Research Article

A Grammatical Error Correction Model for English Essay Words in Colleges Using Natural Language Processing

Juan Long

School of Humanities, Hunan City University, Yiyang, Hunan, China

Correspondence should be addressed to Juan Long; longjuan@hncu.edu.cn

Received 11 May 2022; Accepted 30 June 2022; Published 13 July 2022

Academic Editor: Abid Yahya

Copyright © 2022 Juan Long. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Natural language processing technology is a theory and approach for exploring and developing successful human-computer communication. With the rapid growth of computer science and technology, statistical learning methods have become an important research area in artificial intelligence and semantic search. If there are errors in the semantic units (words and sentences), it will affect future text analysis and semantic understanding, eventually affecting the whole application system performance. As a result, intelligent word and grammatical error detection and correction in English text are a significant and difficult aspect of natural language processing. Therefore, this paper examines the phenomena of word spelling and grammatical errors in undergraduate English essays and balances the mathematical-statistical models and technology solutions involved in intelligent error correction. The research findings of this study are represented in two aspects. (1) In nonword mistakes, four sorts of errors are studied: insertion, loss, replacement, and exchange between letters. It focuses on nonword mistakes and varied word forms (such as English abbreviations, hyphenated compound terms, and proper nouns) produced by word pronunciation difficulties. This paper utilizes the nonword check information to recommend an optimum combination prediction method based on the suggested candidate list for actual word errors, and the genuine word repair model is trained. This approach is 83.78% accurate when used with actual words with spelling errors in the context. (2) It verifies and corrects sentence grammar using context information from the text training set, as well as grammatical rules and statistical models. In addition, it has investigated singular and plural inconsistency, word confusion, subject, and predicate inconsistency, and modal (auxiliary) verb errors. It includes sentence boundary disambiguation, word part-of-speech tagging, named entity identification, and context information extraction. The software for checking and fixing sentence grammatical mistakes presented in this article works on English texts with difficulty levels 4 and 6. Furthermore, this work obtains a clause correctness rate of 99.70%, and the system’s average corrective accuracy rate for four-level and six-level essays is more than 80%.

1. Introduction

Presently, English is among the most extensively spoken languages around the world, yet for some English speakers, their native language is not English. Numerous grammatical errors frequently arise in the writing process and using English due to the varying linguistic habits and cultural backgrounds of individuals in different areas [1]. Such grammatical errors will cause significant problems for users and data readers. However, if we utilize human mistake correction indiscriminately, there would be a lot of labor, making it impossible to satisfy the expectations of English users in the information era [2]. Unfortunately, the error repair capacity of English grammar in traditional English translation software is very poor, and an effective English grammar error correction system that can detect and correct must be created. The use of computer programs for automatic grammar correction of English can understand the practical importance of high efficiency and accuracy, and many language teachers have the task of correcting simple grammar to correct more difficult and complex pedagogical errors [3]. As computer technology and artificial intelligence evolve, computers have become one of the most effective and powerful tools for understanding and changing the world. One of the first difficulties to solve is how to successfully connect with computers for computers to perform human...
activities and needs. Natural language is the best way for humans to express themselves clearly [4, 5]. Natural languages include languages such as Chinese, English, and Korean. Besides, Natural language is an essential instrument for human communication, and thought crystallizes our collective wisdom. People have attempted to teach computers to read and understand natural language [6], that is, using natural language to communicate successfully between humans and computers. Those computers can comprehend and convey intentions and strategies in natural language writing which means that computers can understand concepts and meanings supplied by people in natural language [7]. In natural language creation, the first step is natural language understanding. It has three parts: language modeling, application algorithm design, implementation, and evaluation. The modeling description of language is the creation of a computer-processable mathematical model of natural language. The main objective of algorithm design is to turn a mathematical model into a model instance that a computer can use. With the use of suitable test indicators, application algorithms may be evaluated and assessed in terms of function and performance. Psychology, cognitive science, mathematics and logic, and computer science are all studied in natural language processing technology [8].

Based on current computer application theory and technology, natural language processing technology has advanced significantly in recent years. Semantic search engines, machine translation systems, text summarization systems, and question-answer systems are some examples of common uses. Among these, sound and text are the two most significant information transmitters in human cognition. Thus, natural language processing technology is separated into two main research branches: denoising, recognition, synthesis, and speech signal synthesis. While the primary goal of speech signal processing is to comprehend what is being said, it is also to process and evaluate what is being spoken. The text currently accounts for 70–80% of knowledge, according to [9]. Words are the most fundamental semantic unit in English text processing for speech and language processing. The precise spelling of words directly affects the subsequent text analysis and, ultimately, the overall performance of the actual application system. Improve students’ writing abilities so they can learn without teachers. This topic’s main objective is to explore “Key Technologies of Automatic Review and Scoring of College English Compositions” to intelligently repair word spelling and sentence grammatical mistakes in college essay writing. Inconsistent single and plural nouns, parts of speech confusion, inconsistencies in subject and predicate, and mistakes in modal (auxiliary) verbs are examples of grammatical errors. In addition to two essays in overseas journals and conferences in related areas, the system’s thorough design and code writing are now complete. Correction of word spelling and phrase grammar is frequently used in many fields. They underpin text and voice analysis studies, including part-of-speech tagging, named entity identification, semantic analysis, and syntactic analysis. Automatic question answering and text classification are examples of natural language processing systems that use this method.

Writing is a key indicator of English proficiency since English is the most widely spoken language. Using computer natural language processing to automatically correct word spelling and phrase grammatical problems in students’ writing skills would be extremely beneficial. This saves the instructor time and energy by allowing them to focus on the composition’s overall structure and narrative content. As computerized scoring technology for English essays improves in objectivity and accuracy, spelling and grammatical errors will become key assessment indicators. This paper examines the phenomena of word spelling and grammatical errors in undergraduate English essays and balances the mathematical-statistical models and technology solutions involved in intelligent error correction.

1.1. Offering of This Paper. The offering of this research work is listed as follows:

(i) This paper first recognizes and corrects nonword spelling problems in words, discovering that spelling errors in popular word-form terms are caused by resemblance and comparable sounds. Furthermore, the approach based on edit distance is used for words with an unusual morphology, taking into consideration the features of their writing form and vocabulary capacity. The accuracy and operational efficiency of the system are enhanced by classifying the form of words and executing various inspection and correction procedures.

(ii) The proposed method builds a genuine word disambiguation model by extracting contextual collocation characteristics and contextual word features and accomplishes automatic correction of real-word errors in English text using part-of-speech marked four-level and six-level model essays as a training corpus. The optimal combination prediction algorithm based on the recommended candidate list is adopted, which focuses on solving the true word check and error correction with contextual semantic errors.

(iii) Sentence boundary disambiguation is achieved by identifying the point in the sentence where the sentence boundary is likely to occur, and pattern matching is conducted on the predefined disambiguation rule based on the context elements in the sentence assumption boundary. The condition’s hypothetical border is removed, and the function of disambiguating sentence boundary is completed.

(iv) This paper examines and corrects grammatical faults in sentences. Furthermore, it proposes a method based on the combination of grammatical rules and statistical models to check automatically and correct widespread preposition errors in sentences, incomplete sentence components, inconsistent singular and plural nouns, word confusion, inconsistent subject-predicate and modal (auxiliary) verb errors, and other grammatical errors.
1.2. Organization of the Remaining Sections. Section 2 contains linked work of academics who studied our selected topic, Section 3 is based on my suggested approach for the grammatical error repair model, Section 4 covers our experimental work during our study and its outcomes, and Section 5 is the paper’s conclusion.

2. Related Work

Checking and correcting spelling errors is now an important part of text processing. This technology is not only used in commonly typed text checking but also widely used in Optical Character Recognition (OCR) and Online Handwriting Recognition [10]. The spell check of a word can be used as a separate functional unit to process text strings, or it can be embedded into the system as a part of the application function. Foreign countries have begun to use computer natural language processing technology for English language processing as early as the 1960s. Spelling errors in the text were checked and corrected. The IBM Thomas J. Watson Research Center first implemented a TYPO English spelling checker on IBM/360 and 370 using UNIX systems in 1960. In 1971, Ralph Gorin of Stanford University A spell checking program Spell [11] is implemented on DEC-10. GNU Aspell, AGREP, and the spell checker integrated into the commercial word processing software Microsoft Word have also received extensive attention and research in recent years [12]. In addition, in the search field, to improve the understanding of search engines and the effectiveness of the query, the spell check of the input keywords has become a key part of search engine technology. Spelling checking of words in China started late, but it has developed very rapidly. Reference summarized the types of errors in word spelling and related methods for these errors, focusing on the advantages and disadvantages of different methods for discussion and analysis. The authors of [13] designed and implemented an English composition spell check system to check the text for nonword and real-word errors. In addition, the check of true word errors is realized by a method based on mathematical-statistical models. As for the current mainstream word correction methods, there can be two main categories: automatic correction and interactive correction. However, there are still many areas for improvement in the current spell check system. Specifically, it can be investigated from three aspects: the accuracy of the word error detection and correction algorithm, the execution speed, and the scope of application. Therefore, based on previous research results, it is very valuable to summarize the shortcomings of existing methods and explore more effective solutions.

At present, the technology for automatic grammar checking of English sentences is mainly based on rule-based analysis methods and statistical-based analysis methods. The idea of the rule-based analysis method is based on the theoretical basis of linguistics and behavior, through the artificial construction of grammatical rules and the establishment of a grammatical rule database, and then the syntax is checked according to different constraints. Its advantage is that it can be based on specific purposes and needs. The construction of rules has strong pertinence; the statistical analysis method is based on the training results of the large-scale corpus, and the mathematical-statistical model is built through the method of machine learning to check the grammatical errors of the sentence. It overcomes the rule-based problem of limited grammatical coverage and conflicting rules that exist in the method [14]. Statistical models also have problems such as dependence on the corpus and possible data sparseness. Foreign countries have conducted extensive research on grammatical errors in English sentences. ICICLE system uses the constructed grammatical error recognition rules to detect the grammatical abnormalities in the sentence and gives relevant prompt information [15]. The scholar in [16] used the N-ary grammar model to check grammatical errors, established a list of candidates’ recommended words, and prioritized the candidate words according to random context-free grammar. Similarly, in [17] the authors focused on the study of preposition errors, using the British National Corpus (BNC) as the training set, extracting the context of the prepositions. After that, it established feature vectors and applied them to the generated maximum entropy model to make preposition errors check. Among the contextual features of prepositions, the contribution of contextual word features in the extraction window is greater than that of collocation features, named entities, etc. [18]. In [19], the authors proposed a method based on examples and the introduction of negative rules to achieve grammar checking. The research found that 55% of sentences with grammatical errors are local grammatical errors and 18% are global grammatical errors. In English writing, different students have different professional backgrounds and knowledge levels, which make the types and reasons for grammatical errors also diverse. A good grammar checker should not only make full use of contextual structure information and lexical information. As well as semantic information for grammatical inspection, it is also necessary to have better robustness to adapt to the input of different sentence patterns. There are still many areas for improvement and perfection in this area in the future. Therefore, this paper examines the phenomena of word spelling and grammatical errors in undergraduate English essays and balances the mathematical-statistical models and technology solutions involved in intelligent error correction.

3. Methodology

3.1. The Importance of Grammar Checking Research for College English Essay Words. Because of the tremendous elasticity and ambiguity of ordinary language, English is a representative example of various dictionaries, complicated grammar, and broad use situations, which makes computer automated mistake detection and correction more challenging. Another significant factor influencing the advancement of grammatical mistake corrections is a lack of relevant material. It is quite hard to build a library filled with grammatical mistakes. The current major grammatical error repair research approaches are all based on statistical computer vision, which needs a huge corpus for model
training and validation [20]. Nevertheless, with colleges and research institutes focusing on this problem, the absence of literature has been considerably remedied, creating a solid foundation for future study.

### 3.2. Framework Design of English Grammar Check for College English Essay Words

The design of the grammatical error-correcting system model [21] is depicted in Figure 1 based on the findings of the functional application of analytical needs. The primary unit of grammar error detection and correction consists essentially of three functional units: computational modeling, training of model, and model error detection and correction, with model error correction helping as the algorithm’s basic function [22]. The primary objective of data handling is to preprocess the actual corpus data, arrange the treated corpus data, train the unit, and obtain the typical dataset. Training of model entails training the information in the corpus, storing the learned characteristics in the database, and using them in subsequent testing and comparison [23]. The error correction system maintained in the training database is used to tie the input sentence syntax and production of the right sentence. The error correction facility layout may receive the user’s error correction demand in real time, evaluate it using the corpus’s error correction mechanism, and provide the right materials to the user.

### 3.3. Employment of Grammar Error Correction

First, learning is performed under the features of grammatical error repair, and a user-submitted English mistake correction application is obtained. Verify whether the given parameters are acceptable before proceeding to the next phase to divide the sentence. The previously trained error correction model is then utilized to repair grammatical errors. If the last sentence’s mistake correction is finished, the error repair sentences that were resumed to the division are combined. If the phrase is simple enough, the error correction layout can be applied immediately without any need for sentence splitting. Whenever the user is dissatisfied with the system’s grammatical error repair or if there is an improved method to fix it, the alteration proposal is supplied back to the model. As previously stated, we will filter user-submitted change recommendations; thus the prior feedback proposal filtering methodology will be applied in the feedback proposal function. We develop feedback ideas from two perspectives, similar to how we repair grammatical errors. The first is the feedback filtration interface itself, which has a work flowchart. The other is called flow between units, which is illustrated via sequence diagrams. The feedback filtration interface is introduced first. Verify whether the query parameters are lawful, and then finish immediately, as per the syntax error repair method. For the error correction mechanism, the chances of the error correction report and the actual scheme change statement are determined.

### 3.4. Analysis of Grammar Check and Correction Methods

An important use of natural language processing technology is the automated assessment and repair of English sentence grammar. It can assist students, particularly those whose first language is not English, in mastering English writing. As a result, both local and international experts have conducted extensive research on this topic. Some useful study findings have been obtained. The following is a brief overview of some of the approaches used.
3.4.1. Method Based on Syntactic Analysis Tree. The method based on a syntactic analysis tree analyzes the components of the sentence to be checked in a bottom-up or top-down manner and detects grammatical errors in the sentence. This method uses the data structure of the stack to match the syntactic structure. A sentence is considered a grammatical error if the grammatical analyst is unable to evaluate the whole match of the sentence according to grammar rules or constructs an error tree. Because more than one syntactic tree is frequently formed while parsing some ambiguous words, several probability statistical techniques are included in determining the likelihood of the generated syntactic tree during syntactic parsing, therefore enhancing the accuracy of syntactic analysis. The disadvantage of this method is that it requires a large number of traversal and stacking operations on the sentence components. As the length of the words in the sentence increases, the time and space overhead of the system at runtime will increase sharply, which greatly reduces the syntactic analyzer. Moreover, due to the flexibility of English sentence writing, it is difficult for syntactic analysis rules to cover all kinds of different writing situations, which can easily lead to error analysis of sentence components and cause false positives. Furthermore, if the syntactic analyzer discovers grammatical faults when analyzing sentence components, providing efficient grammatical correction procedures to remedy grammatical errors becomes challenging.

3.4.2. N-Gram Grammar Model. An N-gram is a continuous sequence of n elements from a particular sample of language or speech. Depending on the application, the elements might be letters, sentences, or base pairs. N-grams are often extracted from a textual or audio corpus. An N-gram language model, on the other hand, predicts the likelihood of a word given the previous N−1 words, where feature represents the cooccurrence probability of feature and target word, respectively. Mutual information can be used to measure the degree of correlation between two events. Therefore, the following formula is defined to calculate the correlation between contextual features and the current word:

\[
\text{Relativity (feature, target word)} = \log \left[ \frac{P(\text{feature}, \text{target word})}{P(\text{feature}) \times P(\text{target word})} \right]
\]

where feature represents the context feature, target word represents the current word, \( P(\text{feature}, \text{target word}) \) represents the cooccurrence probability of feature and target word, and \( P(\text{feature}) \) and \( P(\text{target word}) \) represent the respective probabilities of feature and target word, respectively. Through the study of mutual information, the following conclusions are drawn:

(i) If \( \text{Relativity (feature, target word)} > 0 \), it indicates that feature and target word are positively correlated, and there is a clear and credible combination relationship between the two, and the combination degree of the relationship between the two increases with the value of Relativity (*) and enhanced.

(ii) If \( \text{Relativity (feature, target word)} = 0 \), it indicates that the combination of feature and target word does not have a clear and credible relationship.

(iii) If \( \text{Relativity (feature, target word)} < 0 \), it indicates that feature and target word are negatively correlated, and the two have a weak degree of combination, and the possibility of cooccurrence is very low, and the degree of combination of the two relationships increases with the value of Relativity (*).
4. Experimental Work and Result

4.1. System Overall Design. The two primary research objectives of this paper are spelling errors in words and grammatical problems in sentences. The basic unit of study for words is letters, whereas the basic unit of research for sentences is words. That is, word research is synonymous with sentence research. The development of word research is one of the foundations of sentence research. For English composition, it can record the content of writing through keyboard input and written form. The misspellings of words produced by them in writing are not the same. The key study presented is in textual form to make the problem more relevant. From the analysis in the previous chapters, we can see that the detection and correction of nonword errors need to consider the diversity of word forms and various reasons for nonword spelling errors. True word errors should be detected and corrected using context and semantic accuracy. There are two types of spelling errors in context. In the study of grammatical error correction, the object of this article is college English essays. As a result, the sentence’s boundary is derived by implementing boundary disambiguation rules. The technique is targeted at correcting pupils’ common grammatical errors after assessing their common grammatical errors. Preposition errors, incomplete sentence components, inconsistencies in single and plural nouns, word confusion, subject-predicate inconsistencies, and errors in modal (auxiliary) verbs are among the most significant forms of faults that have been verified and corrected. Therefore, based on the above analysis, the overall system inspection process is shown in Figure 2. The flowchart is divided into two sections: spelling processing and grammatical processing. The spelling processing module includes nonword error processing and true word error processing. The grammar processing module first needs to disambiguate the input English text and then uses a combination of statistical models and rules to check for common grammatical errors.

4.2. System Experiment and Analysis

4.2.1. Experimental Data Set. The test content in this section includes three parts: nonword error handling, true word error handling, and sentence grammar handling. Here, English texts of different levels of difficulty are selected to test the impact of the selection of training corpus on the construction of the dictionary. The non-English major college students’ essays (ST3 and ST4) in the Chinese learner corpus are used as test examples of nonword error processing and sentence grammar processing. Table 2 explains the main data source of the test error text. According to this table, four topics were selected, 120 pieces in total. These 120 essays have nonword errors and grammatical errors to varying degrees and have been manually marked with errors. In addition, we extracted from the 4th and 6th level exercises some commonly misspelled words that are used as supplementary test examples for nonword error handling. For the true word error test, the remaining 20% of the sentences related to the confusion set in the training corpus are used as test examples.

Table 2: The main data source of the test error text.

<table>
<thead>
<tr>
<th>Essay topic</th>
<th>Source</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practice makes perfect</td>
<td>ST3</td>
<td>30</td>
</tr>
<tr>
<td>Global water shortage</td>
<td>ST3</td>
<td>30</td>
</tr>
<tr>
<td>Our view on job</td>
<td>ST4</td>
<td>30</td>
</tr>
<tr>
<td>Our view on environment</td>
<td>ST4</td>
<td>30</td>
</tr>
</tbody>
</table>
4.2.2. Test Results and Analysis. Choose three types of training texts with different levels of difficulty (elementary essays for elementary school students, sample essays for grades 4 and 6, and essays selected by Reuters). For each type of text, find out the accuracy of correcting nonword errors of common morphology under different text quantities, and then compare which of the three types of text can obtain the best value. Obviously, at this accuracy rate, we match the text type and text quantity for the selection of dictionary training text. Figure 3 depicts the particular experimental findings.

In Figure 3, it can be found that when four-level and six-level model subjects are selected as a training text, the effect of correction is better than the other two types of text. Furthermore, the accuracy of the correction reaches a maximum value of 75% when the number of texts is 80 as it meets the fourth and sixth levels. The breadth and habit of writing functional words, as well as statistical word frequency features, may be effectively used for spell checking short essays produced in grades 4 and 6. Therefore, this system uses four-level and six-level sample essays as training texts, the number of which is 80, which is used for word frequency statistics in ordinary morphological dictionaries. The effect of the model essay on elementary school students in the test is the most unsatisfactory. The main reason is that the words used by elementary school students are too simple and single. As for Reuters’ selected essays, the wording is biased toward the writing habit of using English as the mother tongue, which is still somewhat different from writing in levels 4 and 6 writing. Based on the above analysis and considering that the main research object of this thesis is college English essay writing, it is reasonable to select the fourth and sixth level model essay as the training text.

4.2.3. Syntax Error Handling

(1) Sentence Boundary Disambiguation. In order to check and correct the grammatical errors of sentences, one of the primary issues is to process the text sentence. This system uses a rule-based sentence method. By adding hypothetical boundaries to the input text, then use the method of rules to realize the boundary disambiguation and realize the function of clauses. The test examples are derived from 120 articles in the experimental data, and the specific punctuation distribution is shown in Table 3.

From Table 3, we can see that if the question mark, exclamation mark, and period are all directly used as the sentence boundary, the sentence accuracy rate is only 66.62%. The result of the sentence is the basis of sentence grammatical error analysis; therefore, the assumption of sentence boundaries is a necessary condition for disambiguation. The following is the sentence result of the test case based on the method of boundary disambiguation rule adopted by this system. Table 4 explains rule-based sentence boundary disambiguation results.

From the above results of disambiguation, it can be known that the clause accuracy rate is 99.70%, and the boundary disambiguation effect is obvious.
(2) Sentence Grammar Check and Correction. Through researching and analyzing the grammatical errors in college grade four and six essays, this system uses a combination of N-gram grammar models and artificial grammar rules to focus on solving common grammatical errors in writing. We extracted 110 test cases in which these grammatical errors occurred from the text of the experimental data. Some of the test cases contained more than one type of grammatical error. In the test, we mainly examined the system’s error detection and error correction accuracy of sentence grammatical errors. The specific test results are shown in Figure 4.

From the data in Figure 5, it can be seen that the average correction accuracy rate of this system for four-level and six-level essays is over 80%. Through the observation of the data, the system is in the error detection and error detection of the two types of errors in preposition errors and inconsistencies in the singular and plural of nouns. The accuracy of the correction is not high enough.

5. Conclusions

With the rapid development of the world economy and technology, international communication is increasing day by day. In this era of development, English plays a pivotal role in communication between people because English writing is a very effective way of communication. If there are spelling errors and grammatical errors in words, it will greatly reduce the reading efficiency. Therefore, if natural language processing technology is used, the spelling errors and grammar errors that occur are automatically corrected by computers. The automatic inspection and correction of errors have become a major problem in text research. Based on the research of "Key Techniques for Automatic Criticizing and Scoring of College English Compositions," this paper takes English assignments of level 4 or 6 as the main research object, focusing on the analysis of nonword errors, true word errors, and grammatical errors in student writing. An in-depth discussion on related issues is carried out. The test results of the system experiment show that the system has a certain application value. By categorizing grammatical errors, summarizing their respective characteristics and reasons for errors, and combining them with previous research results, the system can provide suitable checking and correction strategies for different grammatical errors, achieving the desired effect.

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 he has no conflicts of interest.

Acknowledgments

The paper was supported by Hunan Social Science Achievement Evaluation Committee (No. XSP22YBC531).

References


