joint multi-aspect essay feature encoders based on pre-trained language models. The encoders are adopted to represent the essay topic, structure, and linguistics, respectively. The multi-aspect essay grades are calculated based on these representations.
(3) We combine the multi-aspect essay grades with the attention mechanism [20], and integrate them into an overall grade for each essay.
(4) The method is verified against real datasets and the experimental results show that it outperforms other methods on the Chinese essay grading task. We also analysis the influence of each aspect on the essay grading results. The experiment on the material essay grading dataset shows the practicability of our model in general exam scenarios.

The rest of this paper is organized as follows. Section 2 presents the related works. Section 3 introduces existing expert essay grading standards and the datasets. Section 4 presents the Chinese essay grading method based on multi-aspect expert knowledge. Section 5 evaluates our model on real datasets. We conclude our paper in Sect. 6.

2 Related Work

The classical automatic essay grading works mainly use machine learning methods to analyze the intrinsic essay features and predict the English essay grades. in 1966, Ellis Page [3] developed the first automatic essay grading system, which uses surface features such as the number of words in the essay to make judgments, and does not involve the semantic part of the essay. It can perform batch review, which greatly improves the efficiency of essay grading. In the 1990s, the essay grading system added features such as vocabulary, grammar, syntax, and semantic similarity to the essay content, such as

Intelligent Essay Assessor [4] and Electronic Essay Rater [5] etc. In recent years, machine learning has been gradually applied to essay grading tasks. In 2006, Rudner developed a essay grading system IntelliMetricTM [6], which extracting features from multiple aspects such as grammar, syntax, text content, etc., greatly improves the consistency with manual review, and can review essays in multiple languages. Some works use the Naive Bayes method to merge the nearest neighbor classification and other statistical methods to convert the essay scoring problem into a text classification problem, such as [7–9] etc.

Recently, many essay grading works adopt deep learning methods. Dasgupta et al. [10] chose to consider the representation vectors of words and sentences, and use the convolutional neural network of the attention-pooling mechanism on the word vectors to extract the internal relationship between the essay contents. Dong et al. [11] construct a hierarchical convolutional neural network review model to examine essay sentence structure and article structure. Wang et al. [12] used an enhanced model framework combined with the second weighted Kappa coefficient to review the essay. Tay et al. [13] proposed an improved model for the LSTM structure. By modeling the hidden state at different time steps, they improved the memory problem that exists when using recurrent neural networks such as RNN and LSTM to process long texts. Alikaniotis et al. [14] mainly consider the influence of vocabulary in the essay. They generate special word representations by learning the contribution of specific vocabulary to the essay score, thereby improving the effect of the model. Jin et al. [15] proposed a two-stage deep neural network model. The first stage uses data under non-current topics to train a shallow model, and the second stage is an end-to-end model for scoring.

However, the current related works mainly focuses on English essays. The study of Chinese essay grading tools has theoretical significance and application value for in-depth exploration.

3 Quantifying Expert Knowledge for Essay Grading

3.1 The Expert Standards for Essay Grading

We first present a few existing expert standards for essay grading. In practice, there are a few standards for professionals to grade essay, which represent the expert knowledge on essays. We choose two authoritative standards: (1) “The Chinese Curriculum Standards for Compulsory Education” officially issued by the Ministry of Education in 2018. This framework is to guide cultivating the writing ability of students in primary and middle school. We list some key points in Table 1. (2) “The Essay Grading Standards for College Entrance Examination” officially issued by the Examination Institute. We list some key points in Table 2.

3.2 The Proposed Metrics Based on Expert Standards

In this Section, we propose a set of quantified metrics based on these standards that cover three aspects of topic, structure, and linguistics. Based on the above authoritative standards on essay grading, we propose the unified metrics that cover three aspects of an essay:

Table 1. Some key points of The Chinese Curriculum Standards for Compulsory Education.

Student grade	Education goals
1–2	• Cultivate interest in writing and observation skills
3–4	• Cultivate interest in writing • Able to communicate in short letters and notes • Accumulate and use language materials, use novel words and sentences
5–6	• Able to write simple documentary essays and imaginative essays, with specific content and sincere emotions • Able to segment reasonably on demand • Fluent sentences, correct writing
7–9	• Use reasonable expressions according to needs. Reasonably arrange the order and details of content, and express your meaning clearly. Enrich the content using association and imagination • Enrich content in narrative writing; clear in expository writing; well-founded in argumentative writing; able to do practical writing according to daily needs

Table 2. Some key points of The Chinese essay grading standards for 2019 college entrance examination.

Basic level	Advanced level
Fit the topic	The topic is profound; Going deep into essence through phenomena, revealing the inner relations of things; the viewpoints are enlightening
Fit the essay type requirements	The essay is rich in content. The argument is substantial, vivid and connotative
Sincere emotions, healthy thoughts, fluent language and intact structure	The essay is full of literary talent. It is appropriate in expression, flexible in sentence structure, good at using rhetorical skills, expressive in sentences
Rich in content and clear topic	The essay is innovative. The viewpoints and examples are novel, the ideas are new and clever, the reasonings and imaginations are unique

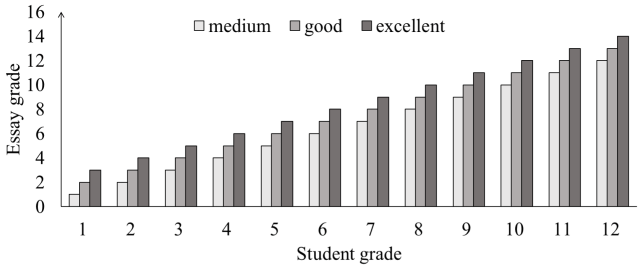
(1) Linguistics proficiency. It considers the fluency and rationality of sentences in an essay. It reflects whether the writing follows the usual usage of language. This metric can be quantified on different levels, namely sentence level and document level.
(2) Topic consistency. It judges whether the essay is related to the given topic and whether there is always only one definite topic.
(3) Structure rationality. It reflects the author's logical thinking and the ability to organize materials. Taking argumentative essays as example, its common structure is to

give the argument at the beginning, gradually introduce materials and examples to prove the argument, and summarize at the end.

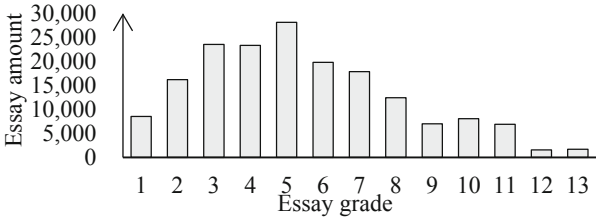
3.3 The Dataset of Essay Grading on Primary and Middle School

The existing essay grading datasets have some shortcomings. Many publicly available Chinese essay datasets only include excellent essays, and lack of those at different levels. In addition, these datasets usually only include attributes such as essay score and topic, lack of diversified attributes such as essay classification and student grade, which are helpful for grading essay. We address these by crawling the essays with multiple attributes from http://www.leleketang.com/zuowen/ to generate the dataset of essay grading on **p**rimary **a**nd **m**iddle school (PAM for short).

We collected essay data from grade 1 to grade 12. The metadata crawled includes: the title; the essay content; the essay classification, which is used to describe the style and content of the essay, and can represent the topic to a certain extent, such as narration, writing of people, writing of scenery and so on; the student grade; excellent words and sentences; the essay grade, including four grades of excellent, good, medium and poor. In order to ensure the uniformity of the essay grade in the dataset, we refer to the mapping method proposed by [16]. This method regards the student grade as the essay grade if the essay is marked medium, adds one to the student grade as the essay grade if the essay is marked good, and adds two if marked excellent, as shown in Fig. 1(a). The distribution of essays in each grade after the mapping method is shown in Fig. 1(b).



(a) Mapping relationship between student grades and essay grades.



(b) Essay amount in each essay grade.

Fig. 1. PAM dataset statistics.

3.4 Linguistics Indexes and Topic Features for Essay Grading

We adopt linguistics indexes to grade essays in the linguistics aspect. We first analyze the correlation between each index and the essay grade. Then we select several indexes from the candidates as the essay grading features in the linguistics aspect, which are more correlated with the essay grades. The candidate linguistic indexes are introduced from L2SAC [17], as shown in Table 3. Different from the grading object in [17], the research subject in this paper is Chinese essay, therefore, some index definitions are different from the original meanings, such as T-unit [18].

Table 3. Linguistics index definitions and calculation methods.

Index	Definition	Calculation method
Sent	Sentence number	The sentence number in an essay
Clause	Clause number	The clause number in an essay
MLS (MLT)	Average sentence length at word level	Word/sentence number
MLC	Average clause length at word level	Word/T-unit number
C/S (T/S)	Average clause number in sentences	Clause/sentence number
CT/T	Average T-unit complexity in T-units	T-unit complexity/number
CN/C	Average compound noun number in clauses	Compound noun/clause number
VP/T	Average verb number in T-units	Verb/T-unit number
MLCC	Average clause length at character level	Character/clause number
MLSC (MLTC)	Average sentence length at character level	Character/sentence number

Table 4. Performance comparison of different essay grading models on PAM dataset.

Model	QWK	SCC	PCC
ATT + CNN + Bi-LSTM (Ours)	**0.7503**	**0.7324**	**0.7742**
ATT + CNN + GRU	0.6379	0.6319	0.6691
ATT + Bi-LSTM	0.7199	0.7068	0.7445
CNN + Bi-LSTM	0.5095	0.5952	0.6216
ATT + CNN	0.6419	0.6341	0.6825
Bi-LSTM	0.7120	0.6972	0.6972
Bert	0.6152	0.6506	0.6629

We calculate the above indexes on the PAM dataset, and verify the Pearson, Spearman and Kendall correlation coefficients [19] between the index values and essay grades, respectively. The results are shown in Fig. 2.

Table 5. Performance comparison of different essay grading models on EGC dataset.

Model	QWK	SCC	PCC
ATT + CNN + Bi-LSTM (Ours)	**0.6339**	**0.6265**	0.6606
ATT + CNN + GRU	0.6265	0.5151	**0.6817**
ATT + Bi-LSTM	0.6067	0.5122	0.6675
CNN + Bi-LSTM	0.6129	0.4649	0.6510
Bi-LSTM	0.6042	0.4820	0.6586

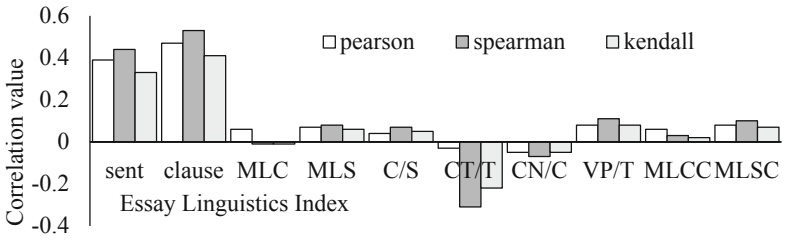


Fig. 2. Correlation coefficients between linguistics indexes and grades in PAM dataset.

Then we choose the indexes with high correlation coefficients and consistent positive and negative values as the indexes of the quantification method, including sent, clause, MLS, CT/T and MLSC. The calculation formula is:

$$\hat{y}_{e_i} = \alpha_1 sent + \alpha_2 clause + \alpha_3 MLS + \alpha_4 (CT/T) + \alpha_5 MLSC \tag{1}$$

Among them, $\hat{y}_{e_i}$ is the quantitative score calculated using multiple indexes. α is the weight corresponding to each index, equals to its corresponding Spearman correlation coefficient. In detail, $\alpha_1 = 0.448$, $\alpha_2 = 0.534$, $\alpha_3 = 0.102$, $\alpha_4 = -0.315$, $\alpha_5 = 0.114$.

Considering the essay grading in topic aspect, we train a neural network model to predicate the grades based on the topic features, as shown in Fig. 3. In order to obtain the vector representing the topic features of the essay, we use the LDA topic model to train the distribution of the essay in the implicit topic space, and use the distribution as the input of the neural network model, and finally gets the topic score of the essay after activation of the ReLU function.

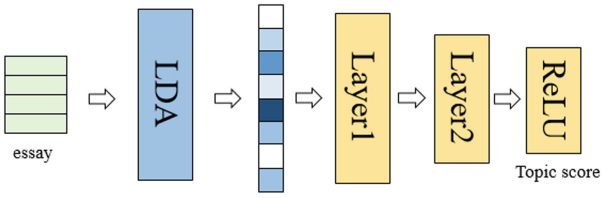


Fig. 3. Topic aspect features based essay grading model.

When the number of labeled essay is limited, we can use the linguistics indexes and topic model mentioned above to calculate the score of the unlabeled essay, thus obtain more labeled data for training the multi-aspect essay grading model, which we will introduce next.

4 The Chinese Essay Grading Method Based on Expert Knowledge

We propose a grading model based on the metrics mentioned in Sect. 3.2, shown in Fig. 4, which is divided into three parts: topic, structure, and linguistics, represented by the yellow, orange and green parts respectively. We will introduce them in detail in the following.

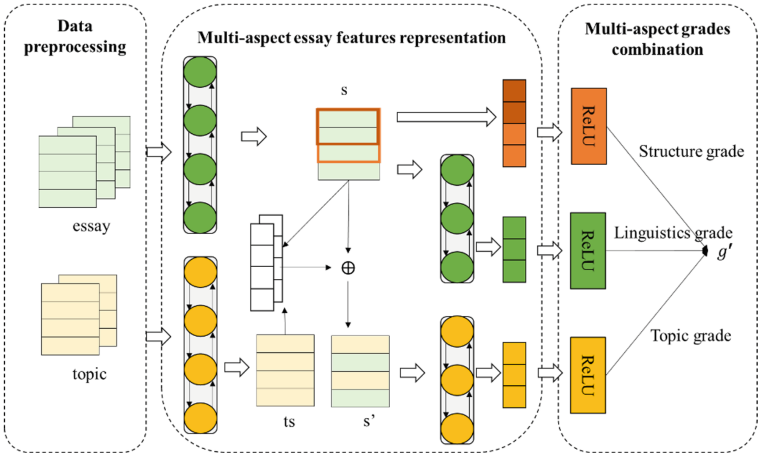


Fig. 4. The essay grading model based on multi-aspect expert knowledge.

4.1 The Topic Part of Essay Grading Model

Considering that the pretrained word vector contains abundant semantic and syntactic information, which is beneficial to the model. We introduce the pretrained vector to generate the lookup table $W^{|V|\times d}$ in the embedding layer of the sentence encoder, where |V| is the size of vocabulary V, d is the embedding dimension. The input is an essay $x = s_1, s_2, \ldots, s_{n_{sent}}$, where s_i represents the i-th sentence of the essay, n_{sent} is the number of sentences. s_i can be represented as a sequence of word embeddings $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_{n_{word}}$, where $\boldsymbol{e}_k$ represents the embedding of the k-th word in the sentence, n_{word} is the number of words in s_i.

We use the bidirectional long and short-term memory network (Bi-LSTM [20]) as the text encoder, obtain the sentence and document vectors constructing the sentence-level and document-level Bi-LSTM respectively. The word sequence passed through the embedding layer, enters a forward and a backward LSTM layer respectively to obtain

hidden states $\overrightarrow{\boldsymbol{h}}_i$ and $\overleftarrow{\boldsymbol{h}}_i$. Combine them to get the hidden state $\boldsymbol{h}_i$ of the i-th word in the sentence. For the j-th sentence, we use the hidden state $\boldsymbol{h}^j_{last}$ of the last word in the sentence as the sentence vector, denoted as $\boldsymbol{s}_j$,s shown in formulas (2–4).

$$\boldsymbol{h}^i_j = \overrightarrow{LSTM}\ (\boldsymbol{h}^j_{i-1}, e^j_i) \tag{2}$$

$$\boldsymbol{h}^i_j = LSTM\ (\boldsymbol{h}^j_{i-1}, e^j_i) \tag{3}$$

$$\boldsymbol{s}_j = \boldsymbol{h}^j_{last} = \overrightarrow{h}^j_{last} \cdot \overleftarrow{h}^j_{last} \tag{4}$$

In many cases, based on a given text material, students should choose a topic within the scope of the material and write an essay developed closely around the topic. We adopt attention mechanism [21] to take the topic consistency into account. First, we process the given material, denoted as $\mathrm{t} = ts_1, ts_2, \ldots, ts_{n_{sent}}$, $ts_j = w_1, w_2, \ldots, w_{n_{word}}$. We use sentence-level Bi-LSTM to obtain the sentence vector $\boldsymbol{ts}_j$, use attention mechanism to calculate attention value w_{ij} between $\boldsymbol{s}_i$ in the essay and $\boldsymbol{ts}_j$ in the material, finally get a new sentence vector $\boldsymbol{s}'_i$ containing topic consistency information, as shown in (5)–(7).

$$w_{ij} = \boldsymbol{ts}_j \cdot (\boldsymbol{W}_t \cdot \boldsymbol{s}_i + b) \tag{5}$$

$$\alpha_{ij} = \frac{\exp(w_{ij})}{\sum_{k=1}^{n} w_{ik}} \tag{6}$$

$$s'_i = \sum_i^n \alpha_{ij} \cdot s_i \tag{7}$$

In which, $\boldsymbol{W}_t$ is a weight matrix. We use $\boldsymbol{s}'_i$ as input to the document-level Bi-LSTM, obtain the essay vector $\boldsymbol{d}$, use $ReLU$ to predict the topic consistency score at the document level, as shown in (8):

$$g_t = ReLU\,(\boldsymbol{d}) \tag{8}$$

4.2 The Structure Part of Easy Model

The structure of an excellent essay should be clear and structured, especially argumentative essay, which should be able to put forward and gradually prove their point of view. Inspired by the paper [22], we use convolutional neural network (CNN) to extract the structure rationality on sentence level. Let $\boldsymbol{s_i} \in \mathbb{R}^{k_{sent}}$ denote each sentence embedding of the essay. All the embeddings are concatenated together to generate a 2D tensor $\boldsymbol{s}_{1:n_{sent}}$ of shape $n_{sent} \times k_{sent}$, n_{sent} denotes the number of sentences in the essay. We select different convolution kernel size $n_{ks} \in \{2, 3, 7\}$, use $\boldsymbol{W}_{ks} \in \mathbb{R}^{n_{ks} \times k_{sent}}$ denotes convolution kernel. The convolution formula is shown in (9):

$$c^j_i = f_i\left(\mathbf{W}^j_{ks} \cdot \boldsymbol{s}_{i:i+n_{ks}-1} + b\right) \tag{9}$$

where j represents the index of convolution kernel, c_i^j represents the i-th element of the 1D tensor obtained by convolution of $\boldsymbol{s}_{1:n_{sent}}$ using the j-th convolution kernel, $i \in [1, n_{sent} - n_{ks} + 1]$. We perform the maximum pooling operation on all $\boldsymbol{c}^j$, get a vector $\boldsymbol{v}_{sent}$ representing structure rationality, and use $ReLU$ function to get the structure rationality score g_s, as shown in (10), $\boldsymbol{W}_s$ is the weight matrix, and b is the bias.

$$g_s = ReLU(\boldsymbol{W}_s \cdot \boldsymbol{v}_{sent} + b) \tag{10}$$

4.3 The Linguistics Part of Essay Grading Model

The linguistics expression of the essay is also important to essay grading, which can reflect the student's ability of choice of words and building of sentences. we introduce the linguistics part, which is used to evaluate the essay in terms of linguistics expression.

We mentioned in Sect. 4.1, we obtained sentence vector $\boldsymbol{s}_j$ encoded by the sentence-level Bi-LSTM. In this part, $\boldsymbol{s}_j$ is directly input into a document-level Bi-LSTM, and the linguistics expression vector $\boldsymbol{d}'$ is obtained. We use formula (11) to calculate the linguistics score:

$$g_e = ReLU\left(\boldsymbol{d}'\right) \tag{11}$$

In addition, we introduce the interpretability algorithm proposed in the paper [16] to extract excellent words and sentences in the essay, as shown in formula (12).

$$\mathrm{P}\left(r_{m_j}^j = 1 | \boldsymbol{h}_{m_j}^j, \alpha_{m_j}^j, \boldsymbol{s}_j, \beta_j, \boldsymbol{d}\right) = \sigma(\boldsymbol{W}_1 \boldsymbol{h}_{m_j}^j + w_2 \alpha_{m_j}^j + \boldsymbol{h}_{m_j}^{j\,T} \boldsymbol{W}_3 \boldsymbol{s}_j + w_4 \beta_j + \boldsymbol{s}_j^T \boldsymbol{W}_5 \boldsymbol{d} + b) \tag{12}$$

For a sentence s_j, m_j is the word index with the largest attention value in s_j, $\alpha_{m_j}^j$ is the corresponding attention value, $\boldsymbol{h}_{m_j}^j$ is the word's hidden state. β_j is the attention value calculated between s_j and the essay, $\boldsymbol{s}_j$ is the sentence vector. $\boldsymbol{d}$ is essay vector.$\boldsymbol{W}_1 \boldsymbol{h}_{m_j}^j$ denotes word content information. $w_2 \alpha_{m_j}^j + \boldsymbol{h}_{m_j}^{j\,T} \boldsymbol{W}_3 \boldsymbol{h}_j$ denotes the importance of word for s_j. $w_4 \beta_j + \boldsymbol{h}_j^T \boldsymbol{W}_5 \mathbf{d}$ denotes the importance of s_j for the essay. Judge whether the word or sentence is excellent from the importance of it to the essay, r_i^j denotes whether the i-th word in s_j is extracted or not.

4.4 Multi-aspect Essay Grading and Optimization Objective

Given an essay x, the output of the model is a set of scores $g = \{g_{\mathrm{t}}, g_{\mathrm{s}}, g_{\mathrm{e}}, g'\}$, g_{t} represents the topic consistency score, g_{s} represents the structure rationality score, g_{e} represents the linguistics proficiency score, g' represents the final score. After scoring from three aspects respectively, we get the g' by combine them linearly as shown in formula (13).

$$g' = \alpha_1 g_t + \alpha_2 g_s + \alpha_3 g_e + b \tag{13}$$

In terms of the loss function, the final loss function is the sum of the cross-entropy loss of excellent words and sentences extraction and the mean square error loss of the score, as shown in formula (14).

$$J(\theta) = \text{CrossEntropy}\left(r, r^{'}\right) + \text{MSE}(g, g^{'}) \tag{14}$$

5 Experiments

5.1 The Dataset of Essay Grading on College and Experimental Settings

In addition to the PAM dataset mentioned in Sect. 3.3, we also conduct experiments on the same cognitive level Chinese **e**ssay **g**rading dataset from some **c**ollege (EGC for short). The dataset contains the essay attributes and the grading information, gives material and requires students to write essays according to the material. The EGC dataset contains 49,642 essays, and the scope of the grades is [0,14]. The essay grade distribution is shown in Fig. 5. There is no essay with the score of 13.5 or 14, and the highest is 13. The average grade in EGC is 9.28, and the variance is 5.80.

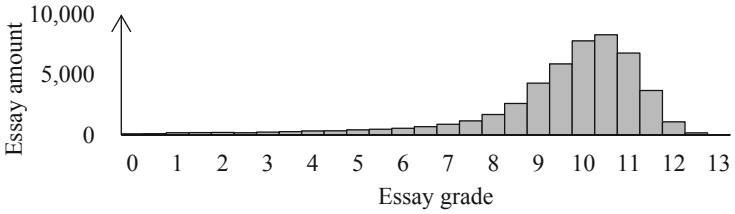


Fig. 5. The EGC dataset statistics.

We set both the maximum number of sentences and the maximum number of words to 50, fill the insufficient part with 0 and cut off the excess part of the sentence. As for the training parameters, the number of epochs is set to 15, the number of batches is set to 32, both the dimension of the sentence vector and the essay vector are 64.

5.2 Experimental Results and Analysis

We conduct experiments on PAM and EGC dataset. The model's name is a collection of encoders' short names. Variation models are used for ablation study by removing or changing some encoders. Bi-LSTM and Bert [23] are adopted as baselines. We use metrics of Quadratic Weighted Kappa (QWK for short) [24], Spearman's correlation coefficient (SCC for short) and Pearson's correlation coefficient (PCC for short). The results are shown in Table 3 and Table 4 below.

Because most of the essays in PAM dataset come from students' daily writing exercises, there are no materials given. So, we use essay's title as the topic-related material. The results shown in Table 3 indicate that our method has achieved the best results on

PAM dataset. Moreover, removing any part will cause a degradation of performance, because the model structure designed in Sect. 4 are based on the expert knowledge. Also, it can be seen that the most important parts of the model are the linguistics and topic part. We guess the reason is that, the CNN used in structure part pays more attention to the consistency of sentences. However, in reality, experts pay more attention to whether the context can be logically connected. It can't extract logical relations well.

We also conduct experiments on EGC dataset to explore whether our model can be applied to exam scenarios, the results are shown in Table 4. Since there is no information about excellent words and sentences in EGC dataset, the first term of loss function in formula (14) is abandoned. It can be seen that the performance of the model significantly reduced, but our method is still the best. According to the above results, we find that our model can be used in exam scenarios (Table 6).

Table 6. Performance comparison of using given material to learn topic features on EGC dataset.

Model	QWK	SCC	PCC
ATT + CNN + Bi-LSTM (Ours)	**0.6171**	0.4756	0.6584
ATT + CNN + GRU	0.6309	**0.4948**	**0.6783**
ATT + Bi-LSTM	0.5365	0.4671	0.4671
CNN + Bi-LSTM	0.5390	0.4858	0.6311
Bi-LSTM	0.6099	0.4910	0.6628

We continue to conduct experiments on EGC dataset using the given material as input to the topic part rather than the essay's title, the results are shown in Table 5. Comparing Table 5 and Table 4, we found that using text materials as the topic consistency content is slightly worse than using the title. The possible reason may be that, the given material usually contains multiple implicit topics. Students often focus on a certain topic when writing. Therefore, integrating the whole material into the text vector may cause the essay has poor correlations with other topics, cause the poor topic consistency score. In addition, the material in EGC dataset is classical Chinese, some words are not included in the pretrained dictionary. So, some features are lost when encoding the material.

6 Conclusion

Based on expert knowledge, we integrate existing essay grading standards, propose metrics from three aspects: topic, structure, and linguistics. Based on these metrics, we first give a quantification grading method using a set of indexes and topic model for low resource situation. Second, we propose a multi-aspect essay grading model. The model uses Bi-LSTM as encoder to generate text vectors, and extracts features from topic, structure and linguistics to grading the essay. We designed complete experiments to verify the effectiveness of our model on PAM and EGC dataset. In addition, we designed experiments to analyze the influence of different topic materials.

References

1. Ministry of Education the People's Republic of China: Compulsory Education Chinese Curriculum Standard. People's Education Press, Beijing (2011). ISBN: 9787303133178
2. General College Admissions Unified National Examination Outline in 2019. Higher Education Press, National Education Examination Authority (2018)
3. Page, E.B.: Grading essays by computer: progress report. In: Proceedings of the Invitational Conference on Testing Problems (1967)
4. Foltz, P.W., Laham, D., Landauer, T.K.: The intelligent essay assessor: applications to educational technology. Interact. Multim. Electr. J. Comput. Enhan. Learn. **1**(2), 939–944 (1999)
5. Burstein, J.: The E-rater® scoring engine: automated essay scoring with natural language processing. In: Shermis, M.D., Burstein, J. (eds.) Automated Essay Scoring: A Cross-Disciplinary Perspective, pp. 113–121. Lawrence Erlbaum Associates, Mahwah (2003)
6. Rudner, L.M.: An evaluation of the IntelliMetric essay scoring system. J. Technol. Learn. Assess. **4**(4), 3–21 (2006)
7. Rudner, L.M., Liang, T.: Automated essay scoring using Bayes' theorem. J. Technol. Learn. Assess. **1**(2), 1–22 (2002)
8. Larkey, L S.: A text categorization approach to automated essay grading. Automated Essay Scoring: A Cross-Disciplinary Perspective, 55–70 (2002)
9. Larkey L.S.: Automatic essay grading using text categorization techniques. In: 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 90–95. ACM, Melbourne (1998)
10. Dasgupta, T., Naskar, A., Dey, L., et al.: Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In: 5th Meeting of the Association for Computational Linguistics, pp. 93–102. ACM, Melbourne (2018)
11. Dong, F., Zhang, Y., Yang, J.: Attention-based recurrent convolutional neural network for automatic essay scoring. In: 21st Conference on Computational Natural Language Learning, pp. 153–162. Association for Computational Linguistics, Vancouver (2017)
12. Wang, Y., Wei, Z., Zhou, Y., Huang, X.: Automatic essay scoring incorporating rating schema via reinforcement learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 791–797. Association for Computational Linguistics, Brussels (2018)
13. 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 Press, New Orleans (2018)
14. Alikaniotis, D., Yannakoudakis, H., Rei, M., et al.: Automatic text scoring using neural networks. In: the 54th Annual Meeting of the Association for Computational Linguistics, pp. 715–725. The Association for Computer Linguistics, Berlin (2016)
15. Jin, C., He, B., Hui, K., et al.: TDNN: a two-stage deep neural network for prompt-independent automated essay scoring. In: 56th Annual Meeting of the Association for Computational Linguistics, pp. 1088–1097. The Association for Computer Linguistics, Melbourne (2018)
16. Yifei, G.: Explainable essay grading method based on expert knowledge. Shandong University (2020)
17. Lu, X.: Automatic measurement of syntactic complexity in child language acquisition. Int. J. Corpus Linguist. **14**(1), 3–28 (2009)
18. Na, H.: A corpus-based study on the syntactic features of primary school students' compositions. Shanghai Normal University (2014)
19. Jäntschi, L., et al.: Pearson versus Spearman, Kendall's Tau correlation analysis on structure-activity relationships of biologic active compounds (2005)

20. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Comput. **9**(8), 1735–1780 (1997)
21. Wang, Yequan, et al. "Attention-Based LSTM for Aspect-Level Sentiment Classification." Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 606–615
22. Kim Y.: Convolutional neural networks for sentence classification[C]. In: Empirical Methods in Natural Language Processing, pp. 1746–1751. The Association for Computer Linguistics, Doha (2014)
23. 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
24. Cohen, J.: Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. Psychol. Bull. **70**(4), 213–220 (1968)