Hindawi
Security and Communication Networks
Volume 2023, Article ID 9869702, 1 page
https://doi.org/10.1155/2023/9869702

Retraction

Retracted: An English Writing Grammar Error Correction Technology Based on Similarity Algorithm

Security and Communication Networks

Received 10 October 2023; Accepted 10 October 2023; Published 11 October 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets 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 own 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] Q. Li, "An English Writing Grammar Error Correction Technology Based on Similarity Algorithm," *Security and Communication Networks*, vol. 2022, Article ID 3690789, 8 pages, 2022.

Hindawi
Security and Communication Networks
Volume 2022, Article ID 3690789, 8 pages
https://doi.org/10.1155/2022/3690789

Research Article

An English Writing Grammar Error Correction Technology Based on Similarity Algorithm

Qingyu Li

Hainan College of Foreign Studies, Wenchang, Hainan Province 571321, China

Correspondence should be addressed to Qingyu Li; liqingyu@hncfs.edu.cn

Received 16 May 2022; Revised 16 June 2022; Accepted 30 June 2022; Published 20 July 2022

Academic Editor: Mukesh Soni

Students are the focus of English writing instruction, which emphasizes their initiative and creativity in the classroom, fosters the development of their self-directed learning and comprehensive language skills, raises their overall English level and mastery ability, and cultivates well-rounded English learners. It is to get yourself out of the dilemma of Chinese English. This study proposes an automatic error correction method for English writing grammar based on a similarity algorithm, which is applied to students during the English writing process. To address the issue of grammatical errors, this study proposes an ontology and grammar rules based on the English writing corpus. Our experimental results demonstrate that our highlighting algorithm can effectively correct grammatical errors in written English, demonstrating the algorithm's efficacy.

1. Introduction

Natural language [1–4] is the polar opposite of artificial language and is the language that the general public uses to communicate with one another on a daily basis. Artificial languages are frequently distinguished by self-created vocabularies, strict grammar, and a limited range of ideograms, and thus, classified as belonging to a language category that is more difficult to become accustomed to, but not difficult for the general public to comprehend. Individuals struggle to establish a sense of the first language as they age because natural language is inextricably linked to the entire social culture and evolves over time. Furthermore, humans can naturally communicate in natural languages due to their syntactic and semantic flexibility. Natural language, on the other hand, is the most difficult to master due to the plethora of exceptions, variations, and commands that it contains.

Natural language processing technology has advanced at a rapid pace in recent years, thanks to the continual growth of computer technology. Calculating text similarity [5–7] is a crucial part of the natural language processing process. Text similarity computation is a technique used in natural language processing to determine the degree to which the semantics expressed in distinct texts are similar. Text similarity calculation and related technologies have been applied in a variety of fields, including machine translation, information retrieval, text classification, automatic summarization, public opinion analysis, semantic sentiment analysis, dialogue systems, paper duplication checking, and others.

The calculation of text similarity is also commonly employed in the practice of patent search. When undertaking semantic retrieval, some researchers provide a ranking to the similarity of linked comparison texts ranging from high to low. When doing semantic retrieval, there are some comparable documents that are more relevant than others and should be displayed in a preferential manner. In an intelligent retrieval system, semantic retrieval assesses the similarity between documents and then returns the documents with the closest text semantics, based on the similarity between the documents in the collection. Text similarity, on the other hand, has some practical uses in the field of English writing. These models, which include the term frequency-inverse document frequency (TF-ID) model [8, 9], the latent semantic indexing (LSI) model [10, 11], and others, are common text similarity calculation models that are commonly used in automatic scoring systems, web search, and DNA sequence matching.

Traditional English composition instruction places little importance on student feedback. In this method, students

are expected to memorize a large number of words and sample essay templates, or to outline the entire examination's content, as opposed to taking the examination itself. The dominant position of the students is not reflected and acknowledged, and students are afraid to share their views on the educational content, resulting in a composition that is identical to the one that was previously written. There are no novel concepts or innovations. In the course of writing, students are more likely to commit a variety of grammatical errors. Using a similarity algorithm to train your English writing skills will provide you with a new perspective on English writing. The rapid development of society has led to a rise in the use of network multimedia instruction, which enables students to interact with teachers and situations while also gaining knowledge from those interactions. Instead of merely preparing lesson plans, teachers are now responsible for devising a variety of activities to pique students' interest in learning, creating and distributing courseware videos, and directing students to an environment that corresponds to the learning theme. Permitting children to write essays about marine life, for instance, is a fantastic idea. After completing a composition assignment in English, we arrange for students to visit an aquarium to experience something more engaging and vibrant than the classroom description. Teachers are responsible for stimulating their students' senses through a variety of vivid and vivid methods, whereas students exercise their subjective initiative to perceive and accept knowledge. This multidimensional, multimodal, and multisensory integration also serves to increase students' writing motivation. In contrast, the grammatical errors committed by the students during this procedure posed significant correctional challenges for the teachers. To address this issue, we have developed an algorithm-based, error-correcting approach for English writing grammar that is capable of effectively correcting students' grammar errors in English writing assignments.

2. Related Work

In recent years, the input to English learning has never been stronger, as demonstrated by the increase in reading and vocabulary skills. However, the output of English language learning, such as writing and speaking, has not improved proportionally. Writing is an essential component of the English language learning process and also serves as evidence of one's overall language proficiency. This course aims to improve students' oral and written communication skills so that they can effectively communicate in their future professional and social relationships. In contrast, the children's current writing skills significantly lag behind their other English skills. A small percentage of students have passed the CET-4 and CET-6 writing levels. Improving the writing skills of students has emerged as a pressing issue in the English education in our country, and it must be addressed immediately. Students' writing can attain the level of writing in their native language as a result of the development of writing techniques or procedures during the acquisition of the mother tongue and the elimination of grammatical difficulties [12–14].

Grammar is considered to be one of the most crucial components of language learning for second-language learners, especially for those who are just starting out. Grammar knowledge is essential for learners in Malaysia because English was widely used as a medium of teaching in secondary schools until Malay was adopted as the official language in 1981. The previous study has demonstrated that using students' writing as a starting point for addressing grammatical ideas is the most effective method of promoting a learner's knowledge of writing grammar. The researchers came to the conclusion that teaching punctuation, inflection, and sentence structure in a writing environment is more beneficial than approaching the topic in isolation and teaching separate techniques. Teachers can assist students in revising and editing their writing by focusing students' efforts to detect and rectify problems in grammar usage when they rewrite and edit their writing. According to some experts, a teacher who notices a large number of pupils writing sentences with erroneous modifiers can use examples of student writing to deliver a brief lecture on the subject. For the purpose of revising, teachers can urge students to share their own drafts with their classmates. Integrating grammar instruction into the revision and editing process allows students to apply what they have learned instantly, helping them to recognize the relevance of grammar to their own writing and improve their writing.

The theory of error analysis (EA) [15–18] is one of the most significant theories in the field of second-language acquisition. In order to assess the errors committed by L2 learners, it is necessary to interpret the errors found by comparing the norms acquired by the learners with the norms of the target language. It is the study of inappropriate forms created by persons who learn a language, particularly a foreign language, which is considered EA in language instruction by certain researchers. In particular, EA refers to the investigation of linguistic ignorance, the analysis of what people do not know, and the investigation of how they attempt to cope with their ignorance. Education via imitation (EV) is a language learning strategy that focuses on the mistakes that students make and assists instructors in understanding the language acquisition process. A number of academics have attempted to identify acceptable correction approaches that can aid in the effective learning and teaching of English since varied faults are considered as a means to an end. The reason for this is that writing allows one to examine language abilities, recall skills, and critical-thinking skills. Some academics utilized EV to look for faults in a corpus of 72 papers written by 72 people, which was analyzed by the researchers. In the past, mistakes were considered defects that needed to be fixed when writing. While some academics believe that errors are unimportant, others disagree, stating that the faults themselves are significant. He believes that systematic mistake analysis can assist teachers in determining the form of reinforcement that will be most effective in the classroom setting.

Generally speaking, a corpus [19–21] is a huge number of papers that people have gathered and organized for use in a certain field. For the study of word similarity calculation utilizing large-scale corpora, many researchers have

used traditional mutual information approaches, which are based on the principle of mutual information. Some researchers employ correlation entropy [22–24] to determine the similarity of two words. In order to calculate the distance between words, some researchers employ more complicated probabilistic models. According to others, the bag-of-words technique, which estimates word similarity by building word context vectors and computing the cosine value of the included angle between the vectors, is preferable.

Some authors explain how to calculate the distance from a new angle using a distance-based method. A shorter path between two words indicates that they are related, and the traversal process does not need or alters the path in a significant way. The author proposes an independent calculation technique to demonstrate this. There is some optimization in the final product. On the basis of HowNet, several academics have suggested a method for computing lexical-semantic similarity between words. It is recommended that, while comparing the semantic expressions of two concepts, you use this method, which follows the principle that the overall similarity is equal to the weighted average of partial similarity. The method of determining the semantic distance according to the upper-lower connection and converting it is used to determine the similarity of two sememes. Wikipedia [25] is a collaborative information base built on the Web 2.0 technology of the internet. As a large-scale corpus with a semantic dictionary function, Wikipedia can be regarded as a valuable semantic data resource in the scientific study because of its unique information organization and information organization. When it comes to word similarity calculations, Wikipedia is a better resource base than a search engine because it has a more reasonable knowledge structure and a broader coverage than WordNet. Another group of researchers has discussed word similarity computation approaches that use several information sources, including the structured semantic information of the semantic dictionary and the information content of the corpus. The similarity between two words is calculated, and trials have shown that this method is preferable to the old method of measuring similarity between two words. According to certain researchers, the calculation methods of semantic similarity between terms based on ontology may be classified and summarized from three perspectives: the information-based method, the distance-based method, and the hybrid method. On the basis of this, a hybrid technique based on directed acyclic graphs and intrinsic information is developed, which is described below. This method avoids the problem of corpus analysis and has a reasonable degree of accuracy because it does not require corpus analysis. Others have proposed the algorithm of word similarity from the perspectives of semantic and statistical fusion, respectively, while still others have implemented the classic approaches based on HowNet and mutual information in a comprehensive manner. The outcomes demonstrate that the algorithm is capable of producing outcomes that exceed public expectations. Using the HowNet network, researchers have proposed a context-based word similarity algorithm. By constructing membership functions to calculate the fuzzy importance of word context information, this algorithm can effectively address the issue of data noise.

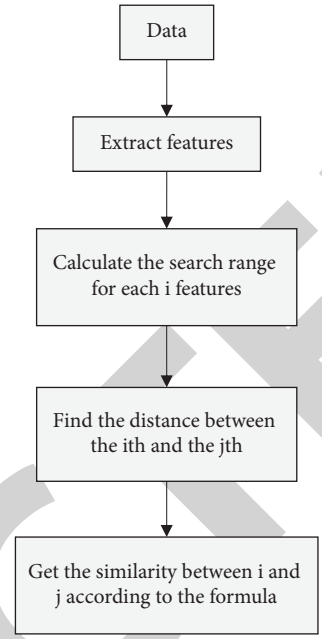

Figure 1: Average degree algorithm flowchart.

3. Research Design

3.1. Data Processing

3.1.1. Data Sources. According to the authors, the dataset for this work was obtained from a data collection of students' English practice compositions at a Beijing institution. The dataset contains both the original text of the English composition and artificial identification tags for grammatical faults.

3.2. Algorithm Introduction

3.2.1. Similarity Calculation Method. It is possible to calculate similarity using a variety of methods. The most widely used are modified cosine similarities and Pearson's similarities, among other things.

(1) Cosine Similarity Has Been Corrected

$$\mathrm{Sim}(i, j) = \frac{\sum_{u \in U_{ij}} \left(r_{ui} - \bar{r}_i\right)\left(r_{uj} - \bar{r}_j\right)}{\sqrt{\sum_{u \in U_i} \left(r_{ui} - \bar{r}_i\right)^2} \sqrt{\sum_{u \in U_j} \left(r_{uj} - \bar{r}_j\right)^2}}. \quad (1)$$

There exist several datasets that have assessed sentences i and j, with U_i and U_j being the datasets that have evaluated sentences i and j, respectively, and with $U_{ij} = U_i \cap U_j$. In this equation, r_{ui} and r_{uj} denote the scores of data u on sentences i and j, respectively, while $\bar{r}_i$ and $\bar{r}_j$ denote the average scores of sentences i and j.

The algorithm flowchart is shown in Figure 1.

(2) Pearson's Correlation Coefficient

$$\mathrm{Sim}(i,j) = \frac{\sum_{u\in S}\left(r_{ui} - \bar{r}_i\right)\left(r_{uj} - \bar{r}_j\right)}{\sqrt{\sum_{u\in S}\left(r_{ui} - \bar{r}_i\right)^2}\sqrt{\sum_{u\in S}\left(r_{uj} - \bar{r}_j\right)^2}}, \quad (2)$$

where S denotes the dataset of sentences I and j that share the same label.

To determine the neighbors, it is important to sort the similarity of the sentences and identify the N sentences with a high similarity as neighbors after calculating the similarity. When all sentence neighbors have been filtered out, the average weighting approach is applied to forecast the final score. When generating suggestions, the average weighting strategy takes into account the score of all sentences based on the data in a complete manner. When there are a large number of sentences, this method is appropriate. When there are only a few sentences, this method cannot forecast well. Its calculating formula is as follows:

$$P_{ui} = R_i + \frac{\sum_{j\in N}\mathrm{Sim}(i,j)\times\left(R_{uj} - \bar{R}_j\right)}{\sum_{j\in N}|\mathrm{Sim}(i,j)|}, \quad (3)$$

where $\mathrm{Sim}(i, j)$ represents the degree of similarity between sentence i and phrase j. N is a set of sentences that are the closest neighbors to i.

3.2.2. Collaborative Filtering Algorithm. This study provides a collaborative filtering algorithm based on phrase attribute grouping and similarity optimization that makes use of the combination of sentence attributes. We adjust the threshold, calculate the attribute distance between the target sentence and each cluster center, filter out clusters that are smaller than the threshold, and look for neighbors in these clusters. Finally, we use the improved similarity calculation method to obtain the target sentence I and the sentences within the search range. We determine and predict the neighbor set based on the similarity between the two. The sentences are clustered using the K-means technique, which is based on the attribute feature matrix for each sentence. In the classic K-means technique, the Euclidean distance between two sentences is used to determine the attribute distance between the sentences, and new cluster centers are generated by continuous iteration until the clusters become stable. In addition, because the statement attribute matrix contains only Boolean values, the Euclidean distance cannot accurately capture differences in characteristics between different phrases; thus, it is difficult to find new cluster centers by averaging over multiple sentences while doing iterations. As a result, using the K-means method, the following operations are carried out in this study.

(1) As a measure of the difference in attribute distance across sentences, one can use the Jaccard distance. The Jaccard distance is mostly used to determine how similar two samples are to one another. It is possible to determine the similarity between two 0-1 type samples by comparing the proportion of distinct elements in the two sets. The following is the calculating formula:

$$d_j(A,B) = 1 - J(A,B) = \frac{|A\cup B| - |A\cap B|}{|A\cup B|} = \frac{M_{10} + M_{01}}{M_{10} + M_{01} + M_{11}}. \quad (4)$$

M_{11} is the number of attributes in both statements whose values are 1, and it is the number of attributes that share this property. M_{01} represents the number of attributes where the value of the attribute is either 0 or 1, depending on which statement is being considered. M_{10} is the number of attributes, one of which has a value of 1, while the other has a value of 0, and both are statement attributes.

(2) During the iterative process of clustering, rather than using the numerical-average approach, the new cluster center will utilize the phrase that has the shortest attribute distance and is the same sentence as the cluster center. We calculate the attribute distance $(\mathrm{dist}_{i,1}, \mathrm{dist}_{i,2}, \ldots, \mathrm{dist}_{i,n})$ between each sentence in the class and the same sentence, and make a note of the attribute distance $\mathrm{Dist}(i) = \mathrm{dist}_{i,1} + \mathrm{dist}_{i,2} + \cdots + \mathrm{dist}_{i,n}$. For instance, if there are k classes $(C_1, C_2, \ldots, C_k)$, we calculate the attribute distance $(\mathrm{dist}_{i,1}, \mathrm{dist}_{i,2}, \ldots, \mathrm{dist}_{i,n})$. We take the statement that has the shortest attribute distance and $\mathrm{Dist}(i)$ and use that as the new cluster center. Next, we finish the updating of all class centers and lastly iterate until the cluster is stable. The algorithm is relatively complicated, but it can be executed offline, which means that it does not negatively impact the efficiency of the computation.

The clustering process performed by the modified K-means algorithm results in the production of k classes $(C_1, C_2, \ldots, C_k)$ and cluster centers $(\mathrm{cc}_1, \mathrm{cc}_2, \ldots, \mathrm{cc}_k)$. If you simply search inside the same category, the only phrases you will be able to recommend are those that have features that are comparable to those of the data. This lacks originality and makes it difficult to mine the data for its potential interest. This study determines the sentence attribute threshold on the basis of clustering, filters out the categories whose attribute distance is within the threshold range, and then searches for neighbors within the categories that were filtered out. The following are the specific measures to take:

Step 1: we configure the distance threshold δ for the attribute. It is important that the value of δ be established with reference to the current circumstances.

Step 2: we determine the attribute distance between the target sentence and each of the cluster centers, and then record the attribute distance of the ith sentence using the formula $\mathrm{Disa}(i) = (d_1, d_2, \ldots, d_k)$.

Step 3: for each of the target sentence's attribute distances $(d_1, d_2, \ldots, d_k)$, if the condition $dj < \delta$, $j \in 1, 2, 3, \ldots, k$ is

met, then the jth class will be classified as the neighbor search for the target phrase i within the range $S(i)$.

In this manner, the search range of the neighbors of the target sentence can be restricted, and the computation time can be correspondingly lowered, leading to an improvement in the algorithm's efficiency to some degree.

After the sentences have been clustered and the classes whose attribute distance is less than the threshold that have been filtered away, the search range $S(i)$ of each sentence is obtained. This range contains the classes that are separated from the target sentence by a variety of attribute distances. These classes and target phrases do not share a similar connection that is weighted equally, and clusters that have near-attribute distances are more comparable to one another in terms of their objective qualities. The standard way of calculating similarity only takes into account the objective similarity when scoring, and it does not take into account the differences in the characteristics of the sentences. This study offers the notion of attribute weight based on the clustering of sentence attributes. Additionally, it combines the classic Pearson's similarity calculation approach with a new method in order to optimize the similarity calculation. The distance between two attributes has a bearing on the weight of the attribute, but that bearing is negative. The shorter the distance, the more comparable the subjective and objective characteristics are. It is determined by the attribute distance between the category of the sentence and the target sentence, and the attribute weight is normalized before being incorporated into the calculation.

$$\text{weight}(i, j) = \frac{\text{dis}_{\max} - \text{d}_j}{\text{dis}_{\max} - \text{dis}_{\min}}, \quad j \in S(i). \tag{5}$$

Among them, the class with the biggest attribute distance will have weight = $\text{dis}_{\max} - \text{dis}_{\min} = 0$, resulting in that all sentences in this class have a similarity of 0 with the target sentence. To avoid this, we improve it.

$$\text{weight}(i, j) = \frac{1 + \left(\text{dis}_{\max} - \text{d}_j\right)}{1 + \left(\text{dis}_{\max} - \text{dis}_{\min}\right)}, \quad j \in S(i), \tag{6}$$

where $\text{dis}_{\max}$ represents the maximum attribute distance that can exist between the target sentence I and the cluster center while the search range is being used. $\text{dis}_{\min}$ refers to the minimum distance for an attribute. d_j represents the distance in attributes between the class center to which the sentence j belongs and the sentence that is being targeted. After making the necessary changes, the weight $\in [0, 1]$.

Pearson's similarity calculation method is paired with the attribute weight that is determined by the category of the sentence, and this combination is then integrated with the classic sentence-based collaborative filtering process. The concluding formula for determining degrees of similarity is as follows:

$$\text{Sim}^a(i, j) = \text{Sim}(i, j) \times \text{weight}(i, j). \tag{7}$$

The complete similarity is determined by using weighted summation, and this is performed on the basis of doing an analysis of the factors that have an impact on semantic similarity. The idea behind the improved algorithm is as follows: when constructing the ontology model, factors such as depth, type, and node density are introduced. The weight factor is set to adjust its influence on the structural similarity. The method of weighted calculation is used to solve the shortcomings of geometric quantitative calculation. These are the three main components of the ontology model. Concept attributes, information amount, and information structure are presented as potential influencing factors of semantic similarity. Additionally, an experienced professional's knowledge and a method based on trial and error are utilized in order to calibrate the overall similarity weight that is selected. The following is the formula for performing a comprehensive calculation of semantic similarity:

$$\begin{aligned}\text{sim}(c_1, c_2) = {} & \omega_1 \cdot \text{sim}(c_1, c_2)_{-\text{Struct}} + \omega_2 \cdot \text{sim}(c_1, c_2)_{-\text{Att}} \\ & + \omega_3 \cdot \text{sim}(c_1, c_2)_{-IC},\end{aligned} \tag{8}$$

where $\text{sim}(c_1, c_2)$ stands for the total semantic similarity between nodes c_1 and c_2, and c_1 and c_2 are the two nodes being compared. The adjustment coefficients $\omega_1 + \omega_2 + \omega_3 = 1$ of semantic distance, attribute, and information amount, respectively, are represented by the variable ω_i $(i = 1, 2, 3)$.

3.2.3. Evaluation Indicators. The mean absolute error (also known as MAE) is a benchmark that is utilized in the process of evaluating the quality of the recommendation. When the MAE is lower, the accuracy of the item prediction rating is higher. If we assume that the predicted data score set is $(p_1, p_2, p_3, \ldots, p_n)$, the matching actual data score set is $(q_1, q_2, q_3, \ldots, q_n)$, and N denotes the number of missing items in the score matrix, then the MAE is determined by taking the difference between the two score sets. The formula for the computation is as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{m} |p_i - q_i|. \tag{9}$$

The precision rate is an indication of the proportion of the positive examples that are actually positive examples or the proportion of the total number of grammatical errors that are correctly identified. This can also be stated as the proportion of the total number of grammatical errors that are correctly identified. When it comes to performance, precision is directly correlated to how well anything is performed. N stands for the total number, Ru for the list of u, and Lu for the set of sentences including u. With the set of sentences containing u, we mean the set of u from the test set whose highest score value is higher than its score mean. The formula for accurate computation is as follows:

$$\text{precision} = \frac{1}{N} \sum_{u \in U} \frac{|R_u I L_u|}{|R_u|}. \tag{10}$$

4. Results

This section focuses on the prediction accuracy of the similarity English grammar error correction algorithm (OUR) in this work, and comparing and analyzing other algorithms included in this study. The specific types of IB-CF included in this are IB-CF, support-vector machine (SVM), convolutional neural network (CNN), and random forest algorithm (RF). One of them is IB-CF, which is a classic item-based collaborative filtering method. This means it is a collaborative filtering algorithm that has not been improved in any way and simply uses the score matrix to compute the similarity prediction score. This algorithm determines the items to be clustered using the clustering approach described in this study, while the neighbor prediction scores are determined using the more traditional Pearson's method of calculating similarity. Figure 2 depicts the MAE results of our method.

Following the debugging of the algorithm parameters in this study through experiments, the algorithms discussed above are compared based on the two evaluation indicators of MAE and accuracy. The specifics of the investigation are shown in Figure 3.

Experiments on the processed dataset are carried out in order to verify the influence of different δ attribute distance thresholds on the algorithm presented in this study, and the results are shown in Figure 3.

On the basis of the determination of the number of neighbors and the number of clusters in Figure 3, it can be observed that, as the value of δ increases, the MAE of the algorithm first increases and then drops, reaching a minimum value when $\delta = 0.8$. Figure 2 illustrates this phenomenon. When the δ value is more than 0.8, the MAE begins to grow, making $\delta = 0.8$ the ideal threshold for this dataset. As a result, the technique described in this study requires $\delta = 0.8$.

Figure 3 shows that as the number of neighbors in different algorithms increases, the overall trend of the MAE value decreases, and it hardly changes once it reaches a certain value, whereas the MAE value of the OUR algorithm is the lowest under a different number of neighbors conditions, indicating that the OUR algorithm's prediction is accurate. As the number of neighbors in different algorithms increases, the overall trend of the MAE value decreases, at the highest level possible. The MAE values of SVM and CNN are lower than those of RF when the number of neighbors is varied, indicating that item attribute clustering is superior to the recommended technique of item score clustering and that the MAE of the IB-CF approach is superior to that of the CNN. The CNN technique outperforms the item attribute clustering method proposed in this study, indicating that the item attribute clustering method is superior. When different numbers of neighbors are considered, the MAE of the OUR algorithm is lower than that of the RF algorithm, indicating that the similarity optimization strategy proposed in this study can greatly improve the algorithm's accuracy.

Figure 4 shows that as the number of neighbors increases, so does the precision of each algorithm, and the system as a whole tends to become more stable after reaching

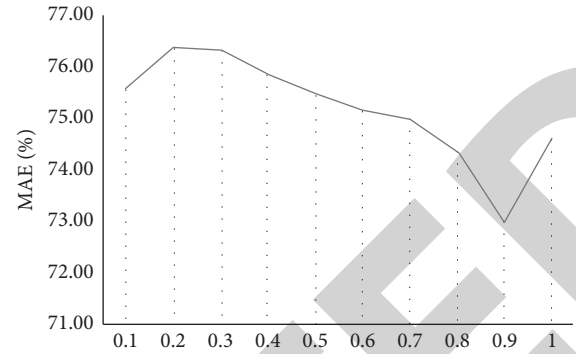

Figure 2: Results of MAE of our method.

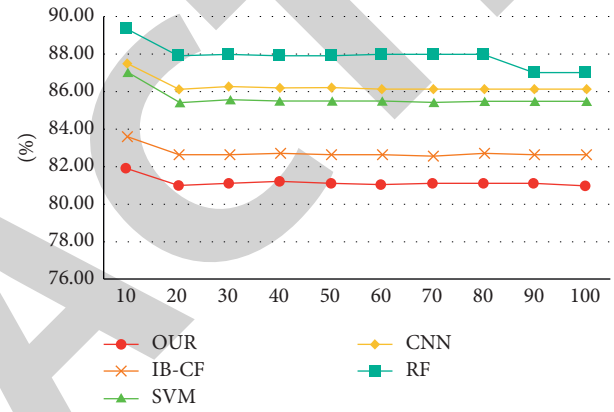

Figure 3: MAE comparison of different algorithms.

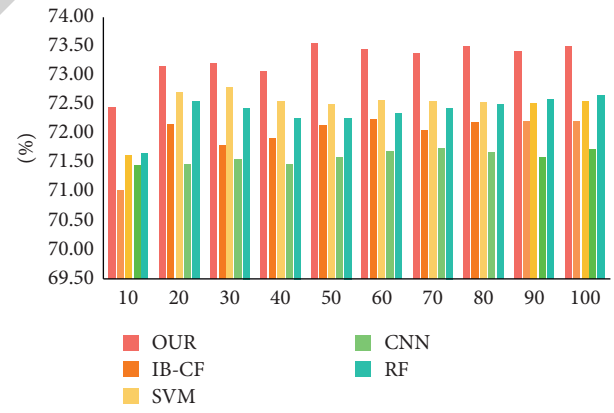

Figure 4: Precision comparison of different algorithms.

a certain maximum value. It has been demonstrated that the OUR algorithm predicts more accurately than other algorithms when the number of neighbors is 50 and that it is the most accurate when the number of neighbors is varied. This indicates that the OUR method predicts more accurately than other algorithms.

5. Conclusions

Because the statistic-based method is dependent on the corpus used, the calculation amount is large, the calculation method is complicated, and there are issues with sparse data

and data noise. The method based on semantic resources, on the other hand, is constrained by the semantic dictionary being used and cannot reflect objective reality. The most recent work in this field shows that using ontology knowledge to compensate for the issues that statistical algorithms face with data sparseness and data noise can result in more objective and accurate computation results. As background information, a sensible combination of a dataset corpus and a semantic dictionary can comprehensively consider a variety of semantic relationships between words. As a result, different types of semantic information can complement each other's advantages, resulting in improved precision in the results of word similarity calculations. The fundamental differences between algorithms based on statistics and those based on semantics extend to their underlying principles. As a result, additional research and practice for the various fusion technologies are required.

The practice of using a language is inextricably linked to everyday communication, which is why the goal of learning English should be to use it rather than to do well on a test. The original purpose of writing in English was to assess students' level of mastery at each stage of their English learning, to allow students to reflect on themselves, to identify and fill any gaps in their knowledge, and to gradually improve students' English learning level and mastery ability so that they could avoid the trap of Chinese English. We develop an automatic error correction approach for English writing grammar based on a similarity algorithm. This method accurately identifies errors, is useful for dataset classification, and investigates the grammatical errors of English writing that are included in the data. The calculation of our algorithm's accuracy in comparison to the accuracy of other algorithms demonstrates the superiority of the algorithm that we created. In the future, we will design automatic error correction systems for other languages.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] T. Winograd, "Understanding natural language," *Cognitive Psychology*, vol. 3, no. 1, pp. 1–191, 1972.

[2] K. R. Chowdhary, "Natural language processing," *Fundamentals of Artificial Intelligence*, vol. 2, pp. 603–649, 2020.

[3] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language toolkit [M]*, O'Reilly Media, Inc, Sebastopol, California, 2009.

[4] C. D. Manning, M. Surdeanu, and J. Bauer, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of the 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60, Spring USA, Naperville, IL, USA, May 2014.

[5] H. Pu, G. Fei, and H. Zhao, "Short text similarity calculation using semantic information," in *Proceedings of the 2017 3rd International Conference on Big Data Computing and Communications (BIGCOM)*, pp. 144–150, IEEE, Chengdu, China, August 2017.

[6] J. Wang, W. Xu, and W. Yan, "Text similarity calculation method based on hybrid model of LDA and TF-IDF," in *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, pp. 1–8, Association for Computing Machinery, New York, NY, USA, December 2019.

[7] G. Chen, X. Shi, M. Chen, and L. Zhou, "Text similarity semantic calculation based on deep reinforcement learning," *International Journal of Security and Networks*, vol. 15, no. 1, pp. 59–66, 2020.

[8] H. Christian, M. P. Agus, and D. Suhartono, "Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 7, no. 4, pp. 285–294, 2016.

[9] A. A. Hakim, A. Erwin, and K. I Eng, "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach," in *Proceedings of the 2014 6th international conference on information technology and electrical engineering (ICITEE)*, pp. 1–4, IEEE, Yogyakarta, Indonesia, October 2014.

[10] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, Spring USA, Naperville, IL, USA, 1999.

[11] 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.

[12] A. Pawley, "Some problems in proto-oceanic grammar," *Oceanic Linguistics*, vol. 12, no. 1/2, pp. 103–188, 1973.

[13] L. White, "Universal grammar: is it just a new name for old problems," *Language transfer in language learning*, vol. 5, pp. 217–232, 1992.

[14] G. Erbach and H. Uszkoreit, "Grammar engineering: problems and prospects," *CLAUS report*, vol. 5, p. 1, 1990.

[15] L. Cheng, P. Ben, and Y. Qiao, "Research on automatic error correction method in English writing based on deep neural network," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 2709255, 10 pages, 2022.

[16] S. Hasyim, "Error analysis in the teaching of English," *K@ ta lama*, vol. 4, no. 1, pp. 62–74, 2004.

[17] P. R. Bevington and D. K. Robinson, *Data Reduction and Error analysis*, McGrawHill, New York, NY, USA, 2003.

[18] M. H. Al-Khresheh, "A review study of error analysis theory," *International Journal of Humanities and Social Science Research*, vol. 2, pp. 49–59, 2016.

[19] S. Atkins, J. Clear, and N. Ostler, "Corpus design criteria," *Literary and Linguistic Computing*, vol. 7, no. 1, pp. 1–16, 1992.

[20] J. Zhang, "Data-driven teaching model design of college English translation using intelligent processing technology," *Wireless Communications and Mobile Computing*, vol. 20229 pages, Article ID 6559772, 2022.

[21] R. Reppen, *Building a corpus*, pp. 31–37, The Routledge handbook of corpus linguistics, England, UK, 2010.
[22] S. J. Gu, C. P. Sun, and H. Q. Lin, "Universal role of correlation entropy in critical phenomena," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 2, Article ID 025002, 2007.
[23] S. K. Lin, "Correlation of entropy with similarity and symmetry," *Journal of Chemical Information and Computer Sciences*, vol. 36, no. 3, pp. 367–376, 1996.
[24] J. Schindler, D. Šafránek, and A. Aguirre, "Quantum correlation entropy," *Physical Review A*, vol. 102, no. 5, Article ID 052407, 2020.
[25] D. Milne and I. H. Witten, "Learning to link with wikipedia," in *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 509–518, Association for Computing Machinery, New York, NY, USA, October 2008.