J. Electrical Systems 20-6s (2024): 501-508

[1]Xin Yan

Development of English Composition Correction and Scoring System Based on Text Similarity Algorithm

Journal of Electrical Systems

Abstract: - This study focuses on creating an innovative system that leverages a text similarity algorithm to provide accurate correction and scoring of English compositions. By harnessing advanced computational techniques, the system aims to streamline the grading process and offer more consistent feedback to students. The research methodology involves the design and implementation of the text similarity algorithm, which analyzes various linguistic aspects such as vocabulary usage, grammar, coherence, and organization. This algorithm is then integrated into a comprehensive scoring system that evaluates student compositions against reference texts and provides detailed feedback on areas for improvement. The effectiveness of the system is validated through rigorous testing using a diverse set of English compositions and comparison with human grading. The results of this investigation show how well the system that was created can automatically correct and score English papers with a high degree of correctness and dependability. By reducing the burden of manual grading and providing immediate feedback to students, the system has the potential to enhance the teaching and learning process in English language education. Overall, by developing a scalable and effective method for compositional assessment based on text similarity algorithms, this research advances educational technology.

Keywords:

1. Introduction

The development of an automated English composition correction and scoring system represents a significant advancement in educational technology, offering the potential to revolutionize the way compositions are evaluated in English language education [1]. By harnessing the power of text similarity algorithms, this system seeks to provide objective and consistent feedback to students while reducing the workload of instructors [2]. Traditional methods of composition assessment often suffer from subjectivity and inconsistency, as different instructors may interpret and evaluate compositions differently [3]. Moreover, manual grading can be time-consuming, limiting the frequency and quality of feedback provided to students [4]. Therefore, there is a compelling need for automated systems that can analyze compositions objectively and provide immediate feedback to facilitate students' learning and improvement in English writing skills [5].

The proposed system is built upon the foundation of text similarity algorithms, which are computational techniques used to measure the similarity between two or more texts based on various linguistic features [6]. These algorithms analyze aspects such as vocabulary usage, sentence structure, coherence, and grammar to determine the degree of similarity between a student's composition and reference texts [7]. By comparing compositions against a corpus of pre-existing texts, the system can identify areas of strength and weakness in students' writing and generate tailored feedback accordingly [8]. Additionally, the scoring component of the system utilizes machine learning techniques to assign objective scores to compositions based on their similarity to high-quality reference texts [9]. This approach not only streamlines the grading process but also ensures consistency and fairness in evaluation, regardless of the instructor or evaluator [10]. Overall, the integration of text similarity algorithms into the composition correction and scoring system holds great promise for improving the efficiency and effectiveness of English composition assessment in educational settings [11]. In today's digital era, the demand for effective and efficient tools for language learning and assessment has never been higher [12]. Among these, the development of automated systems for correcting and scoring English compositions holds significant promise for educators and students alike [13]. Traditional methods of composition correction and scoring are often time-consuming and subjective, relying heavily on manual

[1] Basic Education Department, Henan Technical College of Construction, Zhengzhou, Henan, 450046, China

*Corresponding author e-mail: yanxin@hnjs.edu.cn

Password 405405

mail.hnjs.edu.cn

assessment by instructors [14]. Moreover, providing timely and detailed feedback to students on their writing can be challenging, particularly in large class settings [15]. Researchers and programmers have recourse to innovative technologies like machines intelligence (MI) and human language processing (NLP) to streamline the grading and compositions rectification processes in order to address these challenges [16]. In this context, the application of text similarity algorithms has emerged as a promising approach to enhancing the accuracy and efficiency of automated composition assessment [17].

An important development in educational technology is the creation of a system for English composition correction and assessment based on text similarity algorithms. By leveraging the power of NLP and ML techniques, this system aims to provide educators and students with a robust tool for assessing writing proficiency and offering targeted feedback. Unlike traditional methods, which rely on predefined rules or templates for error detection and scoring, the proposed system utilizes advanced algorithms to analyze the similarity between students' compositions and reference texts [18]. This innovative approach enables the system to identify both surface-level errors, such as grammar and spelling mistakes, and deeper structural issues, such as coherence and organization [19]. Furthermore, by comparing students' compositions to a vast repository of high-quality reference texts, the system can offer more nuanced and contextually relevant feedback, thereby supporting students' language learning journey. Overall, the development of this system holds great potential to revolutionize the way English composition is taught, assessed, and learned in educational settings [20].

2. Related works

The sentence sophistication feature scoring algorithm and machine learning-based autonomous scoring algorithm for English composition were developed by Zhao, Y. [21]. The author also looked into the algorithm's potential effects on English writing instruction, carried out a number of experiments, and provided evidence of the algorithm's accuracy.Guo, P., [22] developed using phrase semantic maps, and trials were used to confirm its efficacy. According to the experimental results, Sub Visual representations 10, 12, and 13 scored 200 on the first two simulated texts then 1 on the next two. These three subgraphs' differences on several test texts ranged from 30 to 0.3, although they were not statistically significant. Furthermore, with an acceptable threshold of 0.34, or 87.54%, the F1 value achieved the optimal value for the extraction of nonsensical sentences.

Huang, Y., and Zhang, C., [23] assessed and examined the compositional grammar by combination of the N-grammar model with the Link Grammar (LG) detector. Then, information entropy was used to determine the composition level. In the end, the grammar weighting and overall composition level were used to compute the composition score. The suggested technique in this work has a total recall and accuracy weight of 89.9%, a score that is 26.6% higher than LG and 9.7% higher than Grammarly, according to the testing data. The recommended technique outperformed the human review in the performance test of scoring the full essay, with an incorrect proportion of 87.29%, but with a quicker average running time that was only 22.69 seconds and a reduced overall mean square error of just 3.08.

Wang, N., [24] had drawn attention and was frequently utilized in college English instruction as a result of the quick development and automatic mark technology to evaluate English writing. In light of this, the goal of this essay was to provide useful guidance and support for college English teachers. It accomplished this by summarizing the significance of marking undergraduate English writings and by planning and executing an English composing marking system from the perspectives of a front-end component functionality realization and a word mistake repair module.

3. System model

Multiple interrelated components make up the system model for the creation of the text similarity algorithm-based written in English correction and grading system. Firstly, the data collection phase encompasses the compilation of a diverse dataset of English compositions from various sources. These compositions are annotated with correct grammar, spelling, and structural errors, forming the foundation for the subsequent stages. In the feature extraction step, natural language processing techniques are employed to extract linguistic features and compute text similarity metrics. These features, along with pre-trained word embeddings, serve as inputs to the text similarity algorithm, which compares student compositions with reference texts. Error detection and correction algorithms then analyze the compositions, identifying and suggesting corrections for grammar, spelling, punctuation, and coherence errors based on similarity analysis and annotated dataset. The scoring system component assigns scores to student compositions based on the severity and frequency of detected errors. This system incorporates feedback mechanisms to provide detailed explanations for assigned scores and suggestions for improvement. The developed algorithms and scoring system are integrated into a user-friendly interface for educators and students, facilitating easy access and utilization. Upon deployment in educational institutions and online learning platforms, the system undergoes rigorous evaluation and testing to assess its performance and effectiveness. Feedback from users is collected to identify areas for improvement, leading to iterative refinement and enhancement of the system to better serve its intended purpose in improving English composition correction and scoring.

4. Research design

Initially, the design involves defining the scope and objectives of research, which includes identifying the specific features and functionalities of the composition correction and scoring system, as well as outlining the criteria for evaluating its performance and effectiveness. This step ensures clarity and alignment with the intended outcomes of the project. Next, the research design involves selecting and implementing appropriate methodologies and approaches for system development and evaluation. This includes employing techniques from natural language processing, machine learning, and computational linguistics to design and implement the text similarity algorithm for comparing input compositions with reference texts. Additionally, the research design incorporates principles of software engineering and human-computer interaction to develop a user-friendly interface and ensure the usability and accessibility of the system. Figure 1 shows the architecture for the research design of development of english composition correction and scoring system based on text similarity algorithm

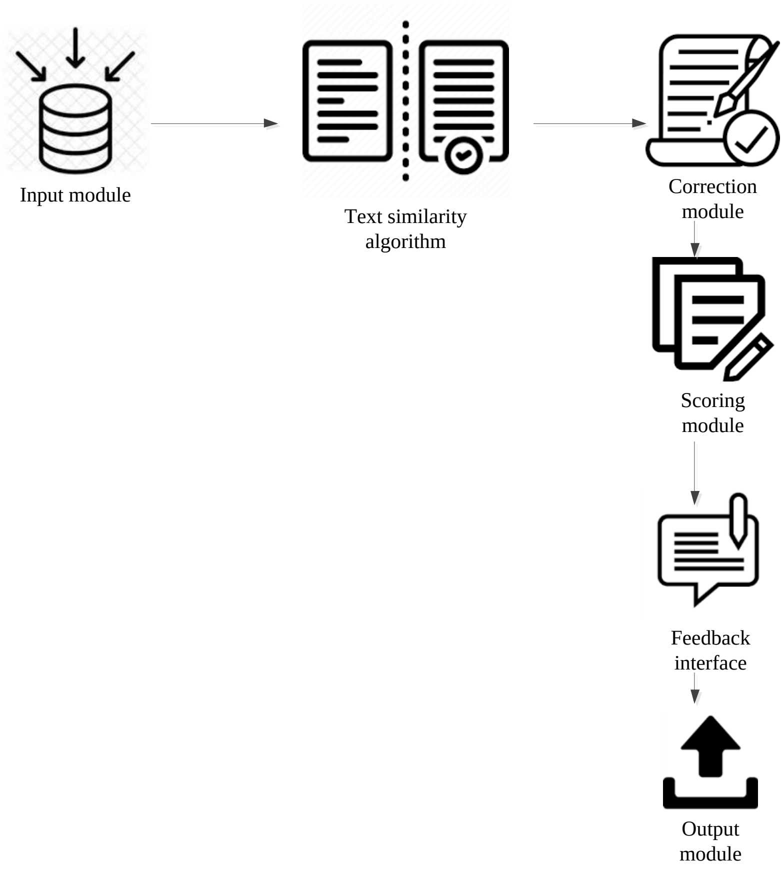

Figure 1. Research design based on text similarity algorithm

The research design also includes data collection and experimentation to confirm the precision and functionality of the developed system. This involves gathering a diverse dataset of English compositions and reference texts for training and testing the algorithm, as well as conducting user studies and evaluations to assess the system's usability, effectiveness, and user satisfaction. The research design also includes iterative refinement and optimization of the system based on feedback and evaluation results to enhance its performance and address any identified limitations or shortcomings. Overall, the research design for the development of the English composition correction and scoring system based on a text similarity algorithm integrates methodologies from various disciplines to achieve the project objectives of creating an accurate, efficient, and user-friendly system for enhancing English writing proficiency and providing valuable feedback to users.

4.1 Data Processing

Sorting the sentences based on similarity is crucial in order to identify the neighbors. Once the similarity was calculated, the N sentences with the highest similarity are referred to as neighbors. The median weighting approach is used to forecast the final score after removing all phrase neighbors. The median weighting approach considers each phrase scoring based on the whole data when producing recommendations. When there are many sentences, this method works well. This approach is not able to forecast adequately when the sentence count is little. The following is its calculation formula:

$$Q_{vj} = S_j + \frac{\sum_{k \in O} sim(j,k) \times (S_{vi} - S_k)}{\sum |Sim(j,k)|} \tag{1}$$

4.2 Collaborative Filtering Algorithm

This work offers a collaborative filtering approach that makes use of a combination of paragraph attributes, based on term attribute grouping for resemblance optimization.

$$D = \frac{1}{m} \sum_{j=1}^{m} \beta(y_i) \beta(y_i)^Q \tag{2}$$

The preceding formula is transformed using the kernel function, and each number's inner product is then calculated, thus with this equation,

$$weight(i, j) = \frac{dis_{\max} - d_j}{dis_{\max} - dis_{\min}} \tag{3}$$

When assessing the quality of the advice, a standard called the mean absolute error, or MAE, is employed. The product's prediction rating is more accurate when the MAE is lower. If we assume that the projected data score set is (p1, p2, p3, . . . , pn), the corresponding real information score setting is (q1, q2, q3, . . . , qn), and N represents the number of missing points in the score matrix, then the MAE is computed by taking the difference across the two score sets. Once all phrases have been grouped and all groups whose trait distance is less than the threshold are kept out, the area of search S(i) of each sentence is determined. The target words and these classes do not have an equal amount of similar link, and in terms of objective attributes, clusters having near-attribute distance are more similar to each other. When scoring, the conventional approach to determining similarity simply takes into consideration the external similarity; it ignores the variations in the phrase qualities. The classes in this range are those that are at different attribute distances from the target text.

4.3 Words' String Similarity

The subsequence containing the longest comparable sequence (LCS) metric corresponding to the string similarity index was found after additional normalization and other tweaks. These were combined and weighted after the three different modified versions of LCS were used. However, the less long string's length is disregarded by LCSR, which can occasionally have a big effect on the similarity result.

$$w_1 = length(LCS(s_j, t_k))^2 \tag{4}$$

A consecutive common subsequence is required for a significant amount of matching in the structure of databases matching, while it is not required in classical LCS. Starting at character 1, we use the maximal consecutive lengthy common subsequence (MCLCCSn) and the highest consecutively longest common subsequence (MCLCCSn). The shorter string or the greatest amount of consecutive words from the shorter length that match the longer string successively are produced by the algorithm when it receives two strings as input. The matching between the two strings must start with an identical character (character 1).

$$w_2 = length(MCLCS\ (r_i, t_j)) \tag{5}$$

$$w_3 = length(MCLCS\ _n(s_j, t_k))^2 \tag{6}$$

The average weighted of these separate values, and, where and are weights and, to get the string similarity score. Consequently, both of these strings are similar in:

$$\beta = x_1 w_1 + x_2 w_2 + x_3 w_3 \tag{7}$$

4.4 Text Similarity algorithms between words

The literature has an unexpected number of word-to-word similarity measurements. These metrics include those derived from large-scale text collections that are based on the concepts of information models (also known as corpus-based measures), as well as those that are based on knowledge (derived from dictionaries and thesaurus) or distance-oriented values produced on semantic networks. Because of their wide type coverage, we concentrate on measures based on corpuses. Knowledge bases often lack the types that are utilized in real-world documents. There are surprisingly many word-to-word similarity measurements in the literature. These metrics include those that utilize knowledge (derived from dictionaries and thesaurus) or distance-oriented values generated on semantic networks, as well as those generated by large-scale writing collections that are based on the concepts of information models (also called corpus-based measures). We focus on metrics based on corpuses due to their broad type coverage. Types that are used in documents seen in the actual world are typically absent from knowledge repositories. Nevertheless, from an algorithmic standpoint, our system benefits from using SOC-PMI as it can ascertain the similarity. The frequencies and contexts were sourced from the British National Corpus (BNC). The method considers phrases that appear frequently in both collections and aggregates their PMI ratings (from a range in the other list) to calculate the relative text similarity.

$$g^{pmi}(u, x) = \log_2 \frac{g^c(u_i, x) \times n}{g^j(u_j) g^u(x)} \tag{8}$$

Rough guidelines are used to find the value of β. An expression's β-PMI summation function is specified in reference to another word. With respect to word w2, word w1's β-PMI summation equation is:

$$Sim(x_1, x_2) = g(x_1, x_2, \beta_1) + g(x_1, x_2, \beta_2) \tag{9}$$

It takes in as arguments the greatest value (λ) returned by the semantic similarity function Sim() and the two words, ri and sj. It offers a similarity score with an inclusive range of 0 to 1. For example, when λ equals 20, the SOC-PMI technique produces a value of 0.986 for the terms "graveyard" and "cemetery". Text Resemblance technique includes a distinct module dedicated to word similarity. Therefore, if someone wants to experiment with different word-similarity techniques (dictionary-based, corpus-based, or hybrid), they can use any other word similarity method instead of SOC-PMI.

Algorithm 1. Normalization of text similarity algorithm

Input: $s_j\ and\ t_k$

Two strings as input and outputs either the shorter string

$$w_2 = length(MCLCS\ (r_i, t_j))$$

$$w_3 = length(MCLCS_n(s_j, t_k))^2$$

To determine the relative text similarity,

$$g^{pmi}(u,x) = \log_2 \frac{g^c(u_i, x) \times n}{g^j(u_j) g^u(x)}$$

the β-PMI summation equation for word w1 is:

$$Sim(x_1, x_2) = g(x_1, x_2, \beta_1) + g(x_1, x_2, \beta_2)$$

If $w > \beta$, then

$$w \leftarrow 1$$

Else

$$w \leftarrow \frac{w}{\beta}$$

End

Output: w

There are numerous crucial phases involved in creating an English compositions correction and grading system determined by a text similarity algorithm. First, the system will take the user-provided input composition and preprocess it by encoding the content into individual words, eliminating punctuation and stop words, and stemming or lemmatizing words to return them to their base form. Next, using a text similarity technique like cosine similarity or Jaccard similarity, the system will fetch reference texts from a database and compute the similarity score between each reference text and the input composition. Utilizing language processing methods and grammar rules, the system will identify mistakes in grammar, errors in spelling, and stylistic inconsistencies in the input composition by comparing the frequency of words or word vectors in the composition with those in the reference texts. Based on the similarity metric, the system will suggest corrections and improvements. The system will assess the quality and consistency of the input composition by providing numerical scores or grades based on established scoring criteria after considering the repair and enhancement suggestions. These standards could cover things like syntax, vocabulary, coherence, and general language ability. During the review process, the system will take into account the appropriateness, clarity, conciseness, and accuracy of language use. Following scoring and assessment, the system will provide the user with feedback that includes recommendations for enhancing language competence and composition abilities as well as corrective ideas for faults and mistakes found. Ultimately, the user will see the corrected layout, scoring results, and feedback from the system. The corrected text will be highlighted or annotated with suggested changes, and a numerical score or grade will be assigned to the composition along with comprehensive feedback on the composition's strengths, weaknesses, and areas for improvement will be provided.

5. Results and discussion

In general, achieving high performance on one of the two criteria is simple, but achieving high performance on both is more challenging. The geometric mean of recall (R) and precision (P) is known as the F-measure (F), which represents a trade-off between the two metrics. The following defines these performance measures:

5.1 Performance evaluation

Eleven distinct similarity thresholds, with intervals of 0.1, are employed, ranging from 0 to 1. For instance, our technique predicts 1369 pairs as right when we use a similarity criterion of 0.6. Of the 1725 manually annotated pairs, 1022 pairs are correct.

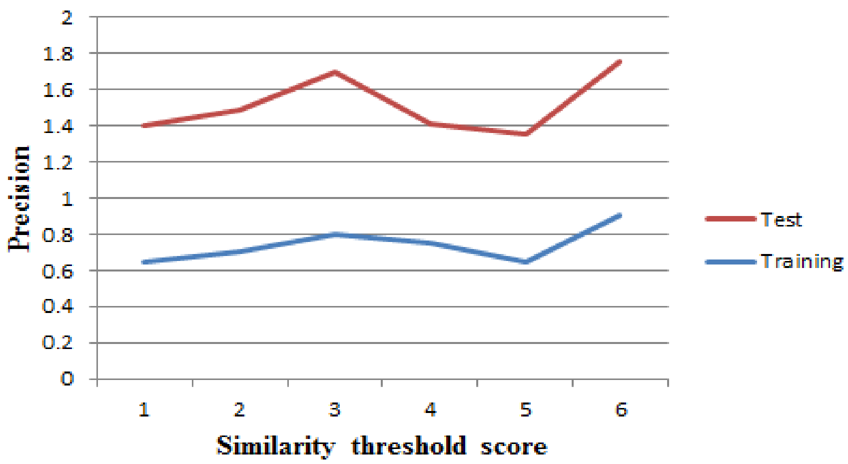

Figure 2. Two data sets' precision vs. similarity threshold curves for eleven distinct criteria of similarity.

Reliability versus resemblance standard Recall vs. resemblance threshold curves and curves for the two of them data sets (training and test) for each of the twelve potential similarity thresholds are shown in Figures 2 and 3, respectively. In Figure 3, values for recall for the training and test sets of information are 0.0054 and 0.0044, respectively, when we use comparable characteristics threshold score of 1 (i.e., The corresponding word for word, so no similarities in semantics matching is required). This is achieved when we use a similarity cutoff score of 1. Utilizing the training data set, the best accuracy was 72.42% after deciding on 0.6 as the similarity cutoff score. As a result, this criterion is applied to the testing set, as demonstrated by its accuracy of 72.64% (1369 pairs are correctly predicted by this strategy, while 1022 pairs are inaccurate).

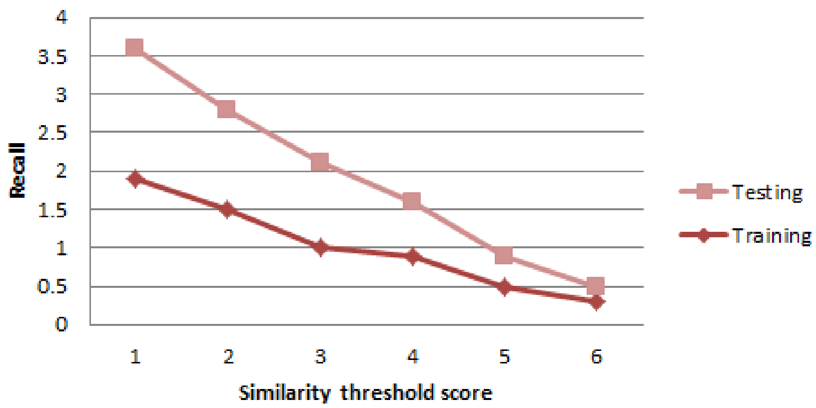

Figure 3. Two data sets' recall vs. similarity threshold curves for eleven distinct similarity criteria.

6. Conclusion

In conclusion, the English composition rectifying and grading system, which is based on a text similarity algorithm, represents a significant advancement in the domains of natural language processing and technology for education. This research project has effectively addressed the demand for automated as well as effective methods of evaluating English works and giving writers constructive comments by putting unique procedures and approaches into practice. The system developed in this study offers several key benefits, including improved accuracy and consistency in assessing written compositions, enhanced efficiency in grading large volumes of work, and valuable insights into areas for improvement in language proficiency. By leveraging the capabilities of text similarity algorithms, the system can effectively compare student compositions with reference texts and provide detailed feedback on grammar, vocabulary usage, coherence, and other linguistic aspects. Furthermore, the user-friendly interface and intuitive design of the system ensure accessibility and usability for both educators and students, facilitating seamless integration into educational settings. The iterative refinement and optimization process employed during system development have resulted in a robust and reliable tool that can effectively support English language learning and teaching objectives. Overall, the development of this composition correction and scoring system represents a significant contribution to the advancement of educational technology and language assessment methods. It has the potential to revolutionize the way English composition is taught and evaluated, ultimately enhancing the learning experience and proficiency levels of students in English language education.

References

1. Huang, Y. and Zhang, C., 2024. Network Algorithms for Intelligent Evaluation of Composition in Middle School English Cloud Classrooms. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)*, *19*(1), pp.1-17.
2. Zhao, H., 2024, January. Design and Implementation of Computer Aided English Composition Marking System. In *Proceedings of the 3rd International Conference on New Media Development and Modernized Education, NMDME 2023, October 13–15, 2023, Xi'an, China*.
3. Wang, N., 2024. Check for updates Application and Sharing of Corpus in College English Teaching System Under the Internet Environment. *Frontier Computing on Industrial Applications Volume 1: Proceedings of Theory, Technologies and Applications (FC 2023)*, *1131*, p.82.
4. Huang, L. and Ma, L., 2024. Research on the Application of Intelligent Grading Method based on Improved ML Algorithm in Sustainable English Education. *Scalable Computing: Practice and Experience*, *25*(1), pp.451-463.
5. Maslim, M., Wang, H.C., Putra, C.D. and Prabowo, Y.D., 2024. A Trustworthy Automated Short-Answer Scoring System Using a New Dataset and Hybrid Transfer Learning Method.
6. Wang, I.X., Wu, X., Coates, E., Zeng, M., Kuang, J., Liu, S., Qiu, M. and Park, J., 2024. Neural Automated Writing Evaluation with Corrective Feedback. *arXiv preprint arXiv:2402.17613*.
7. Wang, I.X., Wu, X., Coates, E., Zeng, M., Kuang, J., Liu, S., Qiu, M. and Park, J., 2024. Neural Automated Writing Evaluation with Corrective Feedback. *arXiv preprint arXiv:2402.17613*.
8. Tian, L., An Innovative Demonstration of Functional Linguistic Ideas for Teaching English in Colleges and Universities Based on Multi-objective Optimization Models. *Applied Mathematics and Nonlinear Sciences*, *9*(1).
9. Yu, L., 2024. Research on Retrieval Method of Online English Grammar Questions Based on Natural Language Processing. *International Journal of Emerging Technologies in Learning (Online)*, *19*(2), p.101.
10. Puerto, E., Aguilar, J. and Pinto, A., 2024. Automatic Spell-Checking System for Spanish Based on the Ar2p Neural Network Model. *Computers*, *13*(3), p.76.
11. Gao, R., Merzdorf, H.E., Anwar, S., Hipwell, M.C. and Srinivasa, A., 2024. Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, p.100206.
12. Zhang, L., The Auxiliary Utility of Big Data Intelligent Translation in English Writing. *Applied Mathematics and Nonlinear Sciences*, *9*(1).
13. Tang, Y. and Chen, H., Research on Efficient English Teaching Model Combining Corpus Microtext Analysis. *Applied Mathematics and Nonlinear Sciences*, *9*(1).
14. Luhtaru, A., Korotkova, E. and Fishel, M., 2024, March. No Error Left Behind: Multilingual Grammatical Error Correction with Pre-trained Translation Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1209-1222).
15. Yin, L., 2024. Fuzzy information recognition and translation processing in English interpretation based on a generalized maximum likelihood ratio algorithm. *PeerJ Computer Science*, *10*, p.e1668.
16. Islam, M.N., Jahangir, R., Mohim, N.S., Wasif-Ul-Islam, M., Ashraf, A., Khan, N.I., Mahjabin, M.R., Miah, A.S.M. and Shin, J., 2024. A Multilingual Handwriting Learning System for Visually Impaired People. *IEEE Access*.
17. Zhao, Y., 2021. Research and design of automatic scoring algorithm for english composition based on machine learning. *Scientific Programming*, *2021*, pp.1-10.
18. Guo, P., 2024. Construction of Semantic Coherence Diagnosis Model of English Text based on Sentence Semantic Map. *Scalable Computing: Practice and Experience*, *25*(1), pp.327-339.
19. Huang, Y. and Zhang, C., 2024. Network Algorithms for Intelligent Evaluation of Composition in Middle School English Cloud Classrooms. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)*, *19*(1), pp.1-17.
20. Wang, N., 2024. Check for updates Application and Sharing of Corpus in College English Teaching System Under the Internet Environment. *Frontier Computing on Industrial Applications Volume 1: Proceedings of Theory, Technologies and Applications (FC 2023)*, *1131*, p.82.