Investigating the Validity and Reliability of a Comprehensive Essay Evaluation Model of Integrating Manual Feedback and Intelligent Assistance

https://doi.org/10.3991/ijet.v18i04.38241

Jiangyong Zhao[1], Yanwei Li[2(✉)], Wei Feng[2]
[1]Department of Public Foreign Language, Shijiazhuang University of Applied Technology, Shijiazhuang, China
[2]School of Foreign Languages, Shijiazhuang University, Shijiazhuang, China
1101658@sjzc.edu.cn

Abstract—How to respond effectively and efficiently to students' writing and to maximize the potential of feedback to promote student writing skills deserves careful consideration. The use of intelligent algorithms to assist teachers in the manual evaluation of students' essays is of practical value and importance in the context of "artificial intelligence + education". The existing intelligent evaluation techniques are subject to the interference of many factors such as openness of questions and students' language expression abilities. For this reason, this study conducts a study on comprehensive essay evaluation method with intelligent assistance and manual feedback and on reliability and validity tests. Before the intelligent evaluation, the semantic integrity of students' essays is analyzed, and a semantic integrity analysis model of students' essays based on *BERT* model is constructed. A fusion similarity algorithm for essay answer key points is proposed by extracting these characteristics that have an impact on the evaluation results, such as technique preferences, paragraph content and paragraph topic of the essays. The *Siamese* and *ESIM* networks are combined to propose an intelligent evaluation model for students' essays, and the model framework and working principle are described in detail. The experimental results verify the effectiveness of the constructed model.

Keywords—intelligent assisted evaluation of essay, comprehensive evaluation, reliability and validity test

1 Introduction

In China, English is the most popular foreign language and students spend a lot of time and effort learning English writing, but many of them still fail to write satisfactory English essays [1–4]. Chinese English teachers are heavily involved in evaluating and giving feedback to students' essays due to the large size of their classrooms [5–11]. The evaluation of essays has nothing to do with the order of the scoring points of the answers, rather, it is more subjective and flexible than the evaluation of other objective questions, thus more difficult to finish [12–16]. How to respond effectively and

efficiently to students' writing and maximize the potential of feedback to promote the improvement of students' writing ability deserves careful consideration [17–19]. This pre-school study gave us a better understanding of how to better use feedback to help students revise their text and thus to improve their writing skills. Therefore, the use of intelligent algorithms to assist teachers in the manual evaluation of students' essays is of practical value and importance in the context of "artificial intelligence + education".

Automatic evaluation systems are becoming more common in English writing and are receiving increased attention. Wang and Huang [20] explores the effects of field cognitive styles and automatic evaluation systems on college writing training. From the perspective of cognitive style differences, the application strategies of automatic evaluation systems in college English writing and teaching are summarized to better achieve the integration of information technology and subject teaching, thus improving students' English writing ability and proficiency. In order to solve the problem of intelligent evaluation of English writing, Wang and Liu [21] proposes an intelligent evaluation method of English writing based on English semantic neural network algorithm. Firstly, it briefly analyzes the research background of English semantic analysis system, describes the techniques related to English distance similarity algorithm, semantic analysis intelligent algorithm structure, word analysis algorithm, sentence lexical analysis algorithm, utterance semantic analysis algorithm and neural network algorithm and finally expounds the database and methods of the English semantic analysis system, providing a guarantee for designing English semantic analysis system. Li [22] applies a combination of teachers' feedback and automatic feedback to high school English writing instruction and explores whether the combination of the two could tackle the drawbacks of the automatic writing evaluation system and solve the traditional problem of evaluation only by teachers. The text feedback system is introduced into the context information which is used to filter the difference between the active contexts, to further reduce the number of participants in collaborative filtering and improve the real-time computational efficiency of the algorithm. Combining with the characteristics of college students' English writing and the evaluation standard reference [23], an efficient evaluation platform is designed by using artificial intelligence technology, which requires the use of convolutional neural networks (CNN) and long short-term memory (LSTM) neural networks to extract features of grammatical, syntactic, and emotional expressions in college students' English writing. In order to better perceive the language sense in English essays and improve the rationality of intelligent marking [24], a quantification method of N-ary language sense value based on correlation analysis and a fitting algorithm of English essay scoring based on rationality enhancement are proposed. The quantification of perceptual value is done by obtaining several components of the essay and calculating their support in the corpus. In addition, word features, sentence features, and chapter structure features are extracted from the papers to match the scores of the English papers.

The research on essay evaluation in China started relatively late compared with that in foreign countries, and the existing intelligent evaluation techniques for essay texts with more complex essay and usage are interfered by many factors such as openness of questions and students' language expression ability, while there is a lack of research on intelligent evaluation methods fully considering the matching degree between key

answer key points and students' essays. For this reason, this study conducts a study on comprehensive essay evaluation methods with intelligent assistance and manual feedback and on reliability and validity tests. In the second chapter, the semantic integrity of students' essays is analyzed before the intelligent evaluation, and a semantic integrity analysis model of students' essays based on *BERT* model is constructed. In the third chapter, a fusion similarity algorithm for essay answer key points is proposed by extracting these characteristics that have an impact on the evaluation results, such as technique preferences, paragraph content and paragraph topic of the essays. In the fourth chapter, the *Siamese* and *ESIM* networks are combined to propose an intelligent evaluation model for student essays, and the model framework and working principle are described in detail. The experimental results verify the effectiveness of the constructed model.

2 Analysis of semantic integrity of essay sentences

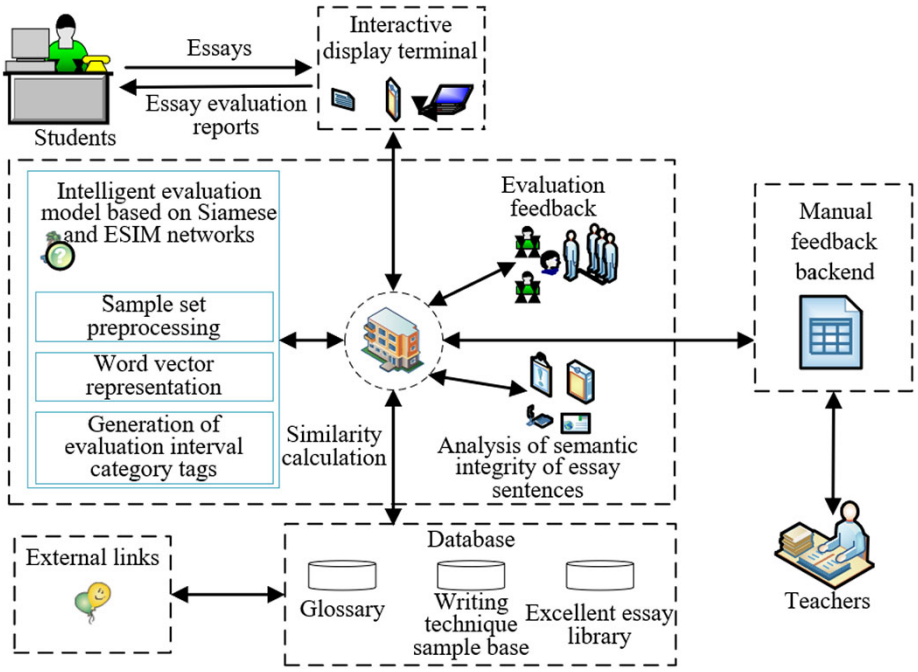

Fig. 1. Framework of intelligent essay evaluation system

To get better evaluation effects, the intelligent evaluation algorithm usually divides the long text into several short texts, calculates the similarity between the short text and the score points of the answer key points, and finally outputs evaluation results based on the similarity score sequence. If there is a sentence in a paragraph that expresses a point of view as completely as the answer key points, and it neither causes ambiguity nor has grammatical errors, then the semantics of the sentence is complete by default

in this article. Therefore, this article analyzes the semantic integrity of students' essays before conducting an intelligent evaluation.

Figure 1 shows the framework of the essay intelligent evaluation feedback system. It can be seen from the figure that the semantic integrity analysis module and evaluation feedback module of essay sentences are the most important modules of the system. In essence, the semantic integrity analysis of students' essays can be regarded as a sequence tagging problem. Traditional tagging sets tend to lead to category imbalance problems, which seriously affect the effectiveness of dividing long texts of essay paragraphs into short texts. In order to solve this problem, the essay samples can be processed by oversampling and undersampling methods. But in order to obtain better sample fitting effect, this article decides to further subdivide the tagging set from the essay samples themselves. This article constructs a semantic integrity analysis model based on *BERT* model and tags the individual characters in the essay paragraphs based on tagging set combined with part of speech, so as to obtain a more detailed and balanced tagging set of part of speech. Assuming that the beginning character of a word is represented by *Y*, the middle character of a word is represented by *M*, and the ending character of a semantically complete sentence is represented by *P*, the following formula shows the expression for tagging set *T*.

$$O = \left\{ \underbrace{Y-Xh, Y-x, \ldots, Y-c}_{39}, \underbrace{M-Xh, M-x, \ldots, M-c}_{39}, P \right\} \tag{1}$$

In this article, each character of the input student essay texts is encoded based on the *BERT* model, the fully connected layer of the model maps the encoding vectors of the characters to a predefined tagging set, and the calculation and output of the probabilities of each tag are realized by the classifier of the model. Through the model constructed in this article, the part of speech distribution sequence of students' essay texts can be obtained. Based on this sequence, this article can scientifically divide students' essay texts to ensure that the divided short texts are composed of multiple semantically complete sentences.

In the semantic integrity analysis model constructed, the presentation layer of a semantic integrity analysis model is a *BERT* model composed of multi-layer *Transformer* encoders. The input and output are the vector representation of students' essay texts and the deep bidirectional representation of each character of the texts, respectively.

The model is set up with a prediction layer consisting of a fully connected neural network and a function classifier *Softmax*, which can realize the mapping from the vector output of the representation layer to the tagging set and the output of the tags corresponding to each character of the students' essay texts. Assuming that the output value of the *i*-th node is represented by c_i and the number of categories is represented by *D*, the following equation gives the expression of the classification function:

$$Soft\max(c_i) = \frac{p^{c_i}}{\sum_{j=1}^{D} p^{c_i}} \tag{2}$$

The model uses a cross-entropy loss function as the objective function. Assuming that the total number of samples is represented by M, the number of categories is represented by D, the predicted probability by sample m for category d is represented by e_{md}, and the matching status is denoted by b_{md}. The following formula gives the calculation formula for this function in the case of multiple classifications:

$$K = -\frac{1}{M}\sum_{m=1}^{M}\sum_{d=1}^{D} b_{md}\log(e_{md}) \tag{3}$$

3 Calculation of similarity of essay answer key points

To solve the problem of temporal changes in students' essay preferences based on a topic model, this article proposes a fusion similarity algorithm targeted the essay answer key points by extracting the characteristics that will have an impact on the evaluation results, such as technique preferences, paragraph content and paragraph topic of the essays. After studying the process of teachers' manual evaluation and semantic integrity analysis of essay sentences, the general steps of calculating the similarity of essay answer key points are obtained. Figure 2 shows the calculation flow of similarity of essay answer key points.

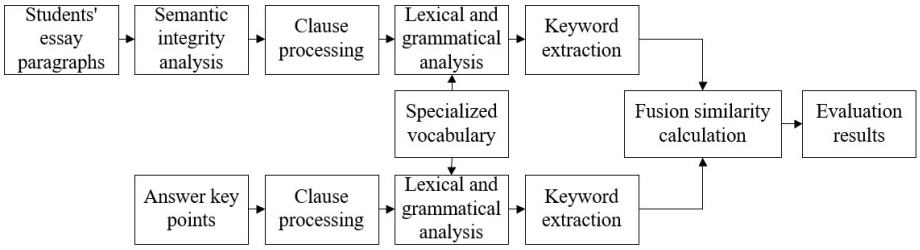

Fig. 2. Calculation flow of similarity of essay answer key points

In this article, *LDa* prototype algorithm is used to study the sentences in students' essays, and the descriptive technique vectors can be used to predict students' writing preferences. Assuming that the set corresponding to the descriptive techniques of students' essay sentences is represented by *CTP*, the following formula gives the similarity calculation formula of students' essay technique preferences:

$$JA(A,B) = \frac{\left|CTP(A)\bigcap CTP(B)\right|}{\left|CTP(A)\bigcup CTP(B)\right|} \tag{4}$$

Assuming that the descriptive technique vector of student essay content text A is represented by A*, and the descriptive technique vector of student essay content text B is represented by B*, the formula for calculating the similarity between content A and content B of students' essays is given by the following formula:

$$CIS(A,B)=\frac{A^{*}B^{*}}{\left|A^{*}\right|\times\left|B^{*}\right|} \tag{5}$$

Assuming that the set of topic entities corresponding to the content of students' essays is represented by *EN*, the following equation gives the formula for calculating the similarity of the topics of passages in students' essays:

$$JAEN(A,B)=\frac{\left|EN(A)\bigcap EN(B)\right|}{\left|EN(A)\bigcup EN(B)\right|} \tag{6}$$

Considering the similarity between essay technique preferences, the similarity between essay contents and the similarity between essay paragraph topics, this article integrates these three aspects and proposes an improved similarity algorithm to characterize the gap between students' essay texts and scoring answer key point through the mutual integration and interaction of these three aspects. Assuming that the weights of the three are represented by μ, ω and γ, the following equation shows the expression of the fusion algorithm:

$$SI(A,B)=\frac{\omega CIS(a,b)+\gamma JA(A,B)+\mu JAEN(A,B)}{\sqrt{\omega^{2}+\gamma^{2}+\mu^{2}}} \tag{7}$$

From the above equation, if only $\omega = 1$ among the three weights, it means that the algorithm only considers the content information of students' essays; if only $\mu = 1$ among the three weights, it means that the algorithm only considers the technique preferences of students' essays; if only $\gamma = 1$ among the three weights, it means that the algorithm only considers the topic information of students' essays. In view of this, this article sets all the three weight values to be 1, that is, considering the similarities of technique preferences, paragraph content and paragraph topic of the essays at the same time. Finally, based on the results of similarity calculation, the evaluation results of essays are given.

4 Implementation of intelligent evaluation of students' essays

To realize the intelligent evaluation of students' essays, this article puts forward an intelligent evaluation model for students by combining *Siamese* and *ESIM* network, including five layers: input layer, embedding layer, convolution layer, interaction layer and prediction layer.

The intelligent evaluation model for student essays sets up the input layer for preprocessing the sample sets of students' essays. Since this article studies the intelligent evaluation task of students' essays based on single character *Siames*e and *ESIM* networks, a dictionary containing all characters needs to be constructed before performing the intelligent evaluation task. As for students' essay paragraphs to be evaluated, the input of *Siamese* network model is the vector splicing representation of "tagging set of students' essay sentence categories-students' essay paragraphs-answer key points" or "tag of students' essay sentence categories-answer key points-students' essay paragraphs".

Assuming that the category tag of the student's essay paragraphs and sentences is represented by k_i, and the dictionary subscripts corresponding to the category tag of sentences, the answer key points, and the word at the current position i of the student's essay paragraphs are represented by w_i, s_i, and x_i, respectively. The lengths of the sentence category tag, answer key points, and student's essay paragraphs are denoted by m_w, m_s, and m_x, respectively, and the vector splicing operator is denoted by [ ], with the former vector splicing as:

$$A_S = \begin{bmatrix} k_1, k_2, ..., k_{m_w}, k_{m_w+1}, k_{m_w+2}, ..., k_{m_w+m_s} \\ w_1, w_2, ..., w_{m_w}; s_1, s_2, ..., s_{m_s} \end{bmatrix} \in \Re^{2\times(m_w+m_s)} \tag{8}$$

The vector splicing of the latter is:

$$A_X = \begin{bmatrix} k_1, k_2, \ldots, k_{m_w}, k_{m_w+1}, k_{m_w+2}, \ldots, k_{m_w+m_x} \\ w_1, w_2, \ldots, w_{m_w}; x_1, x_2, \ldots, x_{m_x} \end{bmatrix} \in \Re^{2\times(m_w+m_x)} \tag{9}$$

The model setting embedding layer is mainly used to represent A_{PS} and A_{PX} through *Word2vec* tool for the word vectors of the input short texts of students' essays A_S with the sentences scoring answer key points A_X. Assuming that the dimension of the word vector is represented by p, the function *Embed* is represented by $\Gamma(\ )$, the calculation formula is as follows:

$$A_{PS} = \Gamma(A_S) \in \Re^{2p\times(m_w+m_s)} \tag{10}$$

$$A_{PX} = \Gamma(A_X) \in \Re^{2p\times(m_w+m_x)} \tag{11}$$

The model sets up the convolution layer mainly for convolutional operations on A_{PS} and A_{PX} to further obtain the corresponding shallow semantic feature representations of A_{DS} and A_{DX} of the paragraphs, words and characters of students' essays. Assuming that the convolution operation is represented by $\Omega(\ )$, the number of convolution kernels is represented by d, the calculation formulas for both are as follows:

$$A_{DS} = \mathrm{ReLU}[\Omega(A_{PS})] \in \Re^{d\times(m_w+m_s)} \tag{12}$$

$$A_{DX} = \mathrm{ReLU}[\Omega(A_{PX})] \in \Re^{d\times(m_w+m_x)} \tag{13}$$

The model sets up the interaction layer as an *ESIM* network, which is mainly used to receive and process the A_{DS} and A_{DX} from the convolution layer. The shallow semantic feature representation successively passes through the double-layer *LSTM* encoding layer, the attention mechanism interaction layer, the double-layer *LSTM* synthesis layer, and then is input to the average pooling layer as well as the maximum pooling layer for processing, and the output is a fixed-length vector u considering the overall semantic relevance between A_S and A_X. Assuming that the multiplicity of dimension expansion after the interaction layer is represented by h and the dimension of the double-layer

LSTM coding layer in the interaction layer is represented by *f*, the formula for *u* is shown as follows:

$$u = ESIM[A_{DS}, A_{DX}] \in R^{hf} \tag{14}$$

The model sets up the prediction layer mainly used to convert *u* into category tag based on functions *Softmax* and *argmax*. Assuming that the function *Drop* is represented by $\Lambda(\ )$ and the dimension of the hidden layer in the prediction layer is represented by *n*, $Q_{N1} \in R^{n \times hf}$, $y_{N1} \in R^n$, $Q_{N2} \in R^{c \times n}$, $y_{N2} \in R^c$, the predefined evaluation interval category tag is represented by *c*, and the final evaluation interval category is represented by *s*, the corresponding calculation formulas are given as follows:

$$a_{P1} = \Lambda(\mathrm{ReLU}(Q_{N1}u + y_{N1})) \in R^n \tag{15}$$

$$a_{P2} = Q_{N2}a_{P1} + y_{N2} \in R^c \tag{16}$$

$$a_P = \mathrm{Soft\,max}(a_{P2}) \in R^c \tag{17}$$

$$s = \arg\max(a_P) \tag{18}$$

5 Experimental results and analysis

To verify the effectiveness of the fusion similarity algorithm, this article decomposes the similarity of the answer key points of essays into 3 parts. As mentioned above, similarities consist of similarities in essay technique preferences, essay content and essay paragraph topic. The comparison experiments are set up with three similarity combinations: *JA+CIS*, *JAEN+CIS*, JA+*JAEN+CIS*. *JA+CIS* model trains students' essay samples and calculates the internal feature information of essay samples. The evaluation performance results for different similarity combinations are given in Table 1. From the table, it can be seen that the combined similarity calculation method of *JA+JAEN+CIS* simultaneously considers the similarity of technique preferences, paragraph content and paragraph topic of the essays to obtain the optimal essay evaluation accuracy, which verifies the effectiveness of the fusion similarity algorithm proposed in this article for student essay evaluation.

Table 1. Evaluation performance results for different similarity combinations

Similarity Combinations		***JA+CIS***	***JAEN+CIS***	***JA+JAEN+CIS***
Sample set number	1	0.56	0.71	0.85
	2	0.42	0.61	0.71
	3	0.592	0.75	0.75
	4	0.61	0.85	0.81
	5	0.85	0.81	0.87
	6	0.71	0.84	0.86
	7	0.42	0.87	0.80
	8	0.46	0.62	0.79

To further verify the effectiveness of the proposed intelligent evaluation model for students' essays based on single character-based *Siamese* and *ESIM* networks, this article sets up comparison experiments, and adds the improved *LSTM* model, *SKIPFLOW* model, *MN* model, and traditional *LSTM* model as comparison objects. The experimental results are shown in Table 2. The average *kappa* of the model used in this article is about 5% higher than that of the other two scoring models. The verification results of independent sample t-test show that there is a significant difference in statistics. The model in this article has significantly improved the evaluation accuracy of students' essays. At the same time, it is proved that adding semantic integrity analysis and fusion similarity calculation plays a very important role in intelligent evaluation of students' essays.

Table 2. Simulation results of different evaluation models

Evaluation Model		**Proposed Model**	**Improved *LSTM***	***SKIPFLOW***	***MN***	**Traditional *LSTM***
Sample set number	1	0.85	0.82	0.86	0.89	0.74
	2	0.5	0.78	0.64	0.75	0.61
	3	0.62	0.71	0.61	0.71	0.69
	4	0.74	0.85	0.84	0.86	0.84
	5	0.71	0.81	0.89	0.81	0.81
	6	0.79	0.83	0.82	0.84	0.83
	7	0.73	0.88	0.89	0.75	0.81
	8	0.61	0.75	0.71	0.61	0.54

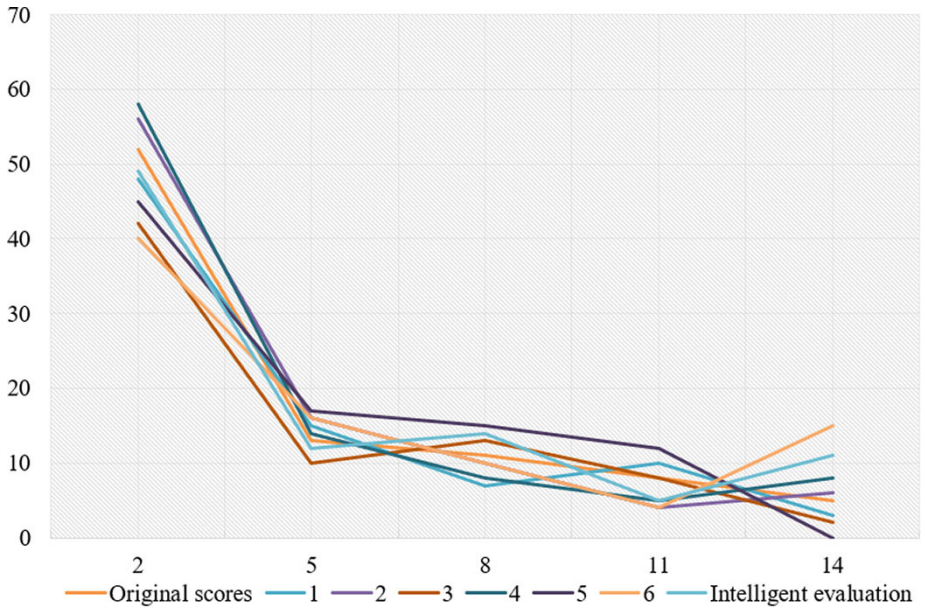

Fig. 3. Evaluation results of different evaluation methods and 6 evaluation teachers

Figure 3 shows the evaluation results of different evaluation methods and 6 evaluation teachers. It can be seen from the figure that there are great differences in the marking standards of different evaluation methods, and in the evaluation results of different evaluation teachers. Table 3 shows the comparison between the average scores of the teachers' manual evaluation and the scores of the intelligent evaluation model. This article conducts linear regression analysis to obtain the R^2 coefficient of 0.9213 and the correlation of the difference of mean scores of 1.09 and 0.924, which shows that the subjective evaluation of manual evaluation has a greater impact on the evaluation results of students' essays, and it is necessary to conduct intelligent evaluation model for assistance, and the evaluation scores obtained are more scientific and objective. Table 3 shows different marking methods and the marking examples of six evaluation teachers.

Table 3. Examples of evaluation results of different evaluation methods and 6 evaluation teachers

	Original Scores	**1**	**2**	**3**	**4**	**5**	**6**	**Machine Scores**
Sample 1	8	2	7	6	10	5	2	4
Sample 2	3	15	7	9	11	5	17	12
Sample 1	*For another, further studies require a large sum of money. This creates an especially heavy burden on students from poor backgrounds. By contrast, working right after graduation enables them to become economically independent.*							
Sample 2	*Since you will live and study in a totally different environment, I would like to make several practical suggestions with regard to your life in our university. First of all, you'd better learn as much Chinese as possible to get along better with Chinese students and teachers.*							

Table 4. The coefficient *Cronbach's a* for each dimension of the intelligent evaluation model

Names	**Number of Items in Evaluation Rules**	**The Coefficient *Cronbach's a***
The fifth-class essays	7	0.625
The fourth-class essays	8	0.847
The third-class essays	8	0.714
The second-class essays	10	0.729
The first-class essays	12	0.715

The standard answer of the intelligent evaluation model consists of 33 key points, and divide the essays into 5 dimensions as follows: the fifth-class essay (3 key points) the fourth-class essay (4 key points), the third-class essay (5 key points), the second-class essay (5 key points) and the first-class essay (7 key points). The internal consistency reliability of the five dimensions is evaluated by calculating coefficient *Cronbach's a*. The calculation results are given in Table 4. It can be seen from the table that the coefficient *Cronbach's a* of the key points standard of the intelligent evaluation model is 0.865, and the internal consistency coefficients of evaluation rules with less than 10 items are all greater than 0.6, and those of the others above 0.7, which verifies that the total reliability coefficient of the intelligent evaluation model is higher.

The article then proceeds to analyze the structural validity of the five dimensions of the answer key point standards. According to the factor load table of the intelligent evaluation model shown in Table 5, it can be seen that the absolute load of the intelligent evaluation model on specific evaluation rules is greater than or equal to 0.28. The results of factor analysis show that the constructed intelligent evaluation model basically conforms to the design principles and design ideas.

Table 5. Factor loads for the total sample of students' essays

Evaluation Rules	*est,std*	*se*	*z*	*P Value*	*ci.lower*	*ci.upper*
The fifth-class essay-1	0.507	0.063	12.369	0.001	0.425	0.627
The fifth-class essay-2	0.417	0.085	10274	0.003	0.469	0.512
The fifth-class essay-3	0.528	0.028	12.241	0.003	0.427	0.697
The fourth-class essay-1	0.501	0.076	10.326	0.017	0.495	0.634
The fourth-class essay-2	0.327	0.027	7.514	0.002	0.263	0.347
The fourth-class essay-3	0.418	0.062	14.295	0.001	0.328	0.517
The fourth-class essay-4	0.324	0.015	8.296	0.027	0.201	0.439
The third-class essay-1	0.281	0.036	5.269	0.001	0.152	0.247
The third-class essay-2	0.625	0.027	13.205	0.036	0.374	0.469
The third-class essay-3	0.428	0.062	15.269	0.024	0.369	0.427
The third-class essay-4	0.674	0.025	16.058	0.009	0.527	0.769
The third-class essay-5	0.538	0.067	11.274	0.032	0.421	0.692
The second-class essay-1	0.431	0.025	16.528	0.024	0.436	0.528
The second-class essay-2	0.537	0.051	13.295	0.071	0.418	0.631
The second-class essay-3	0.425	0.014	11.352	0.029	0.425	0.516
The second-class essay-4	0.683	0.041	18.629	0.074	0.511	0.692
The second-class essay-5	0.502	0.047	16.382	0.001	0.526	0.629
The first-class essay-1	0.638	0.011	19.528	0.003	0.625	0.745
The first-class essay-2	0.403	0.025	11.627	0.021	0.392	0.528
The first-class essay-3	0.251	0.063	6.274	0.084	0.169	0.315
The first-class essay-4	0.528	0.041	19.325	0.017	0.462	0.627
The first-class essay-5	0.625	0.027	25.163	0.002	0.584	0.692
The first-class essay-6	0.618	0.014	26.328	0.014	0.528	0.614
The first-class essay-7	0.784	0.095	22.514	0.038	0.641	0.759

The validity of the intelligent evaluation model answer key point standard is examined using the teacher's evaluation dimension in the manual evaluation process as the validity criterion. Table 6 shows the correlation between manual evaluation and intelligent evaluation model. As can be seen from the table, all the dimensions of the model evaluation standard are positively correlated with the manual evaluation standard, and the positive correlation between the four dimensions of the fifth-class essay, the fourth-class essay, the third-class essay, and the second-class essay and the dimensions of the constructed model evaluation criteria is statistically significant. The correlation of the

first-class essay is not significant due to the differences between the first-class essays and the dimensions of the constructed model evaluation standard. In general, the intelligent evaluation model has better validity than the manual evaluation.

Table 6. Correlation between manual evaluation and intelligent evaluation model

Manual Evaluation The Proposed Model	Internalized Behavior	Externalized Behavior
The fifth-class essay	0.358**	0.284*
The fourth-class essay	0.314*	0.269**
The third-class essay	0.014**	0.027*
The second-class essay	0.169*	0.162**
The first-class essay	0.284**	0.328*

Note: ** and * respectively represent significant parameter estimates at 5% and 10% levels.

6 Conclusions

This study conducts a study on comprehensive essay evaluation method with intelligent assistance and manual feedback and on reliability and validity tests. Before the intelligent evaluation, the semantic integrity of students' essays is analyzed, and a semantic integrity analysis model of students' essays based on *BERT* model is constructed. A fusion similarity algorithm for essay answer key points is proposed by extracting these characteristics that have an impact on the evaluation results, such as technique preferences, paragraph content and paragraph topic of the essays. The *Siamese* and *ESIM* networks are combined to propose an intelligent evaluation model for students' essays, and the model framework and working principle are described in detail. Comparison experiments are set up with the similarity of essay answer key points decomposed into three parts, and the results show the performances of different similarity combinations, which verifies the effectiveness of the fusion similarity algorithm. Comparison experiments are carried out, adding the improved *LSTM* model, *SKIPFLOW* model, *MN* model, and traditional *LSTM* model as comparison objects, and the simulation results of different evaluation models are given to further verify the effectiveness of the proposed intelligent evaluation model for students' essays in this article. evaluation results of different evaluation methods and six evaluation teachers are shown to verify the necessity of manual evaluation to assist the intelligent evaluation model, and examples of evaluations are given. Finally, the internal consistency reliability check, structural validity analysis, and correlation analysis are conducted on the evaluation of the essays divided into 5 classes of the intelligent evaluation model. It can be seen that the intelligent evaluation model has better reliability coefficient, structural validity, and validity scale validity than manual evaluation.

7 Acknowledgement

This paper is supported by the Humanities and Social Science Project of Colleges and Universities in Hebei Province (Grant No.: SZ2022001).

8 References

[1] Wang, Y., Feng, Y. (2022). Developing junior high school students' English writing skills through reading and creating English picture books: An experimental study. In Proceedings of the 5th International Conference on Big Data and Education, Shanghai China, pp. 205–211. https://doi.org/10.1145/3524383.3524417

[2] Zeng, G. (2022). Intelligent test algorithm for English writing using English semantic and neural networks. Mobile Information Systems, 2022: 4484201. https://doi.org/10.1155/2022/4484201

[3] Xia, J., Liu, H., Liu, W. (2021). AI-based IWrite assisted English writing teaching. In International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy, Shanghai, China, pp. 158–165. https://doi.org/10.1007/978-3-030-89511-2_19

[4] Wang, Y. (2019). A study of applying automated assessment in teaching college English writing based on juku correction network. International Journal of Emerging Technologies in Learning, 14(11): 19–31. https://doi.org/10.3991/ijet.v14i11.9411

[5] Liang, J.T. (2022). English abstract writing based on natural language generation. In 2022 International Conference on Information System, Computing and Educational Technology (ICISCET), Montreal, QC, Canada, pp. 203–206. https://doi.org/10.1109/ICISCET56785.2022.00057

[6] Bai, X. (2022). Teaching design of English writing based on UMU. Mathematical Problems in Engineering, 2022: 9075380. https://doi.org/10.1155/2022/9075380

[7] Yu, R. (2020). Reform in the teaching model of English writing in the big data era. In 2020 IEEE 2nd International Conference on Computer Science and Educational Informatization (CSEI), Xinxiang, China, pp. 252–259. https://doi.org/10.1109/CSEI50228.2020.9142506

[8] Wang, C.Y., Huang, H.F. (2020). On applying blended learning to writing for English argumentative essays for Chinese undergraduates. In Proceedings of the 2020 8th International Conference on Information and Education Technology, Okayama Japan, pp. 78–82. https://doi.org/10.1145/3395245.3396436

[9] Ran, X., Hossain, M. (2019). Recommendation algorithm based business English writing training strategy. Journal of Intelligent & Fuzzy Systems, 37(3): 3445–3452. https://doi.org/10.3233/JIFS-179148

[10] Liu, M. (2019). Design of intelligent English writing self-evaluation auxiliary system. Informatica, 43(2): 299–303. https://doi.org/10.31449/inf.v43i2.2783

[11] Gao, S. (2022). Evaluation method of writing fluency based on machine learning method. Mathematical Problems in Engineering, 2022: 1253614. https://doi.org/10.1155/2022/1253614

[12] Terreau, E., Gourru, A., Velcin, J. (2021). Writing style author embedding evaluation. In Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems, Punta Cana, Dominican Republic, pp. 84–93. https://doi.org/10.18653/v1/2021.eval4nlp-1.9

[13] Öncel, P., Flynn, L.E., Sonia, A.N., Barker, K.E., Lindsay, G.C., McClure, C.M., Allen, L.K. (2021). Automatic student writing evaluation: Investigating the impact of individual differences on source-based writing. In LAK21: 11th International Learning Analytics and Knowledge Conference, Irvine CA, USA, pp. 620–625. https://doi.org/10.1145/3448139.3448207

[14] Gao, J. (2021). Exploring the feedback quality of an automated writing evaluation system pigai. International Journal of Emerging Technologies in Learning, 16(11): 322–330. https://doi.org/10.3991/ijet.v16i11.19657

[15] Xu, S., Xu, G., Jia, P., Ding, W., Wu, Z., Liu, Z. (2021). Automatic task requirements writing evaluation via machine reading comprehension. In International Conference on Artificial Intelligence in Education, Utrecht, The Netherlands, pp. 446–458. https://doi.org/10.1007/978-3-030-78292-4_36

[16] Hsu, C.H., Chung, C.H., Venkatakrishnan, R., Venkatakrlshnan, R., Wang, Y.S., Babu, S.V. (2021). Comparative evaluation of digital writing and art in real and immersive virtual environments. In 2021 IEEE Virtual Reality and 3D User Interfaces (VR), Lisboa, Portugal, pp. 1–10. https://doi.org/10.1109/VR50410.2021.00089

[17] Sun, H. (2020). The learning method of Peer Review in College English writing course. International Journal of Emerging Technologies in Learning, 15(5): 156–170. https://doi.org/10.3991/ijet.v15i05.13775

[18] Wilson, J., Ahrendt, C., Fudge, E.A., Raiche, A., Beard, G., MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. Computers & Education, 168: 104208. https://doi.org/10.1016/j.compedu.2021.104208

[19] Zhang, X.P., Wang, S.X., Cao, Y.L., Chen, G.Q. (2020). Application of analytical hierarchy process in teaching quality analysis of English writing. International Journal of Emerging Technologies in Learning, 15(14): 137–150. https://doi.org/10.3991/ijet.v15i14.15359

[20] Wang, P., Huang, X. (2022). An online English writing evaluation system using deep learning algorithm. Mobile Information Systems, 2022: 7605989. https://doi.org/10.1155/2022/7605989

[21] Wang, J., Liu, B. (2022). Intelligent evaluation algorithm of English writing based on semantic analysis. Computational Intelligence and Neuroscience, 2022: 8955638. https://doi.org/10.1155/2022/8955638

[22] Li, J. (2022). English writing feedback based on online automatic evaluation in the era of big data. Mobile Information Systems, 2022: 9884273. https://doi.org/10.1155/2022/9884273

[23] Jian, L. (2022). Construction and application of IWrite artificial intelligence evaluation system for college English writing. Security and Communication Networks, 2022: 1511153. https://doi.org/10.1155/2022/1511153

[24] Gao, L., Xu, J., Fu, X., Li, J. (2022). Construction of intelligent evaluation platform based on random matrix IWrite college English writing. Mathematical Problems in Engineering, 2022: 6402418. https://doi.org/10.1155/2022/6402418

9 Authors

Jiangyong Zhao is an associate professor working in the Department of Public Foreign Language in Shijiazhuang University of Applied Technology, China. His research interests include English competitions and teaching methodology. Three papers are published and 3 text books are published. Email: 2003100492@sjzpt.edu.cn

Yanwei Li is a lecturer in School of Foreign Languages in Shijiazhuang University, China. He graduated from Hebei University in 2007. His research interests include computational linguistics, corpus linguistics and English teaching. He has authored 6 papers and conducted 5 research projects. Email: 1101658@sjzc.edu.cn

Wei Feng is a lecturer in School of Foreign Languages in Shijiazhuang University, China. She graduated from Hebei University in 2008. Her research interests include linguistics and English teaching. She has authored 5 papers and conducted 3 research projects. Email: 1101663@sjzc.edu.cn

Article submitted 2022-10-21. Resubmitted 2023-01-04. Final acceptance 2023-01-08. Final version published as submitted by the authors.