Hindawi
Security and Communication Networks
Volume 2023, Article ID 9804257, 1 page
https://doi.org/10.1155/2023/9804257

Retraction

Retracted: Automatic Scoring of English Essays Based on Machine Learning Technology in a Wireless Network Environment

Security and Communication Networks

Received 5 December 2023; Accepted 5 December 2023; Published 6 December 2023

This article has been retracted by Hindawi, as publisher, following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of systematic manipulation of the publication and peer-review process. We cannot, therefore, vouch for the reliability or integrity of this article.

Please note that this notice is intended solely to alert readers that the peer-review process of this article has been compromised.

Wiley and Hindawi regret that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] F. Zhang, L. Yu, and J. Shen, "Automatic Scoring of English Essays Based on Machine Learning Technology in a Wireless Network Environment," *Security and Communication Networks*, vol. 2022, Article ID 9336298, 9 pages, 2022.

Hindawi
Security and Communication Networks
Volume 2022, Article ID 9336298, 9 pages
https://doi.org/10.1155/2022/9336298

Research Article

Automatic Scoring of English Essays Based on Machine Learning Technology in a Wireless Network Environment

Fuzhuang Zhang,[1] **Lan Yu**,[1] **and Jun Shen**[2]

[1]*School of Foreign Languages, Xinyang College, Xinyang 464000, Henan, China*
[2]*School of Foreign Languages, Xinyang Normal University, Xinyang 464000, Henan, China*

Correspondence should be addressed to Jun Shen; shenjun@xynu.edu.cn

Received 22 March 2022; Revised 23 April 2022; Accepted 27 April 2022; Published 28 May 2022

Academic Editor: Muhammad Arif

The English composition is an important indicator of English learners' overall language skills and is asked in large-scale English examinations, both in China's college entrance examinations and graduate examinations and in the TOEFL, GRE, and IELTS examinations in Europe and the United States. Some automatic scoring systems for English writing have been created in the United States and internationally, however the systems still have issues with generalization, accuracy, and error correction. In this paper, we present a method to improve the accuracy of existing automatic composition scoring systems through deep learning techniques in a wireless network environment. Experiments reveal that the method can accurately assess the quality of English learners' writings, paving the way for the creation of an automated composition scoring system for large-scale machine testing and web-based self-learning platforms.

1. Introduction

Composition is an important indicator of English language learners' language ability [1, 2]. At present, in the field of English language teaching and testing, learners' essays are usually reviewed manually, which is very labor-intensive and difficult to ensure the reliability and validity of the assessment results. In order to improve this situation, scholars at home and abroad have started to use machine learning [3, 4] and natural language processing techniques [5, 6] to automatically assess the quality of learners' compositions by computer. Automated Essay Scoring (AES) systems [7] can be used in large-scale, high-impact language proficiency tests such as the TOEFL and GRE as an aid to verify the reliability of manual scoring and to reassess essay quality if there is a significant difference between the two. The AES system can also be used in a non-testing environment as a web-based self-directed learning platform, providing real-time feedback to students after they have submitted their essays, and providing dynamic assessments to urge them to revise their essays and improve their second language writing.

A wireless network is a wireless local area network (WLAN) that uses radio waves as a medium for information transmission [8]. The traditional English teaching mode is a single "human-person" face-to-face teaching mode, but the teaching activities in the wireless network environment will also produce a "human-computer" online teaching mode [9]. The wireless network will have an impact on the content, organization, location, and technique of teaching, which will be the breakthrough of single face-to-face education. In a wireless network setting, instructional approaches are more imaginative [10]. Constructivist ideals influence instruction in this environment, which primarily uses cooperative/collaborative learning, discovery learning, and independent learning methodologies. Teaching methods change from the traditional single lecture to a more active approach to student motivation. In the teaching of English composition writing, more emphasis will be placed on students identifying their own grammatical flaws and problems in their writing ideas.

This paper focuses on the characteristics of education in the "human-machine" environment of wireless network

teaching. By integrating research methods from the fields of computer science and linguistics, we use machine learning-based algorithms to extract lexical, grammatical, and discourse features of learners' texts and construct scoring models in terms of text complexity, grammatical correctness, and discourse coherence to improve the performance of existing AES systems.

The paper's organization paragraph is as follows: the related work is presented in Section 2. Section 3 analyzes the methods of the proposed work. Section 4 discusses the experiments and results. Finally, in Section 5, the research work is concluded.

2. Related Work

The automated essay scoring system is an automated computerized scoring system that assesses the quality of essays by mimicking the experience and process of human scoring, extracting relevant quantifiable features from a large number of texts that have been scored by expert teachers, and constructing scoring models to simulate the correlation between these features and the level of student writing. The performance of the scoring system is assessed by the consistency of the system scores with the human scores, the closer the system scores are to the human scores the better the performance of the system [11].

Project Essay Grade (PEG) [12] was the first AES system to be developed, at the request of the American College Board, by Ellis Page, and the first version was introduced in 1966. The main feature of PEG is that it focuses on the analysis of the surface structure of language at the expense of the content of the language, and applies mainly the principle of regression in statistics, using a number of easily quantifiable variables related to the essay as independent variables and the essay score as the dependent variable, and scoring the essay by looking at a number of quantifiable factors [13]. The PEG uses the length of the essay to predict the student's expressive ability, the number of different word forms to predict the writer's mastery of word usage, and the variation in word length to predict the writer's vocabulary. Training is done by regression analysis to obtain correlation coefficients between the variables of interest and text scoring, leading to automatic essay scoring.

IEA (Intelligent Essay Assessor) [13] was developed by Knowledge Analysis Technology, a subsidiary of the Pearson Group, in the late 1990s. IEA was the first automated essay scoring system based on latent semantic analysis, a statistical analysis technique, which uses the analysis of essay content as an important reference indicator for scoring. The basic principle of IEA is derived from Latent Semantic Analysis (LSA) [14], a statistical method developed by the psychologist Thomas Landauer, which is a statistical calculation to extract the specific meaning of words and phrases in a given context. It starts by representing the different semantic units of a composition in a high-dimensional semantic space, each semantic unit being a point in this semantic space, and the semantic similarity between two different semantic units is estimated by their relative distance in the semantic space. LSA is a complex statistical technique for knowledge acquisition and representation. It is a statistical model for statistical analysis of the semantics of words, based on the word bag theory, where all the words in a text are put together, and if one of the words changes, then the semantic information of the text will follow. The latent semantic analysis of text believes that the semantic quality of a text is determined by the words in the text, and a text word matrix is created. It is necessary to first remove the dummy words from the text, i.e., words that have no real meaning but occur frequently, because increasing or decreasing the number of these words has little effect on the semantics of the text, but increases the dimensionality of the word text vector and makes the calculation more difficult. In general, the elements of the text-word matrix are the word frequencies of the words, but in some cases, the document frequency and inverse document frequency (TF-IDF) values are used as elements of the text-word matrix [15]. The text vector representing the composition to be tested is compared with the text vector in the semantic space, and the similarity is used as a weighted vector, to sum up the ratings of the training composition, and finally, the semantic rating of the composition to be tested is obtained. The semantic score of the essay to be tested is obtained.

E-rater (Electronic Essay Rater) [16] is a scoring system being used by the Educational Testing Service in the United States, developed by Burstein et al. in the late 1990s as the first AES system to be applied to a large-scale socialized test. The theoretical core of the system is based on natural language processing techniques based on artificial intelligence, and also uses a regression algorithm similar to PEG. Natural language understanding refers to the use of statistical, machine learning, and other research methods to achieve an intuitive understanding of human language by computers, enabling unhindered communication between computers and humans. E-rater uses three main natural language processing tools: syntactic analysis, expository analysis, and thematic analysis. Syntactic analysis aims to parse the text and carefully evaluate the sentence structure, such as virtual voice and other compound phrases, in order to capture the expected types of sentences in the text. In terms of common writing language, the expository analysis identifies the relationships and organization of the different sentences in the text. Thematic analysis analyzes the use of vocabulary in the essay with the aim of assessing the content quality of the essay.

IntelliMetric^{TM} is an automated essay scoring system developed by Vantage Learning [17], whose core technology is based on artificial intelligence theory. The process of manual scoring is mastered by learning from a large number of training sets, constructing different scoring models, cross-validating the models using different test sets, and finally identifying scoring models with reliable performance for scoring essays.

The Bayesian Essay Test Scoring System (BETSY) is an automated text classification-based essay scoring system developed by Lawrence M. Runder at the University of Maryland and funded by the U.S. Department of Education [18]. BETSY integrates content and formal features into one feature set and classifies essays into four levels (excellent,

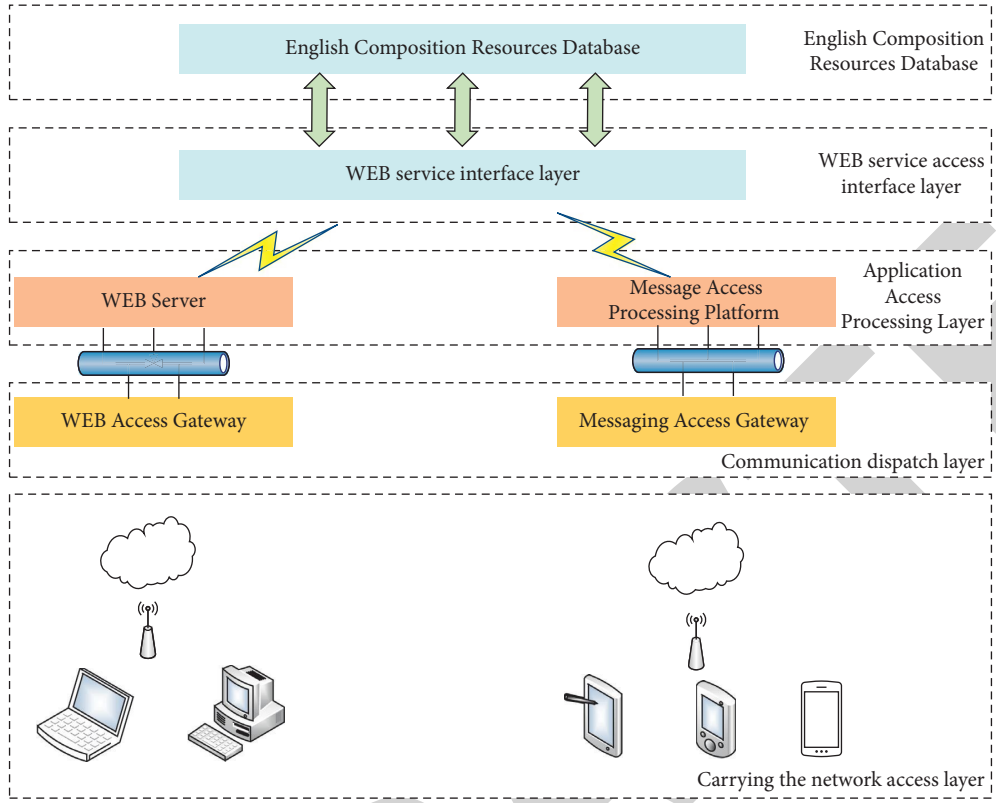

Figure 1: Web framework for a wireless English compositions correction system.

good, pass, and fail). BETSY's classification technique relies heavily on plain Bayesian theory [19]. Although computationally intensive, the Bayesian essay scoring system has the following advantages: it uses the multivariate Bernoulli model (MBM) and the Bernoulli (BM) model, which are premised on conditional assumptions. It is currently the only open-source system that is freely available to users. Not only that, but it also has a good web interface, enabling timely and effective feedback of information.

3. Methods

3.1. Wireless Network System Framework. The overall framework of the wireless network we designed for the English essay scoring system is shown in Figure 1. The system adopts a web service-oriented architecture with hierarchical processing and separation of communication processing and content provision to improve the portability, compatibility, and scalability of the system.

The system consists of five layers from the bottom up the carrier network access layer, the communication dispatch layer, the application access processing layer, the Web Service access interface layer, and the database resource layer. The carrier network access layer refers to the bearer network required for system data communication and consists of wireless communication networks such as GSM and CDMA. The communication dispatch layer enables data communication between the wireless communication network and the IP network, and thus between the system and the wireless communication network. The application processing layer handles two parts: one for access request processing, and one for data collation. The access interface layer focuses on the processing of English teaching resources, making the integrated data conform to the requirements for automatic scoring of wireless English essays by cutting and reorganizing the raw material. The teaching logic is also encapsulated to provide a complete teaching plan building block to the public.

3.2. Automatic English Composition Scoring System Solution Design. To facilitate comparison with previous research, the Cambridge FCE Composition Corpus Training and Assessment Essay Scoring System was used [20]. Figure 2 depicts the whole system, which has four instructional programmer components: data pre-processing, feature selection, model creation, and model evaluation.

3.2.1. Bag-of-Words Feature Extraction and Filtering [21]

(1) Feature Extraction. The set of all N element sequences V is first extracted from the training set, and then each composition in the training and test sets is transformed into a vector of Dimension $|V|$, where $|V|$ represents the sequence type. Assuming $V = \{v_1, v_2, \ldots, v_{|V|}\}$, the text d can be characterized as a vector $d = (c(v_1, d), c(v_2, d), \ldots, c(v_{|V|}, d))$. Where $c(v_{|V|}, d)$ is the frequency of occurrence of the sequence v in the text d. The bag-of-words feature consists of words and Wordiness of length 1~3. For example, the composition "What clothes should I take? How

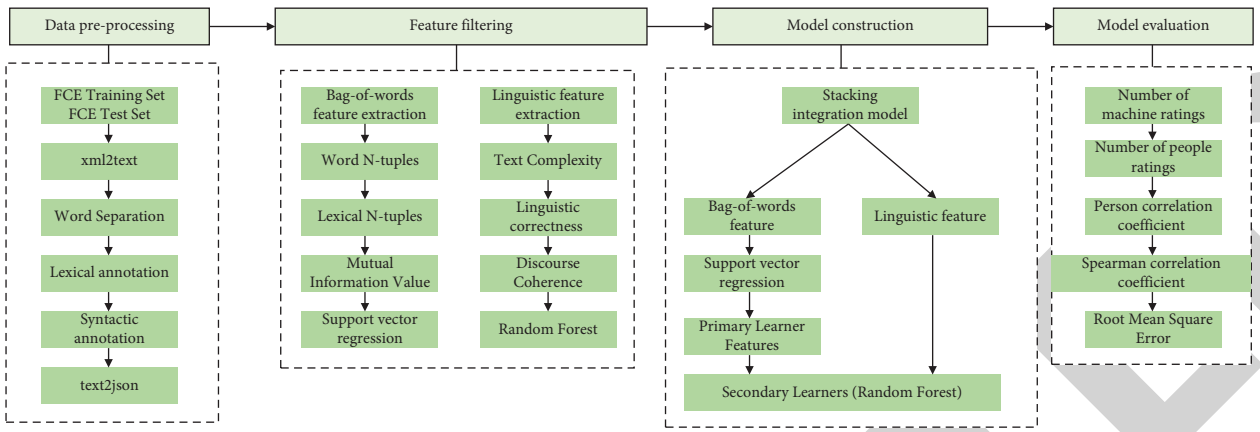

Figure 2: General framework of the AES scoring system.

Table 1: Bag-of-words feature extraction.

Word sequences	Frequency	Part of speech	Frequency
I	2	PRP	4
Should	2	VB	1
was	0	VBD	0
How much	1	MD PRP	3
Should I taken	2	MD PRP VBN	2
Should I take	0	MD PRP VB	0

much money should I take? And how could we meet at the airport?" The words and word properties contained in the essay are shown in Table 1. The part of speech assignments is PRP for pronouns, VB for verb proxemics, and MD for modal verbs.

N sequences represent fixed relationships between words. The variety and number of sequences in a composition varies from level to level, reflecting the accuracy and fluency of the learner's English. For example, the ternary sequence "MD PRP VBN" in the above example detects two cases of misuse of the modal verb "should I taken" in the composition.

(2) Feature Filtering. As shown in equation (1), the original feature set BOW is filtered by the length and mutual information of N sequences, and the feature subset $\mathrm{BOW}_{\mathrm{sub}}$ is obtained. Where: len_v is the length of the word and part of speech. t_{len} is the length threshold. MI_v is the mutual information value of the sequences. t_{mi} Is the mutual information threshold. The values of t_{len} and t_{mi} are set manually and the best values are determined based on SVR model errors.

$$\mathrm{BOW}_{\mathrm{sub}} = \{v \in \mathrm{BOW} | len_v \langle t_{\mathrm{len}} \wedge MI_v \rangle t_{\mathrm{mi}}\}. \quad (1)$$

The variety of N sequences is proportional to the length of the sequence. However, some of the sequences are only specific words that are closely related to the topic of the training essay. If they are not filtered, the generalization ability of the model in predicting essays on different topics will be reduced. The mutual information value was used to select N sequences with high differentiation, and was calculated as follows: firstly, the distribution of the sequences v in high and low scoring essays was counted, and a 2×2 column table was constructed as shown in

Table 2: List of N-dollar sequence distributions.

List of columns	$D_{\mathrm{high\_score}}$	$D_{\mathrm{low\_score}}$
v	n_{11}	n_{12}
v	n_{21}	n_{22}

Table 2. Where $D_{\mathrm{high\_score}} = \{d \in D_{\mathrm{train}} | \mathrm{score}(d) \geq m\}$, $D_{\mathrm{low\_score}} = \{d \in D_{\mathrm{train}} | \mathrm{score}(d) < m\}$. D_{train} is the training set. $\mathrm{score}(d)$ is the score of essay d. m is the median score of the training set. m is the median of the scores of essays in the training set. n_{ij} is the number of high and low scoring essays with or without a particular sequence.

Calculate the MI value of the sequence v according to the following equation.

$$MI_v = \frac{n_{11}}{n}\log_2\frac{nn_{11}}{n_{1+}n+1} + \frac{n_{21}}{n}\log_2\frac{nn_{21}}{n_{2+}n_{+1}} + \frac{n_{12}}{n}\log_2\frac{nn_{12}}{n_{1+}n_{+2}} + \frac{n_{22}}{n}\log_2\frac{nn_{22}}{n_{2+}n_{+2}}, \quad (2)$$

where: $n = n_{11} + n_{12} + n_{21} + n_{22}$ is the total number of essays in the training set. $n_{1+} = n_{11} + n_{12}$ is the number of essays containing sequence v. $n_{+1} = n_{11} + n_{12}$ is the number of high scoring essays. The Mutual Information Value (MI) measures the information gain of the sequence distribution given the text category, with higher MI values indicating a higher correlation between the sequence and the essay scores.

The SVR model requires weighting the sequence frequencies to reduce the weight of common words (e.g., get, make). As shown in equations (3) and (4), both Binary and TFIDF are used to weight the original word frequencies.

$$\mathrm{Binary}(c(v,d)) = \begin{cases} 1, & c(v,d) \geqslant 1 \\ 0, & c(v,d) < 1 \end{cases}, \quad (3)$$

$$TF - IDE(c(v,d)) = (1 + \log(c(v,d))) \times \log\left(\frac{N}{df_v}\right), \quad (4)$$

Table 3: Text surface features.

Feature number	Feature category	Feature code
1	Number of characters	LEN_CHAR
2	Number of words	LEN_WORD
3	Number of punctuation marks	LEN_PUNCT
4	Number of sentences	LEN_SENT
5	Number of paragraphs	LEN_PARA
6	Average word length	LEN_AWL
7	Average sentence length	LEN ASL

Table 4: Text readability features.

Features serial number	Characteristic categories	Feature code	Calculation method
8	Word average syllables	RE_AWS	SYL/N
9	Complex words proportion	RE_CWR	CW/N
10	FOG readability indicator	RE_FOG	0.4 ∗ (ASL + 100 ∗ CWR)
11	FLESCH readability indicator	RE_FLESCH	206.835–84.6 ∗ AWS-1.015 ∗ ASL
12	KINCAID readability indicators	RE_KINCAID	11.8 ∗ AWS + 0.39 ∗ ASL-15.59

where $c(v,d)$ is the original frequency of sequence v in composition d. $\log(N/df_v)$ is the inverse document frequency, also is the logarithm of the ratio of the total number of documents N to the number of compositions df_v containing sequence v.After weighting the training and validation set sample vectors, the SVR algorithm was used to predict the essay scores using the mean square error MSE assessment model shown in equation (5). Where n is the number of samples in the validation set, y_i and $\widehat{y}_i$ are the number of human and machine scores.

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n} \tag{5}$$

3.2.2. Linguistic Feature Extraction and Filtering. Linguistic features include six dimensions of text surface features, part of speech diversity, text readability, syntactic complexity, grammatical correctness, and discourse coherence, with a total of 28 sub-categories.

In this paper, seven types of text length-based surface features are selected to build the scoring model, as shown in Table 3. Earlier AES systems such as PEG were built entirely with surface features, which only considered the form of the text and not the content of the text and were prone to misclassification. In order to avoid these shortcomings, other deep linguistic features need to be introduced to improve the accuracy of the system.

The readability indicators in Table 4 were chosen to assess the written language complexity of English learners. N is the total number of words in the composition among them. The total number of syllables in all words is denoted by SYL. Complex words, or those with more than two syllables, are referred to as CW. ASL is the average sentence length. And AWS is the average syllable length of words. The parameters in the FOG, FLESCH, and KINCAID readability formulas are determined by multiple regression equations. The values of FOG and KINCAID are proportional to the difficulty of the text, which roughly corresponds to the learners' language level. FLESCH measures the readability of the text, which is inversely proportional to the difficulty of the text.

Lexical diversity refers to the ratio of different lexical types T to the total number of words N in the text, as shown in Table 5.

Syntactic complexity measures the quality of writing by analyzing the proportion of each syntactic structure in learners' compositions. The syntactic structures such as clauses (SYN_C), subordinate clauses (SYN_DC), verb phrases (SYN_VP), complex noun phrases (SYN_CN), and parallel phrases (SYN_CN) were automatically labelled using the syntactic analyzer, and then the syntactic complexity was measured by calculating the ratio of the frequency of use of these structures to the total number of sentences in the text (S) as shown in Table 6.

As shown in Table 7, the grammatical correctness of learners' compositions was assessed by detecting spelling (SPELL_E) and complex grammatical errors (GRM_E). The detection of complex grammatical errors is based on chain grammar. The chain grammar consists of a dictionary and an algorithm, which contains the syntactic collocations of words. The algorithm slices the sentences according to the syntactic collocation of the words, and the grammatically correct sentences form a complete link, while the opposite is true, indicating the inclusion of grammatical errors.

As shown in Table 8, the overall and local coherence of the composition was assessed by counting the number of lexical links in the discourse according to the lexical articulation theory. $Links_{local}$ and $Links_{global}$ are the number of lexical links between adjacent and any two sentences in the composition, and N_{sent} is the total number of sentences in the composition.

After extracting the linguistic features, the features were filtered using the Random Forest (RF) algorithm. The RF regression constructs n decision trees using Bootstrap sampling and CART, with each node choosing an optimal feature from m randomly selected characteristics to split the data. When the training set of decision trees was selected

Table 5: Lexical diversity features.

Features serial number	Feature category	Feature coding	Calculation method
13	Word species/word order ratio	TTR	T/N
14	Square root TTR	TTR_ROOT	$T/\sqrt{N}$
15	Logarithmic TTR	ITR_LOG	$\log T/\log N$
16	Continuous sample TTR	TTR_SEG	Mean TTR of consecutive intercept samples
17	Random sample TTR	Feature coding	TTR mean of random intercept samples

Table 6: Syntactic complexity features.

Feature serial number	Feature category	Feature code
18	Proportion of clauses	SYN_C/S
19	Proportion of subordinate clauses	SYN_DC/S
20	Proportion of verb phrases	SYN_VP/S
21	Proportion of complex noun phrases	SYN_CN/S
22	Proportion of parallel phrases	SYN_CP/S

Table 7: Grammatical correctness characteristics.

Feature serial number	Feature category	Feature code
23	Spelling errors/total number of words ratio	SPELL_E/W
24	Spelling errors/total number of sentences	SPELL_E/S
25	Complex grammatical errors/total number of words	GRM_E/W
26	Complex grammatical errors/total number of sentences	GRM_E/S

Table 8: Discourse coherence features.

Feature categories	Feature categories	Feature code	Calculation method
Partial coherence	Partial coherence	COH_LOCAL	$Links_{local}/N_{sent}$
Overall coherence	Overall coherence	COH_GLOBAL	$Links_{global}/N_{sent}$

Table 9: FCE training and test sets.

Category	Training set	Test set
Type of learner's mother tongue	16	14
Type of essay topic	28	4
Number of essays	1141	97
Number of words in essay	878767	75550
Average score	27.81	27.46
Median score	28	26
Lowest score	0	13
Highest score	40	40
Standard deviation	5.49	5.94

using the self-service sampling method, about 35% of the samples did not appear in the data set, which constituted the out-of-bundle (oob) samples, and were used to evaluate the importance of the features, which was calculated as

$$\text{importance}(x) = \frac{\sum_{i=1}^{N_{\text{tree}}} (\text{MSE}(\text{permutate}(\text{oob}_i, x)) - \text{MSE}(\text{oob}_i))}{N_{\text{tree}}}, \tag{6}$$

x is the linguistic feature. N_{tree} is the number of decision trees. MSE is the mean square error of the i^{th} decision tree model predicting the score of the out-of-bundle sample (oob). permutate (∗) function is used to randomize the values of feature x in the out-of-bundle sample. In this paper, the linguistic features with importance greater than 0 are selected to construct the scoring model.

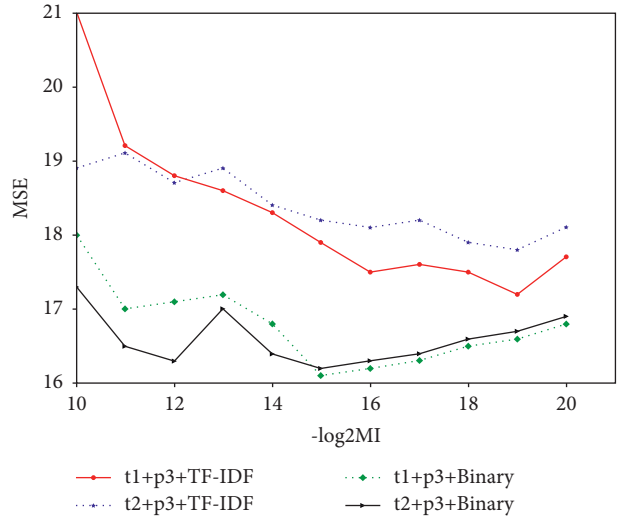

Figure 3: Bag-of-words feature-model error plot.

4. Experiments and Results

4.1. Experimental Data. The scoring model was trained and tested using the open data set FCE English Learner Corpus.

Table 10: Bag-of-words feature filter results.

Bag of words features	Word sequence			Part of speech sequences			$-\log_2$MI	MSE
	Monadic	Binary	Ternary	Monadic	Binary	Ternary		
BOW_A	√			√	√	√	16	16.41
BOW_B	√			√	√	√	15	16.42
BOW_C	√	√		√	√	√	15	16.45
BOW_D	√	√		√	√	√	16	16.51
BOW_E	√			√	√	√	17	16.52

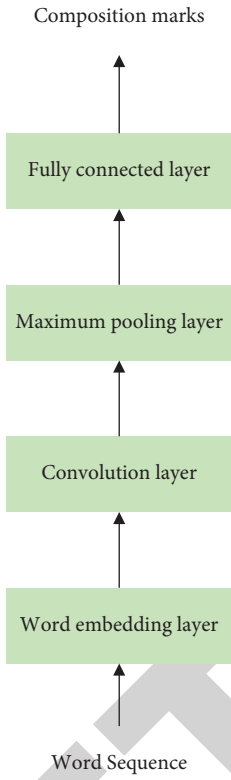

Figure 4: Deep learning scoring model 1 framework.

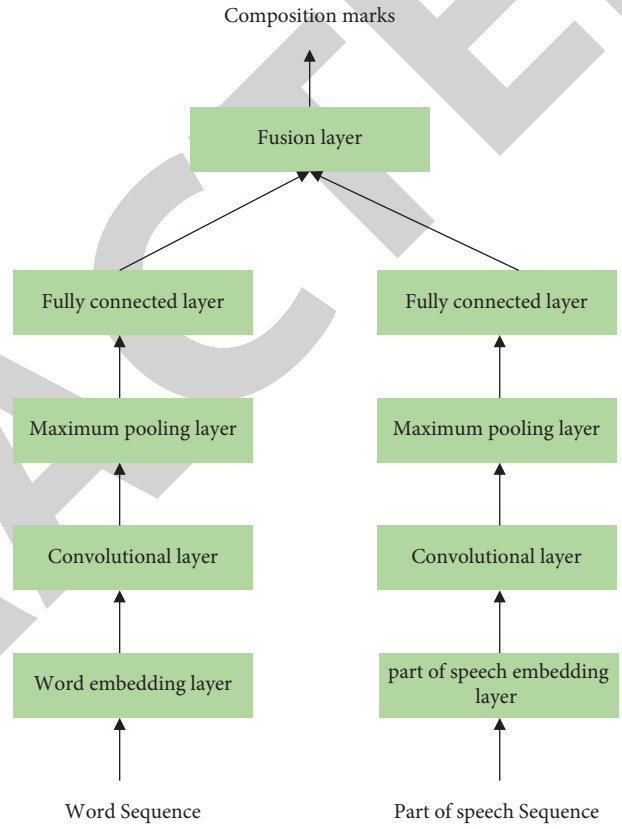

Figure 5: Deep learning scoring model 2 framework.

As shown in Table 9, the corpus consists of 1,141 Cambridge FCE exam essays and 97 test set essays with 950,000 words, each with a hand-corrected score. In addition, the FCE training and test sets are drawn from different years of FCE examinations and do not overlap in terms of writing topics.

Firstly, 90% of the samples from the training data were selected as the training set and 10% as the validation set using the random sampling method. Then, the bag-of-words features were extracted by setting the sequence length and the mutual information value of N elements. Both Binary and TF-IDF were used to weight the training and validation data. The loss function of the model is

$$J(w) = \min \frac{1}{2} w^T w + C \sum_{i=1}^{m} \left(\max\left(0, \left| y_i - w^T x_i \right| - \varepsilon\right)\right)^2. \quad (7)$$

Among them, (x_i, y_i) is the training set sample, $i = 1, 2, \ldots, m$ ($x_i \in R^n, w \in R^n$). The hyperparameter C is the constraint cost parameter and ε is the insensitive loss parameter. After obtaining the model parameters w, the mean square error of the model is calculated using the validation set, and the features are then filtered.

Figure 3 shows the relationship between the type of bag of words (type), the MI value, and the model error. Where t is the word sequence and p is the part of speech sequence. The Binary weighted model has a smaller error than the TF-IDF, and the model formed from a unary word sequence and a unary to ternary part of speech sequence has the lowest error. The five feature combinations with the lowest model errors are listed in Table 10. It can be seen that all features contain unary to ternary word sequences, but not ternary word sequences. There are many types of unary to ternary word sequences, and most of them have low frequencies, which is not conducive to the generalization of the model. In contrast, the part of speech is more frequent and reflects the lexical and syntactic collocations of the learners' written language, providing greater generalization ability.

Table 11: Results of a deep learning-based scoring model evaluation.

Model	Features	r	ρ	RMSE
Baseline model [23]	Word vectors	0.538	0.499	5.033
Model 1 of this paper	Word vectors	0.511	0.556	5.579
Model 2 in this paper	Word vectors + part of speech vectors	**0.542**	**0.575**	**5.004**

After screening the bag-of-words features, a random forest model was constructed using the statistical software *R*, and the importance of the linguistic features was calculated by (6). The number of decision trees in the model was $N_{\text{tree}} = 100$ and the number of randomly selected features was $m = 9$. The importance of the number of paragraphs (LEN_PARA) and the proportion of parallel phrases (SYN_CP/S) was less than 0. After excluding these two types of features, 26 types of linguistic features were finally selected to build the scoring model.

4.2. Deep Learning-Based Model Building and Evaluation. In this paper, two scoring models based on convolutional neural network (CNN) deep learning algorithms were designed, as shown in Figure 4. The experimental parameters of model 1 are as follows. The length of the word sequence in the input layer is the maximum number of words in the essay *dinput_length* = 900. The word embedding layer is a *Word2vec* pre-trained word vector with dimensionality *dword_embedding* = 300. The number of filters in the convolution layer is $h = 20$. Convolutional window length $m = 3$. Maximum pooling layer window length $n = 2$. Fully connected layer dimension *ddense* = 128.

Model 2 enhances the input layer with part of speech sequences in addition to word sequences, as seen in Figure 5. The lexical vector was obtained by training the model with a lexical embedding layer of dimension *dpos_embedding* = 50, and the output of the fully connected layer was then fused with the two types of sequences to predict essay scores. *ReLU* activation functions were used for each layer of Model 1 and Model 2, and the models were trained using the Adam optimizer with *batch size* = 16.

As shown in Table 11, the results of the evaluation showed that the deep network model with the addition of part of speech sequences had the highest accuracy. As mentioned earlier, part of speech sequences contains some shallow syntactic features that reflect the quality of the learner's writing, and a model that incorporates both word and part of speech sequences outperforms a single word vector model.

The evaluation results show that the accuracy of the deep learning-based scoring model is significantly higher than that of the benchmark model. The Pearson correlation coefficient **r**, Spearman correlation coefficient **ρ,** and root mean square error **RMSE** showed that the integrated scoring model built with the bag-of-words feature BOW_A and 26 types of linguistic features in Model 2 outperformed the existing benchmark model based on the FCE dataset.

5. Conclusion

The expanding Internet era gave birth to the online writing education and marking approach. An automatic marking system for English composition based on wireless networks can intensify students' impressions of incorrect language phenomena and help them avoid making the same mistakes again, as well as reduce English teachers' effort. However, the current automatic English essay scoring system suffers from slow scoring efficiency, low accuracy, and weak portability, while the development of artificial intelligence technology, especially the continuous breakthrough of deep learning technology, has effectively overcome such problems. In this paper, we propose an automatic English essay scoring system based on deep learning methods in a wireless network environment.

First, support vector regression was used to filter the subset of bag-of-words features that were highly correlated with essay scores by *N*-element sequence length and mutual information values. Then, the deep linguistic features of the essays were extracted in terms of text complexity, correctness, and coherence. Finally, a deep learning algorithm based on random forest regression was used to fuse the bag-of-words and linguistic features to construct a scoring model. This strategy minimizes the quantity of bag-of-words features, reduces the model's complexity, and refines the variety of linguistic features as compared to existing scoring systems. The composition quality of students was evaluated from a variety of angles, including vocabulary, grammar, and discourse. The findings reveal that the 26 linguistic variables chosen in this study are substantially connected with the composition quality, and that the deep learning-based scoring system surpasses the previous SVR-based scoring system.

Data Availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Xinyang Social Science Federation, Research on the Concept of "Three Causes" of Integrating Ideology and Politics into Courses Based on Outcome-Based Education (OBE) in the Context of New Liberal Arts (Project no. 2021JY067).

References

[1] S. C. Weigle, "English language learners and automated scoring of essays: c," *Assessing Writing*, vol. 18, no. 1, pp. 85–99, 2013.

[2] L. Guo, S. A. Crossley, and D. S. McNamara, "Predicting human judgments of essay quality in both integrated and independent second language writing samples: a comparison study," *Assessing Writing*, vol. 18, no. 3, pp. 218–238, 2013.

[3] Z. Yuan, "Interactive intelligent teaching and automatic composition scoring system based on linear regression machine learning algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 2, pp. 2069–2081, 2021.

[4] O. Lyashevskaya, I. Panteleeva, and O. Vinogradova, "Automated assessment of learner text complexity," *Assessing Writing*, vol. 49, Article ID 100529, 2021.

[5] P. Gamallo Otero, M. Garcia, I. del Río, and I. González López, "Avalingua," *Studies in Corpus Linguistics*, vol. 70, pp. 35–58, 2015.

[6] Z. Wang, H. Huang, L. Cui, and J. J. H. H. N. Chen, "Using natural language processing techniques to provide personalized educational materials for chronic disease patients in China: development and assessment of a knowledge-based health recommender system," *JMIR medical informatics*, vol. 8, no. 4, Article ID e17642, 2020.

[7] M. Uto, "A review of deep-neural automated essay scoring models," *Behaviormetrika*, vol. 48, no. 2, pp. 459–484, 2021.

[8] R. G. Garroppo, M. G. Scutellà, and F. d'Andreagiovanni, "Robust green wireless local area networks: a matheuristic approach," *Journal of Network and Computer Applications*, vol. 163, Article ID 102657, 2020.

[9] W. S. Albiladi and K. K. Alshareef, "Blended learning in English teaching and learning: a review of the current literature," *Journal of Language Teaching and Research*, vol. 10, no. 2, pp. 232–238, 2019.

[10] D. Xu and T. S. Rappaport, "Construction on teaching evaluation index system of track and field general course for physical education major in light of wireless network technology," *Journal of Intelligent and Fuzzy Systems*, vol. 37, no. 3, pp. 3435–3443, 2019.

[11] P. Bamdev, M. S. Grover, Y. K. Singla, P. Vafaee, M. Hama, and R. R. Shah, "Automated speech scoring system under the lens: evaluating and interpreting the linguistic cues for language proficiency," 2021, https://arxiv.org/abs/2111.15156.

[12] E. B. Page, "Computer grading of student prose, using modern concepts and software," *The Journal of Experimental Education*, vol. 62, no. 2, pp. 127–142, 1994.

[13] K. K. Y. Chan, T. Bond, and Z. Yan, "Application of an automated essay scoring engine to English writing assessment using many-facet rasch measurement," *Language Testing*, 2022.

[14] A. Kaur and M. Sasi Kumar, "Performance analysis of LSA for descriptive answer assessment," *Innovations in Computer Science and Engineering*, vol. 79, pp. 57–63, 2019.

[15] T. K. Landauer, D. Laham, and P. W. Foltz, "The intelligent essay assessor," *IEEE Intelligent Systems*, vol. 15, no. 5, pp. 27–31, 2000.

[16] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® V.2," *The Journal of Technology, Learning, and Assessment*, vol. 4, no. 3, pp. 4–15, 2006.

[17] L. Rudner, V. Garcia, and C. Welch, "An evaluation of intellimetricTM essay scoring system using responses to gmatawa prompts," *Retrieved*, vol. 9, 2005.

[18] D. Ramesh and S. K. Sanampudi, "An automated essay scoring system: a systematic literature review," *Artificial Intelligence Review*, vol. 55, pp. 1–33, 2021.

[19] R. Ridley, L. He, X. Dai, S. Huang, and J. Chen, "Prompt Agnostic Essay Scorer: A Domain Generalization Approach to Cross-Prompt Automated Essay Scoring," 2020, https://arxiv.org/abs/2008.01441.

[20] X. Lu and R. Hu, "Sense-aware lexical sophistication indices and their relationship to second language writing quality," *Behavior Research Methods*, pp. 1–17, 2021.

[21] T. Xia and X. Chen, "A weighted feature enhanced Hidden Markov Model for spam SMS filtering," *Neurocomputing*, vol. 444, pp. 48–58, 2021.