Exploration of English Composition Diagnosis System Based on Rule Matching

https://doi.org/10.3991/ijet.v13i07.8603

Xurong Xu
Yancheng Vocational Institute of Industry Technology, Jiangsu, China
xxr5690846@126.com

Abstract—To discuss the grammar checking system and to solve the low accuracy of the system and bad effect, the grammar checking algorithm and the accuracy were improved. On this basis, a grammar checking system was designed and developed. The grammar checking module adopted the grammar checking design based on multi-rules and used multi-layer grammar rules to provide error correction function for common English grammar. To improve the grammar checking accuracy of grammar checking module, the grammatical model based on statistics was introduced by the system. In addition, N-gram model was applied to evaluate the errors found out by the rules grammar. The experimental results showed that the grammar checking system carried out grammar check of domestic students' English composition samples, and the system performance was good. It is concluded that in asynchronous mode, hundreds of sentences can be checked at the same time, and the purpose of the design is achieved.

Keywords—Grammar check, rule grammar, English composition, system design

1 Introduction

China has a large population, and every year, many students come to study English. At the same time, China is also a country with a shortage of educational resources, and the proportion of teachers and students is relatively different from that of developed countries. Every year, many English exams need to be corrected in terms of English composition, which brings a great deal of work to teachers. English grammar is not easy to grasp. If it is expected to skillfully use, it needs for regular writing training. Because the teacher's energy is limited, they often cannot timely feedback the works of students. As a result, they need to conduct the students' English writing grammar checking by computer. At present, the Chinese market has no tools that can effectively check the English compositions, so constructing a suitable auxiliary English composition teaching used by the Chinese educators will bring great benefits to the teaching and research of Chinese English language.

The main work of this paper is to provide the support on grammar check for the English composition assistant marking tools. As everyone knows, English grammar is the framework of English composition, which is also an important reference index in the process of marking English composition. The overall quality of English composition is directly influenced by the mastery of the students' English grammar. Therefore, to provide syntax checking tool can help English essay marking researchers find the grammatical problems in English composition, and it gives a reasonable proposal. It will help reducing the required workload of marking the composition, assist reviewers evaluation of English writing level, and reduce the workload of marking researchers. According to the above requirements, supplementary grammar tools for English grammar mistakes check should be designed and implemented in the process of English writing correction. The purpose is not only to find out whether there is a syntax error, more importantly, to pinpoint specific grammar mistakes made in English composition, for marking reference. English grammar is complex and changeable, and the level of English composition of Chinese students is uneven, which has brought difficulties to the grammar examination. Therefore, we need to redesign the grammar checking system, combine various kinds of grammar checking models, and improve the overall effect of the system, so as to achieve the purpose of grammar checking, and design a more suitable grammar checking system based on this purpose.

2 Literature review

Currently, grammar checkers are mainly based on rule models. Due to the complexity of English grammar, regular grammars cannot express all the grammatical errors. But the rule model is the first choice for all kinds of grammar checkers because of its simple design, intuitive and easy to use. Statistics-based grammar checking has also been the focus of research. In recent years, the related algorithm models tend to be perfect day by day. Google and other companies have provided a good corpus for N meta syntax, convenient for related research. The following is the introduction to several related research projects at home and abroad.

CLAWS (Constituent Likelihood Automatic Word-tagging System) is a kind of words tagging based on statistical model. Doerfel and others as stated in [1] proposed the syntax checking method based on words tagging. His idea is to get a word label and the word tagging of words that is adjacent to it, and then to calculate the probability of the sequence that they make up. If the probability value obtained is less than the predefined threshold, then the advantage of the method with error base statistics is that the handwritten rules are not required.

FLAG (Flexible Language and Grammar Checking) is a platform for specific users to perform grammar checking programs. Dale and others as stated in [2] proposed a grammar checking system used for the wrong grammar during writing process. The system provides different components for formal analysis, words tagging, dicing and topological parsing. FLAG uses trigger rules to detect errors, which are used to find hidden grammatical errors in English text.

In 2016, Shi and others as stated in [3] proposed a Raman spectroscopy for early real-time endoscopic optical diagnosis based on biochemical changes. The diagnosis method is used to check the grammar error and correct it immediately. In 2017, Cheng and others as stated in [4] proposed a ranking causal anomaly for system fault diagnosis via temporal and dynamical analysis on vanishing correlations. The propose is to determine the little mistake that students can't easily find. In addition, Han and others as stated in [5] proposed a fault diagnosis system used for web service composition. The diagnostic system helps teachers to easily and quickly modify their students' compositions.

For the present research results, the following features can be summed up.

First, with the development of corpus linguistics and the construction of corpus, the main goals of Natural Language Processing are gradually transferred to the processing of large and real text. Gabrilovich and Markovitch as stated in [6] discussed the Wikipedia-based semantic interpretation for natural language processing. Therefore, a reasonable, suitable and large-scale corpus for research and application should be established.

Secondly, the method of machine learning is more and more used in the process of Natural Language Processing. Murata and others as stated in [7] studied the machine learning and various textual features in language use process. In addition, the machine learning algorithm that is proposed by Sengupta and others as stated in [8] are fully considered. The further development of machine learning algorithm has changed the traditional way of research in the field of Natural Language Processing. More and more researchers have invested in the research of machine learning algorithm applied in grammar checking. According to the above analysis, syntax checking system in this paper also uses a variety of model fusions and adds the relevant syntax check module automatically based on machine learning, to ensure the error detection of syntax checking system. the Kelly and others as stated in [9] used the repetitive error detection in a superconducting quantum circuit and had a good result. Ferraro as stated in [10] studied the improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation.

To sum up, the existing research still has a system inspection loophole in the text matching of the composition diagnosis system. To solve the problem of grammatical error correction, a detailed analysis based on the rule matching is designed. The specific process is: redesign the grammar checking system, combine various kinds of grammar checking models, and improve the overall effect of the system. Therefore, the purpose of grammar checking is achieved and a more suitable grammar checking system based on this purpose is designed. This method makes up for the shortcomings of the current research on the composition system matching and has the advantage of correctly searching for grammatical errors and correcting the errors accurately.

3 Syntax checking based on multi-layer rule model

3.1 Multi-rule grammar check

The major work of rules grammar check system mainly included, punctuation, word segmentation, tagging, cutting and other pretreatment process of the text, as well as the design and implementation of rule engine.

The grammar checking module is made up of two parts, which are rule-based module and statistical-based module. In this paper, the combination of two parts is used to make a grammatical check of input sentences. The architecture diagram of this module is shown in Figure 1. The whole design is made up of two parts. Text input is entered through the rule grammar module. Firstly, the rule grammar module is evaluated, and the result is fed back to the probability module. Then, the score is calculated by the probability module, and finally whether to modify it is determined.

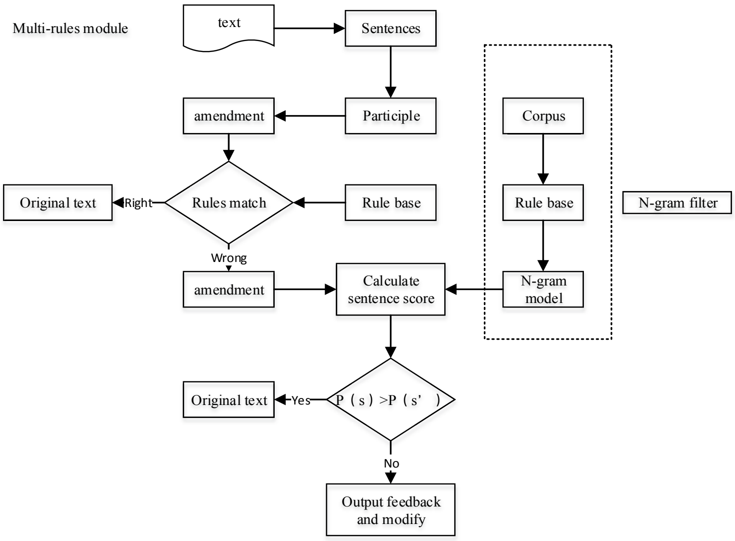

Fig. 1. General design diagram of modules

Before making a grammatical check, a sentence is first tagged with parts of speech corresponding to each word, such as verbs, nouns, adjectives, adverbs or other parts of speech. Different corpus uses different mark set conventions to mark words.

Tagging is usually based on rules and statistics. The common annotation algorithm is Markoff chain, HMM and so on. Because the emphasis of this article is not on the implementation of part of speech tagging, this section does not elaborate in detail. In this paper, we use the open source word tagging. Here, the Stanford parser is used to complete the part of speech tagging process.

3.2 Rules expression

The construction of the rule base is mainly divided into two parts, the collection of rules and the arrangement of the rules. In the process of rule grammar design, the construction of the rule base is the most basic and core work of the rule base.

The current system finishes 1083 grammar rules, mainly from three aspects. One is to learn from other rules grammar checker. Chinese Students Grammar Check System designed. Based on the rule base, in accordance with the format of this system, we complete summary into the system. Secondly, through English Linguistics and education related books, we select the representative books, such as "Bobing English Grammar", "English Grammar Problems Detailed Explanation" and "College Students' English Writing Guide". Through the summary of common errors in English grammar and common classification analysis, we classify, analyze, and sort out more than 300 grammar rules. Thirdly, we search common mistakes from the annotated corpus, and summarize them to find a more representative error, as well as the grammatical errors that are often used wrongly. Moreover, they are incorporated into the rule base as rules. The description of the rules is stored in the form of XML files.

Because most rules in the rule base are manually arranged, although they are highly generalizable, there is no relative conflict between the rules. Therefore, after the establishment of rule base, we need to check the conflict of these rules, including two aspects: first, effectiveness validation; second, mis-judgement verification.

The main purpose of validity verification is to verify whether the rules used can be reasonably detected. For example, if a rule is added, if the rule is correct, the corresponding errors will be identified by the system. To achieve the test of validity, we need to add artificial test cases according to the characteristics of rules. In addition, many test cases are used to describe all kinds of application scenarios of the grammar rules as much as possible. The actual detection results of these rules are evaluated. In the actual work of this article, two to three test cases are usually used for key rules. If the detection rate reaches 80%, it is considered that the rule is effective and can be used. The rules less than 80% are adjusted to find out whether there is a problem of conflict of rules.

On the one hand, it is for rule to judge whether there are errors in the rule, and on the other hand, it also judges whether there are rules failing to point out the wrong cases correctly. The scheme adopted in this paper is solved through a variety of model fusion methods. When making rules, it can be evaluated through the N-gram model, and we can quickly find out the grammar rules with insufficient accuracy. After finding these rules, we will analyze the reasons one by one, and constantly improve the rules.

Table 1. Grammatical structured

Number	Grammatical structured
1	A verb phrase consisting of an auxiliary verb, a verb, and an adverb.
2	A noun phrase made up of articles, pronouns, numerals, adverbs, adjectives, nouns, and the structure of OF
3	An adverb phrase made up of a preposition and a noun phrase
4	A non-finite verb phrase consisting of a noun phrase and a non-finite verb
5	A non-finite verb phrase consisting of an adverb phrase
6	An adverb phrase made up of a prepositional addition phrase
7	A complex noun phrase consisting of a non-finite verb phrase, an adverb phrase, and a noun phrase they modify.
8	Simple sentence
9	Attributive clause, adverbial clause, and noun clause
10	Compound sentence of subject and subordinate and parallel compound sentence

Grammatical errors in English are often embodied in the structure of the sentence. For example, the most common errors of subject and predicate in English sentences, we need to analyze sentence structures, find out the subjects and predicates of sentences, and determine whether they are unified in single and plural form. Some grammatical structures in English sentences are very simple, while some are complex. In some cases, complex grammatical structures include simple grammatical structures.

Based on the above theory, a multi-layer rule model can be designed. A rule base can be represented as a XML file with multiple rules. The rule base is stored in a sequential way, that is, the rules are read in order each time. In this paper, the multi-layer analysis error mechanism is used to arrange the rules according to the grammatical level, and the rule base design can be represented as figure 2. The details of each layer of the rule library are shown in table 2.

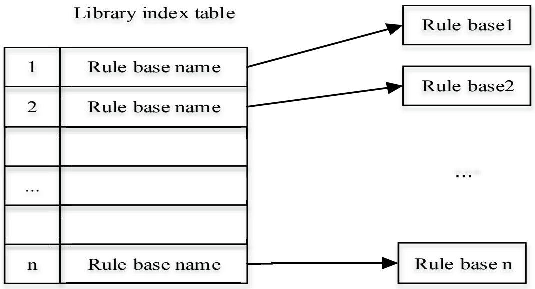

Fig. 2. Rules design diagram

Table 2. Rules content in each layer

Number	Grammatical structure
The first layer	The analysis of predicates with obvious characteristics, such as the features of an auxiliary verb, a modal verb, a tense change, and so on
The second layer	A simple noun phrase, namely a noun phrase that contains only a premodifier, such as an article, a pronoun, an adverb, an adjective, and so on
The third layer	Simple noun phrases for OF connections
The fourth layer	A simple adverb phrase expressing frequency, degree, emphasis and so on
The fifth layer	An adverb phrase composed of a preposition and a simple noun phrase, expressing means, location, time and so on
The sixth layer	Non-finite verb phrases consist of non-finite verb phrases and two, three, four or five layers analysis results
The seventh layer	A complex word phrase, modified by an adverb phrase or a non-finite verb phrase, or connected by OF
The eighth layer	A complex adverb phrase, mainly consisting of the previous analysis results
The ninth layer	A complex verb phrase, consisting mainly of the previous analysis results
The tenth layer	Attributive clauses, adverbial clauses, noun clauses and their predicates
The eleventh layer	A grammatical element, such as a noun or a verb modified by a clause
The twelfth layer	According to the basic structure of the sentence to determine the sentence composition C of main clause

The input parameters of the algorithm are the rule library index table and the current level number. The index table is used to provide index entries to facilitate the retrieval, and the current level number refers to the level used in the current inspection.

The rules are classified according to their grammatical problems. Different categories of grammar rules are imported into a rule base. When searching, the system searches one by one from top to bottom. Therefore, the rule base scheduling algorithm uses a similar multi-level structure and searches at the level from top to bottom. The rules of each layer are matched to the next level until the search is completed. Its algorithm flow is shown in figure 3.

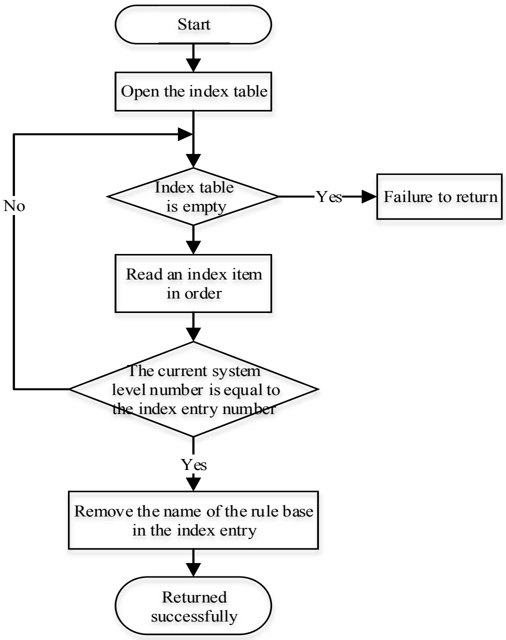
Start
Open the index table
Index table is empty
Yes
Failure to return
No
Read an index item in order
The current system level number is equal to the index entry number
Yes
Remove the name of the rule base in the index entry
Returned successfully

Fig. 3. Rule scheduling algorithm flow chart

The above rule base system is applied to design the grammar checking module based on rules. After a statement is preprocessed, a rule check is carried out. According to the hierarchy of the rule base, the search is carried out by layer, and return to the errors that are checked every layer. According to the actual effect, the design can effectively return the grammatical errors in the statement. The overall flow design is shown in figure 4. The system is designed to conform to the general checking process of regular grammar. In addition, multiple layer design is adopted to reduce the interact influence of syntax between different levels, aiming to reduce the conflict between rules.

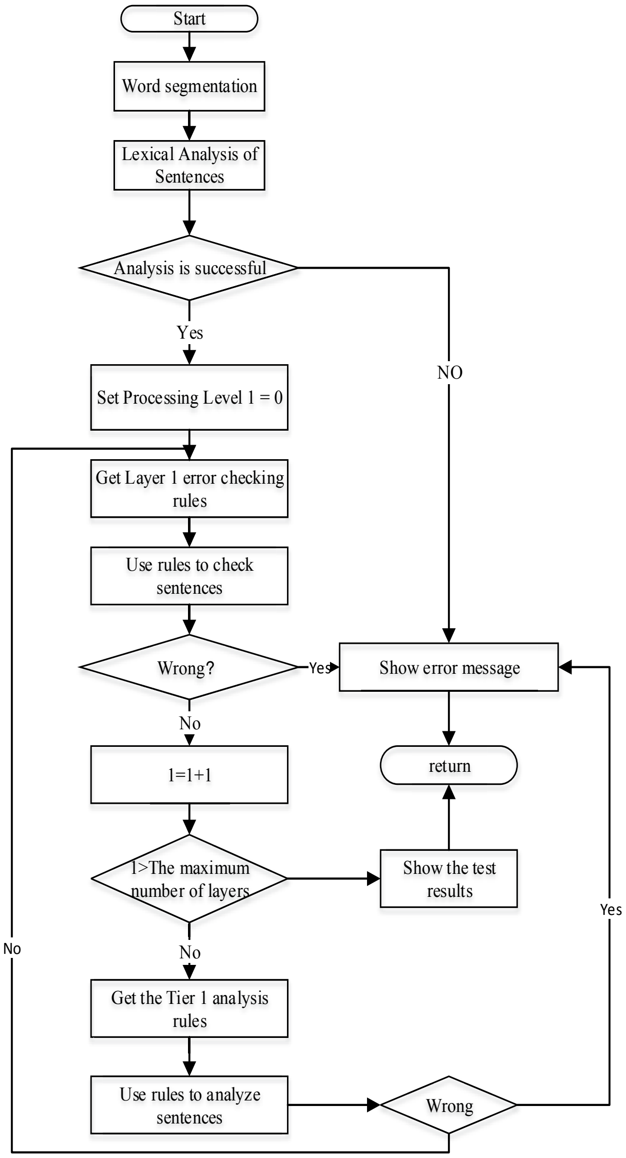
Start
Word segmentation
Lexical Analysis of Sentences
Analysis is successful
Yes
NO
Set Processing Level 1 = 0
Get Layer 1 error checking rules
Use rules to check sentences
Wrong?
Yes
Show error message
No
1=1+1
return
1>The maximum number of layers
Show the test results
Yes
No
No
Get the Tier 1 analysis rules
Use rules to analyze sentences
Wrong

Fig. 4. Multilevel rule grammatical error correction system diagram

3.3 Grammar error detection filtering based on N-gram model

The N-gram grammatical model uses statistical ideas to evaluate the rational existence of a sentence. Input statements are usually processed as strings, and a sentence entered is converted to string S. The N-gram model constructs the probability distribution P(S) of the string S. The probability of the occurrence of string S as a sentence can be expressed by the value of P(S).

If there are 100 statements in the corpus, where the number of name appears is 1, then the probability of the word name occurrence can be equivalent to 1%. Sentence S is made up of the words w_1, w_2,..., and w_n, and then the probability of the occurrence of a string S can be expressed as:

$$P(S) = p(w_1)p(w_2|w_1)p(w_3|w_2 w_1)...p(w_n.|w_1...w_{n-1}) \tag{1}$$

Here, the probability of the occurrence of the i-th word is determined by the i-th word before this word, and usually the former i-th word is called the history of the i-th word. The calculation amount of this method increases exponentially with the increase of i. In this paper, the Markoff chain is used to solve the above problem. The core idea is on depending on the limited word occurring before it.

If the occurrence of a word only depends on the word before it, this belongs to biagram model, namely:

$$P(S) = p(w_1)p(w_2|w_1)p(w_3|w_2 w_1)...p(w_n.|w_{n-1}) \tag{2}$$

If the occurrence of a word only depends on the two words before it, this belongs to trigram model. To ensure the division significance of $p(w_i|w_{i-1})$, the sentences need to add the beginning and end marks. Taking "Today is a good day" as an example, the probability mode of this sentence can be expressed as:

$$P(S) = p(\text{today}| < \text{BOS} >) * p(\text{is}|\text{today}) * p(a|\text{is}) * p(\text{good}|\text{a}) * p(\text{day}|\text{good})$$
$$* (< \text{BOS} > |\text{day})$$

To calculate the probability of the whole sentence, it is necessary to calculate the probability of each condition. The product of the N element grammar can be obtained by the following formula.

$$p(w_n|w_1...w_{n-1}) = \frac{p(w_1...w_n)}{p(w_1...w_{n-}1)} \tag{3}$$

The most commonly used in practice is biagram and trigram, and what are more than four element grammars are rarely used. Because for more than four element grammars, the time complexity is high, but the accuracy is not much higher. In the concrete implementation, this paper uses the N-gram module provided by the SRILM tool to train the Wiki corpus, and then evaluates the input sentences.

4 Evaluation and verification of error detection effect

4.1 Evaluation content

There are many aspects of English grammar, such as verb tense error, verb deletion, subjective-verb agreement, article and preposition error and so on. There are also many special uses, as well as the existence of idioms. Therefore, it is difficult to completely define all the categories involved in English grammar. Of course, in grammar check research, we have summarized and sorted out common grammatical problems, involving 28 kinds, and complete contents are not described.

To ensure the rationality of evaluation, this paper selected the most representative English grammar mistakes from many English grammar problems, as the evaluation main content, to judge the system error detection ability effects. Other grammatical contents are relatively low in the proportion of corpus, and the proportion of different corpus is unstable, so it is unified into other contents. The content of the evaluation in this article is shown in table 3.

Table 3. Grammatical error evaluation content

Categories	Content
Vt	Verb tense error
V0	Verb deletion error
Vform	Verb form error
SVA	Subjective-verb error
AitQrDet	Article error
Prep	Preposition error
Nn	Nouns plural error
Others	

The test mainly uses NUCLE-release2.2 corpus and CLEC corpus.

4.2 Evaluation index

The evaluation of error detection effect uses the evaluation of F-measure, and F-measure is the full name. It is a commonly used evaluation standard in the natural language processing and grammar check field, and the calculation formula is as follows:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{4}$$

In the above formula, P refers to the precision, and R suggests the recall. Their value is between 0 and 1, usually represented by percentage. The calculation formula of accuracy is shown as follows:

$$P = \frac{A}{A+C} \tag{5}$$

The accuracy rate indicates the percentage of the number of sentences marked as correct in the sentences marked as wrong. A represents the number of sentences that are labeled as correct, and C suggests the number of sentences that have no errors but are marked as wrong. The calculation formula of recall rate R is as follows:

$$R = \frac{A}{A+B} \tag{6}$$

The recall rate indicates the percentage of the number of incorrect sentences that are labeled as correct in all the wrong sentences in the text. B represents the number of sentences that are wrong but not annotated in the text. Recall and accuracy are two important concepts and indicators in the grammar inspection system.

At that time, it is the most common F1-Measure. This article uses F1-Measure:

$$F_1 = \frac{P * R}{P + R} \tag{7}$$

Usually we expect that: for the grammatical errors in English sentences, the more detected the better, that is, the better the greater "recall rate". and it is better if the accuracy is better, that is, it is better it it is more relevant in the detected grammatical errors. The two values can be objectively evaluated by the F1 values, and the higher the F1 value is, the better the error correction effect of the system.

4.3 Verification of the error detection effect

For the grammar check system designed in this paper, the current rules have about more than 900 examples built inside, from English grammar related books and the summary of Internet on English grammar; N-gram syntax checking experimental part uses the SRILMN-gram model based on English Wiki, and the actual system uses Google N-gram for the calculation. The article part and the preposition part both adopt Wiki+NUCLE as the training corpus, and the experiment uses NUCLE-rclease2.2 as the test corpus.

The test corpus used by the system comes from CLEC Chinese Learner English Corpus. CLEC corpus is China's first Chinese Learner English Corpus, and it is also the world's first officially open corpus, which consists of a large number of middle school and college English corpus. The editor makes tagging grammar and speech errors of library corpus. In this test corpus, 100 English compositions are selected randomly from CLEC as the test sets, including 1128 sentences, a total number of 1083 errors. As shown in table 4, on this test set, the grammar check module is tested.

Table 4. F-measure of the syntax check module on the CLEC test set

	Precision	Recall	F
Verb error detection	45.51	28.77	35.50
Subjective-verb concord error detection	40.00	21.33	31.71
Noun plural error detection	85.71	54.12	62.74
Article error detection	67.39	42.84	52.38
Preposition error detection	72.13	37.87	49.66
Overall error detection	82.82	25.57	39.07
Your words net.	89.72	48.70	-
GCSCL	67%	-	-

Compared with English grammar checking system (GCSCL system), facing Chinese learners with the use of CLEC test set by Wang Quanbin, it is also randomly selecting 100 articles from CLEC corpus as the test corpus, to check the error of the test set, whose accuracy rate is 67%. In comparison, the system has a certain degree of improvement in terms of its accuracy on this basis.

5 Conclusion

A grammar checking system was designed and developed. The grammar checking module adopted the grammar checking design based on multi-rules and used multi-layer grammar rules to provide error correction function for common English grammar. To improve the grammar checking accuracy of grammar checking module, the grammatical model based on statistics was introduced by the system. The specific conclusion is as follows:

First, the main work of this article is about the grammar check design, which realizes the grammar error checking and error correction, as well as the feedback to the wrong opinions.

Second, the focus of this paper is to improve the accuracy and recall of grammar checking. To improve the accuracy of the grammar checking system, the N-gram grammar module is added to the system.

Third, the rule syntax module is supplemented with GoogleN-gram. As a filter, the grammatical error of the rule syntax module is discriminated, and the unreasonable modification is filtered out. It solves the problem of rule grammatical misinformation and eliminates the two meanings caused by the conflict of rules and grammatical rules to some extent.

Fourth, the syntax check module experiment for English composition aided marking system designed and implemented in the paper compares the domestic several commonly used grammar checkers. F-value is close to sentence grammar check system used often in domestic. In the same test case, the accuracy rate is higher than that of Chinese learners' English grammar checker GCSCL system, and the F value is higher in articles and prepositions checking, which achieves the design accuracy requirement. It can provide an effective syntax checking function for the main system.

6 References

[1] Doerfel, S., Singer, P., Singer, P., Niebler, T., Strohmaier, M., & Strohmaier, M. What users actually do in a social tagging system: a study of user behavior in bibsonomy. Acm Transactions on the Web, 2016, vol. 10(2), pp. 14. https://doi.org/10.1145/2896821

[2] Dale, R. Checking in on grammar checking. Natural Language Engineering, 2016, vol. 22(3), pp. 491-495. https://doi.org/10.1017/S1351324916000061

[3] Shi, H., Chen, S. Y., & Lin, K. Raman spectroscopy for early real-time endoscopic optical diagnosis based on biochemical changes during the carcinogenesis of barrett's esophagus. World Journal of Gastrointestinal Endoscopy, 2016, vol. 8(5), pp. 273-275. https://doi.org/10.4253/wjge.v8.i5.273

[4] Cheng, W., Ni, J., Zhang, K., Chen, H., Jiang, G., & Shi, Y., et al. Ranking causal anomalies for system fault diagnosis via temporal and dynamical analysis on vanishing correlations. Acm Transactions on Knowledge Discovery from Data, 2017, vol. 11(4), pp. 40. https://doi.org/10.1145/3046946

[5] Han, X., Li, B., Wong, K. F., & Shi, Z. Exploiting structural similarity of log files in fault diagnosis for web service composition. Caai Transactions on Intelligence Technology, 2016, vol. 1(1), pp. 61-71. https://doi.org/10.1016/j.trit.2016.03.006

[6] Gabrilovich, E., & Markovitch, S. Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research, 2014, vol. 34(4), pp. 443-498.

[7] Murata, M., Ito, S., Tokuhisa, M., & Ma, Q. Order estimation of japanese paragraphs by supervised machine learning and various textual features. Journal of Artificial Intelligence & Soft Computing Research, 2015, vol. 5(4), pp. 247-255. https://doi.org/10.1515/jaiscr-2015-0033

[8] Sengupta, P. P., Huang, Y. M., Bansal, M., Ashrafi, A., Fisher, M., & Shameer, K., et al. Cognitive machine-learning algorithm for cardiac imaging: a pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy. Circulation Cardiovascular Imaging, 2016, vol. 9(6), pp. e004330. https://doi.org/10.1161/CIRCIMAGING.115.004330

[9] Kelly, J., Barends, R., Fowler, A. G., Megrant, A., Jeffrey, E., & White, T. C., et al. State preservation by repetitive error detection in a superconducting quantum circuit. Nature, 2015, vol. 519(7541), pp. 66. https://doi.org/10.1038/nature14270

[10] Ferraro, J. P., Rd, D. H., Duvall, S. L., Chapman, W. W., Harkema, H., & Haug, P. J.. Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. Journal of American Medical Informatics Association Jamia, 2013, vol. 20(5), pp. 931-939. https://doi.org/10.1136/amiajnl-2012-001453

7 Author

Xurong Xu is with the Yancheng Vocational Institute of Industry Technology, Jiangsu, China.

Article submitted 21 March 2018. Resubmitted 04 April 2018. Final acceptance 07 May 2018. Final version published a submitted by the author.