Hindawi
Mobile Information Systems
Volume 2023, Article ID 4450347, 8 pages
https://doi.org/10.1155/2023/4450347

Research Article

Automatic Scoring Model of English Interpretation Based on Semantic Scoring

Hua Ma

School of Foreign Languages, Chizhou University, Chizhou 247000, Anhui, China

Correspondence should be addressed to Hua Ma; 20151001086@m.scnu.edu.cn

Received 7 August 2022; Revised 11 September 2022; Accepted 23 September 2022; Published 17 April 2023

Academic Editor: Jiafu Su

In order to explore how English interpreting can achieve automatic scoring, the author proposes an automatic scoring model for English interpreting based on semantic scoring. This method recommends key technical problems and solutions based on information represented by semantic scoring, and explores the research on how interpreting can achieve automatic scoring of oral examinations. Research has shown that, the automatic scoring of English interpretation based on semantic scoring is faster than traditional methods, and the efficiency is improved by about 75%. However, the current automatic scoring of interpreters faces huge challenges. It needs to be tested and improved in more teaching, learning, and testing practice. The automatic scoring of interpretation should consider multiple dimensions such as semantic accuracy, content integrity, expressive fluency, and language authenticity.

1. Introduction

Smart education is the general trend of education development in the Internet era, and computers have become an important tool to assist learning [1, 2]. In the context of English language training and automated assessments, automatic question scoring systems that define candidate answer content such as reading questions and follow-up questions based on recent developments have reached a practical level. There is very little research. The interpretation test is a comprehensive test of foreign language application ability, including foreign language thinking ability and language organization ability. The research and development of an effective automatic scoring system for Chinese-English translation not only provides a platform for students to practice translation but also facilitates teachers' lectures and reduces the pressure of teachers' teaching and scoring. Based on this, by analyzing the scoring requirements of the interpreting test, focusing on the semantic scoring method at the content level of interpreting, a multiparameter Chinese-English sentence-level interpreting automatic scoring model is established as the basis for building an application system.

Taking the semantic scoring of spoken Chinese to English questions as the research focus, this paper introduces a semantic scoring model integrating long-short-term memory neural network and self-attention mechanism, which can be applied to keyword scoring and sentence semantic scoring [3]. The scoring principle of the model is as follows: firstly extract word and sentence features and represent them in a vectorized form, then use a bidirectional long-short-term memory neural network to optimize the feature vector, and then use the self-attention mechanism to obtain the semantic features of words or sentences, and finally the semantic score is calculated by a simple neural network. Experiments show that, compared with the stretchable recursive autoencoder-based semantic scoring model that performs better in semantic scoring, this model has better results in sentence semantic scoring. The average consistency rate between the sentence semantic scoring results and the original scores reached 55%.

Chinese-English translation quality evaluation has been one of the hotspots in the field of automatic Chinese-English translation quality evaluation in recent years. In terms of automatic spoken language scoring, most of the current research focuses on assessing spoken language at the level of

pronunciation quality, such as reading questions and follow-up questions [4]. English reading questions were scored using the most probable linear regression and most probable north probability algorithms with moderate results. However, there is still a lack of effective evaluation strategies for research on question types related to textual content, such as explanatory questions and repetitive questions (keywords, main content of sentences). Although some scholars have carried out relevant research, the actual results of large-scale oral test scores are very limited [5, 6].

Therefore, we provide an automated Chinese-English quality translation method. To evaluate translation quality, we choose three main parameters: semantic keywords, sentence similarity, and speaking ability. In sentence-level Chinese-English translation, the translation of keywords must be meaningful, and the general meaning of Chinese-English sentences must also be accurate. As a spoken language translator, the fluency parameter is also very important, and fluency also reflects the overall level of the translator's spoken language. In the Chinese-English translation question and answer scoring research, researchers generally focus on assessing the accuracy of the Chinese-English translation and the respondents' comprehension of the entire sentence. This is also the main reason for choosing the previous three evaluation parameters. Since many Chinese-English translation questions are the main types of Chinese-English translation questions, automatic scoring of Chinese-English translation questions has practical uses. The framework of the automatic scoring system for Chinese-English spoken translation is shown in Figure 1.

2. Literature Review

Rajagede et al. said that in the 1960s, people began to study controlled automatic assessment of oral quality in the form of university research projects [7]. The world's first large-scale computer-assisted language learning system PLATO system is a programmable learning system for automatic teaching. It was developed in 1959 by the University of Illinois and its business partner, Control Data. Its appearance has greatly promoted the application of computer in foreign language learning. The second-generation CALL system, represented by ALLP of the Massachusetts Institute of Technology, studies the application of computers in the field of education. Liu said that after the 1990s, the research on the third-generation CALL system paid more attention to the application of human-computer interaction and multimedia in language learning [8]. The Stanford International Consulting Institute (formerly known as the Stanford Research Institute) has a research and development group focused on speech research. The VILTS (Voice Interactive Language Training system) system developed by the group is used to test students' intonation and pronunciation fluency. The system uses a posterior probability algorithm and a log-likelihood algorithm to calculate the speaker's pronunciation accuracy, while using the duration score to characterize the speaker's pronunciation fluency. Carnegie Mellon University has designed a special automatic scoring system for the SET-10 oral English test. Xia, L. believes that the system can achieve good results in judging the spoken English proficiency of non-native speakers, but the system does not automatically score the types of open-ended questions [9]. The SCILL algorithm and the simplified posterior probability algorithm jointly developed by the University of Cambridge and the Massachusetts Institute of Technology calculate the pronunciation accuracy. The simplified posterior probability algorithm greatly shortens the calculation time, as a result, system performance has been improved.

Zhang believes that there are not many related research studies on the oral quality assessment of semicontrolled topics in the United States, and the main research representatives are the TOEFL test system and the Pearson Academic English test system [10]. Peng Research on the exploratory use of support vector machines and classification and regression tree algorithms for question and answer scoring methods for the TOEFL test [11]. The study not only found that vector machines have the advantages of quantitative analysis but also found that the classification and regression tree algorithm is very effective in mining the underlying laws of data. The TOEFL test system uses the multiple linear regression method to integrate the four characteristic scoring parameters of intonation, grammar, fluency, and vocabulary diversity to calculate the test taker's score. The scoring system grades the TOEFL test's six test question types in turn. The Pearson Academic Test System, developed by Pearson, selects intonation, fluency, sentence proficiency, and vocabulary as characteristic scoring parameters. Different from the TOEFL test system, the system does not distinguish between question types and scores, and directly calculates the scores of the candidates' four scoring characteristics based on the candidates' answers.

In the 1970s, American researchers began to study translation quality. Translation is a theory put forward by western countries to evaluate the quality of translation. It emphasizes that the translator must fully express the emotion, goal, and meaning contained in the text in the translator's language on the basis of understanding the original text language. After the 1980s, Western scholars tried to quantify the quality of explanations through empirical research, trying to find scientific variables and proportions to evaluate the quality of explanations. Ban and Translators conducted a survey of translators' expectations and found that the most important indicator for translators to measure translation quality is content consistency, followed by translation coherence, translation completeness, and grammatical features correctness [12]. Yuan believes that this study lays the foundation for an empirical study of interpretation quality assessment [13]. Gillier et al. proposed that, at major international conferences such as medicine and law, the audience's views on the interpretation of quality assessment were summarized and case studies were conducted, and it was found that there were differences in the interpretation of the quality assessment standards between audiences and translators. Ismagilov and I interviewed members of the International Conference of Translators, and the study found that audiences and translators place a high value on the accuracy and clarity of translated content [14].

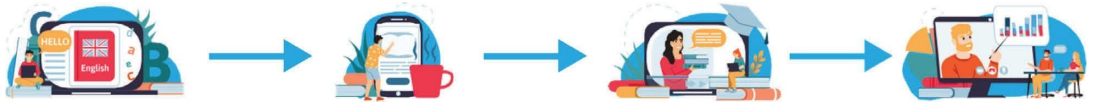

Figure 1: The framework of the automatic scoring system for Chinese-English spoken translation.

Compared with content, their requirements for expressions such as voice and speech speed are not high. Qin et al. introduced a semantic scoring model combining short-term and short-term memory neural networks and self-attention mechanism, which can be applied to keyword scoring and sentence semantic scoring. The scoring principle of the model is as follows: first, extract the features of words and sentences and express them in the form of vectorization, then optimize the feature vector using the two-way short-term memory neural network, and then use the self-attention mechanism to obtain the semantic features of words or sentences. Finally, a simple neural network is used to calculate the semantic score. The experimental results show that the average correlation between the model and the original score is 0.444, compared with the semantic scoring model based on the scalable recursive automatic encoder. The minimum coincidence rate with the original score is 95%. The highest consistency rate with neighboring countries is 74%. The automatic scoring model of Japanese interpretation based on semantic scoring has proved to be practical and has achieved good results [15].

The interpreting test and interpreting ability assessment can not only grasp the interpreting level of students, but also evaluate the teaching quality of teachers. It can be seen that the interpreting test and interpreting ability evaluation can provide an important basis for teaching improvement and have a certain guiding role. From the research status of interpreting quality assessment, it can be seen that information communication has become particularly important in interpreting assessment. In the scoring rules, it is rare to see scoring points that emphasize voice intonation, and more emphasis is placed on the integrity of information expression and the accuracy of information transmission.

3. Methods

3.1. Model Evaluation Indicators. An automatic scoring system is basically a computer model that scores on a rater's answer sheet, and the difference between the system scoring results and the manual scoring results reflects the performance of the automatic scoring system. To create an automated scoring system, you first need a standard set of human scoring data. This data are also known as the raw score of the test taker's answer. The aim of this study is to make the machine scoring results as close as possible to the candidate's initial score. We can evaluate system performance based on the correlation and consistency between the automatic scoring system scoring results and the initial scoring.

Correlation: correlation is an important metric for evaluating system scoring performance, which is used to measure the similarity between machine scores and initial scores in a linear sense. The Pearson correlation coefficient is used to measure the similarity between the machine score and the original score, and the calculation formula is shown in equation (1) as follows:

$$\rho\,\text{human, machine} = \frac{\sum_{n=1}^{N}\left[\left(Sn-\overline{S}\right)\times\left(\text{SR}n-\overline{\text{SR}}\right)\right]}{\sum_{n}^{N}\left(Sn-\overline{S}\right)^{2}\times\left(\text{SR}n-\text{SR}\right)^{2}}. \quad (1)$$

Convenience: the initial score consistency assessment and automatic scoring model have two parameters: the consistency level and the adjacent stability level, based on the initial score and an explicit distribution of the automatic scoring model for different scores [16]. The fitness is the ratio of the number of samples at the same level to the total number of samples, that is, the formula for calculating the ratio of the number of samples with a full score of S to the number of samples N in the second formula as follows:

$$\text{consistency} = \frac{S}{N}\times 100\%. \quad (2)$$

The adjacent consistency rate refers to the ratio of the number of samples whose machine score differs from the original score by one level (less than or equal to 0.5 points) to the total number of samples N, and it can usually be used as an effective indicator for comparing the degree of consistency between the two; the calculation formula is as shown in equation (3) as follows:

$$\text{Adj}_{\text{consistency}} = \frac{S_{+0.5}}{N}\times 100\%. \quad (3)$$

The LSTM storage unit is mainly composed of memory cells, forgetting gates, input gates, and output gates. The forgetting gate is used to screen old cell information and update the current memory cells according to the memory cell candidate information generated by the input gate. The sigmoid activation function in the forget gate processes the input information and outputs a value between [0, 1] [17]. The output value indicates that the old storage unit information is stored, the output value 0 indicates that all the information in the old storage unit is forgotten, and the output value 1 indicates that all the information in the old storage unit is stored. The information state calculation

formula of the forget gate is shown in equations (4) and (5) as follows:

$$f_t = \sigma\left(W_f \bullet \left[H_{t-1}, X_t\right] + b_f\right), \tag{4}$$

$$C_t = \mathrm{f}_t * C_{t-1} + \mathrm{i}_t * \overline{C}_t. \tag{5}$$

The input gate determines what new information can be added to the current memory cell. Each LSTM unit input includes the output Ht-x of the previous unit and the new information input Xr. The sigmoid activation function in the input gate processes the input information and outputs a value between [0, 1]. The output value represents the state in which the current information needs to be retained. Then, use the tanh function to generate new memory cell candidate information. The calculation formulas of the two functions are shown in equations (6) and (7), respectively, as follows:

$$i_t = \sigma\left(W_i \bullet \left[H_{t-1}, X_t\right] + b_i\right), \tag{6}$$

$$\overline{C}_t = \tanh\left(W_C \bullet \left[H_{t-1}, X_t\right] + b_C\right). \tag{7}$$

The output gate determines the output of the current cell information state [18]. As with the previous two gate designs, the sigmoid function is used to process the output result of the input gate and output a value between [0, 1]. The output value is multiplied by the tanh function value of the updated memory cell to obtain the final output result Ht. The calculation formula for the output gate is shown in equations (8) and (9) as follows:

$$\mathrm{o}_t = \sigma\left(W_o \bullet \left[H_{t-1}, X_t\right] + b_o\right), \tag{8}$$

$$H_t = o_t * \tanh\left(C_t\right). \tag{9}$$

It can be seen from the LSTM unit structure diagram and calculation formula that the memory cell C is propagated through a simple linear transformation in the LSTM network, so it can remain in the LSTM model for a long time. By adding forgetting gates, input gates, output gates, and memory cells to the neural unit to screen memory information, the LSTM unit uses memory cells C to retain long-term memory, and the hidden layer H to retain short-term memory, realizing the processing and learning of long sequence data [19].

The sigmoid function can map real numbers to the interval (0, 1), but not centered at zero. In the case that the feature difference is relatively complex or the difference is not particularly large, the sigmoid function is better for text classification. The biggest disadvantage of the sigmoid function is that it is easy to cause the problem of gradient disappearance when backpropagating. The sigmoid function formula is shown in equation (10) as follows:

$$f(z) = \frac{1}{1 + \exp(-z)}. \tag{10}$$

3.2. Examination Interpreting Scoring Criteria. It can be seen from the scoring standard of interpreting that interpreting emphasizes the accuracy of information transmission and the fluency of expressing information. In the exam, grammar and pronunciation and intonation are not tested [20]. Therefore, the information transfer is divided into keyword score and sentence semantic score, and the score parameter at the phonetic level selects fluency. Combined with the opinions of the interpreting teachers, the key words, sentence semantics and pronunciation fluency are determined as the characteristic scoring parameters of the Chinese-English interpreting automatic scoring system, as shown in Figure 2. In order to facilitate the analysis of the experimental results, the original scores of candidates can be divided into four grades, as shown in Table 1.

3.3. Conduct Experiments. The recording of candidates for the interpreting and listening exam in a certain examination room was used as the experimental data set. Select the first 1–5 questions in Volume A and Volume B, a total of 10 Chinese-to-English sentence translation questions as the research object. In the original data, there are 328 candidates in Volume A, that is, each question has 328 recorded data, a total of $328 * 5 = 1640$ data; Volume B has a total of 334 candidates, that is, each question has 334 recorded data, a total of $334 * 5 = 1670$ pieces of data. Since the graders are graded according to the candidates' recordings, the quality of the recordings has a great influence on the grades [21]. In order to reduce the influence of this factor of recording quality, we screen the original recordings. We excluded recordings with no sound or loud ambient noise. In order to reduce the subjective influence of manual scoring, we select the data with the scoring results less than or equal to 0.5 points by two raters for the experiment. After screening, the experimental data are shown in Table 2.

The experimental data are generally divided into two parts, one part is used for modeling and the other part is used to test the model hypothesis. The dataset used to create the model is called the training package, and the dataset used to test the accuracy of the model's assumptions is called the test package. The average sample size of the Chinese-English sentence interpretation for the self-study test is 272, which is a small sample of machine learning data. If the data are simply divided into a test set and a data set, it is difficult to make full use of the data of the few-sample data set. This simple approach to data distribution makes it difficult to accurately assess model predictability. K-cross-validation is a common method for evaluating the predictability of a few-shot model. K-cross-validation refers to randomly dividing the experimental data into K subsets, using each subset as a test set in turn, and combining the remaining subsets into a training set. After the model is trained K times, the average of all test sets is taken as the final calculation result [22]. This method reduces the influence of a single test set and training set division method on the prediction results by calculating the average value of the model prediction performance of each subset. After repeated testing, the author uses 3-fold cross-validation to evaluate the model scoring accuracy, and the basic steps are as follows:

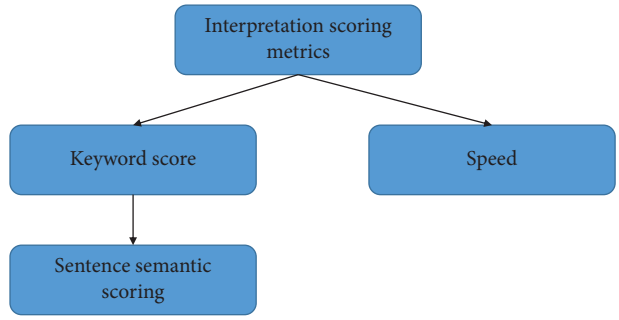

Figure 2: Scoring parameters of the automatic scoring model for interpretation.

Table 1: Scoring criteria and ratings.

Score	Grade	Grading criteria
[1, 2, 5]	A	The key information is accurate, the language is fluent, the vocabulary is used correctly, and the sentences are correct and popular.
[1, 5]	B	The key information is relatively accurate, the language expression is relatively fluent, the vocabulary application is relatively obtained, and the sentence generality is generally grasped.
[1, 0, 5]	C	The key information is not accurate enough, the language expression is insufficient, the sentences are biased, and the overall understanding is average.
[0, 5, 0]	D	Key information is incorrect or irrelevant, language is not fluent, sentences are highly distorted, and general comprehension is poor.

Table 2: Experimental data.

Volume A topics	Number of recordings	Volume B topics	Number of recordings
A_1	270	B_1	278
A_2	281	B_2	261
A_3	273	B_3	253
A_4	289	B_4	275
A_5	284	B_5	262

(1) Divide the data set into 3 subsets with basically the same amount of data,

(2) Use the first subset as the test set, and use the union of the remaining two subsets as the training set,

(3) Use the training set data to train the model, and use the test set data to verify the predictive ability of the model,

(4) Repeat steps 2-3, and take the remaining subset as the test set in turn.

In a small sample data set, if the training set contains the vast majority of sample data, theoretically the trained model can learn more data features [23]. However, at this time, the sample data in the test set will be relatively small, and the evaluation results are prone to large fluctuations, and the reliability will be reduced. If the test set contains more sample data and the sample data of the training set becomes relatively small, the model may not be able to learn the effective features of the data, thus reducing the credibility of the evaluation results. A common practice is to use about 2/3–4/5 of the samples as training data, and the remaining 1/5 to 1/3 of the samples as test data. Therefore, the division of the data set for each subject in this experiment follows the principle of training set: testing set = 7:3 for the experiment.

The experimental data in this study has a total of 10 sentence translation scoring questions, and the average number of samples per sentence is 272, which is a small sample experiment in deep learning. In order to reduce the randomness of the experimental results and improve the reliability of the experiment, during the experiment, 10 questions were modeled and tested, and 3-fold cross-validation was used, that is, each question was tested 3 times, and the last 3 times were taken, the average value of the experiment is used as the experimental result of each question. Three experiments are carried out for each question, and Table 3 shows the data set of each experiment.

4. Results and Analysis

Sentence semantic score is not only to test the overall understanding of the candidate's sentence but also to reflect the candidate's ability to express the sentence. In the manual scoring, there is no specific item of sentence semantics, but the scorer will score the overall situation of the candidate's translation. The rater can directly tell by listening to the recording whether the examinee is speaking a complete sentence or a string of keywords. However, computers cannot easily make such judgments [24]. At the semantic

Table 3: Experimental dataset.

Topic	Number of training sets	Number of test sets
A_1	180	91
A_2	187	94
A_3	182	90
A_4	193	96
A_5	189	95
B_1	185	92
B_2	174	88
B_3	177	84
B_4	183	92
B_5	174	87

level, sentences usually consist of keywords and common words. Keywords are generally words that can affect the meaning of a sentence, consisting of nouns, verbs, and adjectives with specific meanings. The universal word is not decisive for understanding the meaning of the whole sentence, but it is an essential part of the sentence, such as prefixes, conjunctions, and sentences. The author compares the sentence semantic scoring model based on the stretchable recurrent autoencoder neural network with the proposed BiLSTM-AM-based semantic scoring model and analyzes the advantages and disadvantages of the two models in sentence semantic analysis.

Stretchable recursive autoencoders for semantic detection: this model is based on a recurrent neural network and improved with autoencoder, which can extract effective features of sentence generality [25]. Recurrent neural network processes sentence sequence information through tree structure; the basic process is to combine the input sentences according to the order of their network nodes to generate parent nodes, and then process the newly generated parent nodes and other child nodes as input again. The model recurses from bottom to top until all child nodes are integrated, and the characteristics of the last root node are obtained. Moreover, this feature can be thought of as a feature extraction representation for all input nodes. The autoencoder is divided into an encoding layer and a decoding layer. The former obtains another representation of the input data by encoding and compressing the features of the input data. The latter restores the original input by decoding. If the recovery result of the decoding layer is very close to the input data properties, it is assumed that the encoded function can represent an approximation of the input. The URAE-based sentence semantic scoring model studies sentence semantic properties through encoded latent layer neurons. The repetitive neural network of the autoencoder, combined with URAE, reconstructs the function compressed into the parent node to form the child node, filtering out the nature of the error measurement method between the atomic node and the reconstructed child node. If the error is too large, it means that the effect of the adjacent node merging is not good, and it is necessary to continuously traverse and optimize to combine the two adjacent nodes with the smallest reconstruction error.

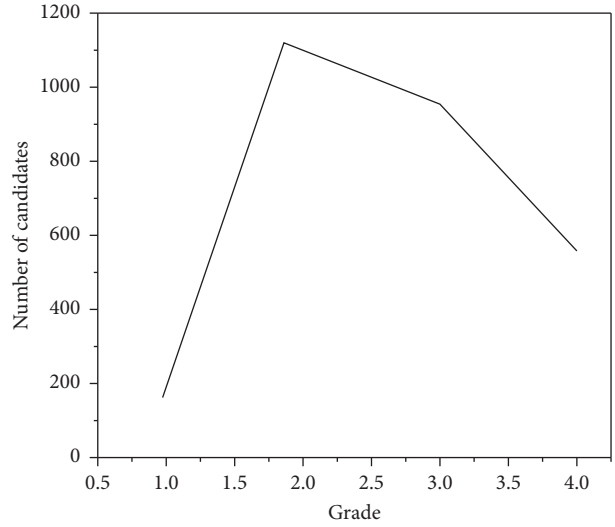

Figure 3: Distribution of manual scoring levels.

To minimize the effect of fictitious subjectivity on grades, the authors used data from answer sheets that were graded independently at multiple points, with the average of the two teachers' scores as the initial score. To confirm the validity of the initial score and the accuracy of the test data, we confirmed the accuracy of the machine scoring model by subtracting the speech data from over 0.5 points and comparing the difference in scores. Two-point scorer 1, scorer 2, or the histogram of the first score follows a normal distribution, with a high number of candidates with moderate scores, and a small number of candidates with excellent and unsuccessful scores. Research shows that, in general, the distribution of student grades follows a normal distribution, which demonstrates the accuracy of student assessments across grades and the reliability of initial grades. Figure 3 shows the distribution of the two scorers and the original score.

In order to test the superiority of the random memory algorithm in melting point, the control variable method was used to conduct comparative experiments under the condition that the speech rate scoring method, the keyword scoring method, and the sentence meaning scoring method were compatible. Stochastic memory algorithms are replaced by linear regression forecasting methods for integral supply. First, we compare the score distribution of the scores of the two models with the score distribution of the original scores. Whether the random memory algorithm or the linear regression prediction algorithm is used to combine the scores, it can be seen that the final score of the two-point model conforms to the normal distribution law. From the four-level distribution of ABCD, the automatic scoring model using the random forest algorithm for score fusion is closer to the original score distribution.

Original scoring protocols is compared using the automated scoring model scoring results and different scoring methods. The correlation between the scores of the two models and the initial score is more than 90%, while the score of the automatic scoring model using the random

memory algorithm is generally 100% adjacent to the initial score. In terms of average transaction performance, the automatic scoring model using the random memory algorithm reached 77.4%, and the automatic scoring model using the linear regression prediction method reached 55.5%. The automatic model using the random forest algorithm has an average agreement rate of 21.9% higher than using the linear regression method.

5. Conclusion

It is undeniable that doubts about the automatic scoring system have never stopped since it appeared in the field of language testing. People may not believe that AI technology can be used for automatic scoring of interpretation, and think that it cannot make correct judgments on the content of interpretation. We believe that, from the perspective of formative assessment, applying AI technology in low-risk teaching assessment or learning diagnosis to help interpreting teachers and students, still a viable evaluation option.

Because there is a certain error in the ASR results, the final score prediction model will also be affected to varying degrees due to differences in recording quality. In the future, with the technical development of the automatic scoring system for interpretation, the speech features of the expression dimension of the CSE interpretation scale can be further added to the algorithm model, which can not only effectively avoid recording quality problems, but also help to give more detailed speech suggestions. In the future, deep learning algorithms will be used to directly model raw speech and textual information, enabling automatic scoring systems to better learn how to utilize textual semantic features for interpretation score prediction. In terms of teaching use, we will continue to enrich descriptions and feedback scenarios. The automatic grading system for interpreting can also provide interpreting students with more personalized and detailed feedback (such as specific error examples and correction suggestions) in the future, establish milestones in the development of interpreting abilities, form the overall trend of learning and personal portraits of students' learning, and teach teachers and students test to provide more reference.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] A. Barwise, M. E. Yeow, and D. K. Partain, "The premise and development of check in—check-in for exchange of clinical and key information to enhance palliative care discussions for patients with limited English proficiency," *American Journal of Hospice and Palliative Medicine*, vol. 38, no. 6, pp. 533–538, 2021.

[2] J. Zhang and F. Wang, "A better medical interpreting service: interpreter's roles and strategies under goffman's participation framework," *International Journal of Translation Interpretation and Applied Linguistics*, vol. 3, no. 1, pp. 1–14, 2021.

[3] Y. S. Kashikar, "The clinical images in the works of jhumpa lahiri: an approach in medical humanities," *College English*, vol. 17, no. 4, pp. 2456–8104, 2020.

[4] N. Voievodina, "Plurilingual education today: a perspective of teaching intercultural competence in German after English to student interpreters/translators," *ARS LINGUODIDACTICAE*, vol. 5, pp. 45–54, 2020.

[5] M. Leather, G. Fewings, and S. Porter, "Outdoor education: the romantic origins at the university of st mark and st john," *History of Education Review*, vol. 49, no. 1, pp. 85–100, 2020.

[6] A. M. El-Zawawy, "Simultaneous interpretation of complex structures from English into Arabic," *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, vol. 35, no. 1, pp. 30–64, 2022.

[7] R. A. Rajagede, "Improving automatic essay scoring for Indonesian language using simpler model and richer feature," *Kinetik Game Technology Information System Computer Network Computing Electronics and Control*, vol. 6, no. 1, pp. 11–18, 2021.

[8] J. Liu, L. Lin, and X. Liang, "Intelligent system of English composition scoring model based on improved machine learning algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 2, pp. 2397–2407, 2021.

[9] L. Xia, D. Luo, J. Liu, M. Guan, Z. Zhang, and A. Gong, "Attention-based two-layer long short-term memory model for automatic essay scoring," *Journal of Shenzhen University Science and Engineering*, vol. 37, no. 6, pp. 559–566, 2020.

[10] 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.

[11] H. Peng and Q. Li, "Research on the automatic extraction method of web data objects based on deep learning," *Intelligent Automation and Soft Computing*, vol. 26, no. 3, pp. 609–616, 2020.

[12] H. Ban and J. Ning, "Design of English automatic translation system based on machine intelligent translation and secure internet of things," *Mobile Information Systems*, vol. 2021, no. 7639, pp. 1–8, Article ID 8670739, 2021.

[13] Z. Yuan, C. Jin, and Z. Chen, "Research on language analysis of English translation system based on fuzzy algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 4, pp. 6039–6047, 2021.

[14] I. I. Ismagilov, A. I. Sabirova, D. V. Kataseva, and A. S. Katasev, "Collection scoring models development and research based on the deductor analytical platform," *Nexo Revista Científica*, vol. 33, no. 2, pp. 608–615, 2021.

[15] W. Qin, "Automatic scoring model of Japanese interpretation based on semantic scoring," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 4, pp. 5809–5819, 2020.

[16] Q. Xiong, "Research on English spoken semantic recognition machine learning model based on neural network and statistics fusion," *Journal of Intelligent and Fuzzy Systems*, vol. 38, no. 6, pp. 7341–7350, 2020.

[17] J. Xu and C. Yi, "The scoring mechanism of players after game based on cluster regression analysis model," *Mathematical Problems in Engineering*, vol. 2021, no. 3, pp. 1–7, Article ID 5524076, 2021.

[18] Y. Wang, "Implications of blended teaching based on theory of semantic wave for teaching English writing in high school," *Journal of Higher Education Research*, vol. 3, no. 2, pp. 166–168, 2022.

[19] L. Duan, K. Yang, and R. Lang, "Research on Automatic Recognition of Casting Defects Based on Deep Learning," *IEEE Access*, vol. 9, pp. 12209–12216, 2020.

[20] X. Li, F. Xu, X. Lyu et al., "Dual attention deep fusion semantic segmentation networks of large-scale satellite remote-sensing images," *International Journal of Remote Sensing*, vol. 42, no. 9, pp. 3583–3610, 2021.

[21] J. Chen and S. Ou, "Research on the construction of the semantic model for Chinese ancient architectures based on architectural narratives," *The Electronic Library*, vol. 38, no. 4, pp. 769–784, 2020.

[22] D. Selva, B. Nagaraj, D. Pelusi, R. Arunkumar, and A. Nair, "Intelligent network intrusion prevention feature collection and classification algorithms," *Algorithms*, vol. 14, no. 8, p. 224, 2021.

[23] J. Hu, Y. M. Kang, Y. H. Chen, X. Liu, X. Li, and Q. Liu, "Analysis of aerosol optical depth variation characteristics for 10 years in urumqi based on MODIS_C006," *Huan jing ke xue Huanjing kexue*, vol. 39, no. 8, pp. 3563–3570, 2018.

[24] R. Huang, S. Zhang, W. Zhang, and X. Yang, "Progress of zinc oxide-based nanocomposites in the textile industry," *IET Collaborative Intelligent Manufacturing*, vol. 3, no. 3, pp. 281–289, 2021.

[25] Q. Zhang, "Relay vibration protection simulation experimental platform based on signal reconstruction of MATLAB software," *Nonlinear Engineering*, vol. 10, no. 1, pp. 461–468, 2021.