Research on Grammar Checking System Using Computer Big Data and Convolutional Neural Network Constructing Classification Model

Fang Sun¹ *, Jing Zhang²

¹School of Aviation Fundamentals Air Force Aviation University Changchun, P. R. China
²Foreign Training Group Air Force Aviation University Changchun, P. R. China

*Corresponding author: 14033114@qq.com

Abstract. This article proposes an automatic grammatical correction method for typos and word order errors that may occur in Chinese writing by beginners. The thesis first constructs heuristic rules and expands the corpus by analyzing the characteristics of different grammatical error types in the data set. Secondly, when the paper uses classification methods to detect grammatical errors, it extracts sentence-level binary and ternary part-of-speech combinations, n-gram models based on part-of-speech statistics and other three types of features to construct single classification and ensemble classification models, and then use convolutional neural the network constructs classification models from different angles. Finally, when the paper adopts the method based on sequence labeling for grammatical error detection, it mainly uses dependency syntax tree features, and realizes grammatical error detection by constructing a conditional random field model. This method can automatically detect grammatical errors while also identifying the sentence the location of the error. On this basis, the paper implements a simple Chinese grammatical error automatic detection system, which can provide help for the optimization of questions and answers in the question-and-answer system.

1. Introduction

The development of international Chinese education is very rapid. Many countries and regions have included Chinese language courses as required or optional courses. Many countries and regions have established Confucius Institutes and Confucius Classrooms. Against the background of the accelerated development of Chinese teaching, Chinese teaching research experts, Chinese teachers and Chinese learners are all looking for better Chinese teaching and learning models. The study of Chinese as a second language teaching has been further in-depth, and the research objects have become more diversified [1]. Among them, Chinese grammar checking is a research topic that has attracted much attention. In this highly information age, people's demand for information interaction between humans and computers is increasing, and the computer's demand for natural language understanding is becoming more and more urgent. The computer's ability to understand natural language will directly affect the level of artificial intelligence, and is closely related to people's daily experience of information life. The description of natural language will inevitably contain various grammatical errors, which makes the
computer's understanding of the original meaning of sentences expressed in natural language deviate, and then affects the computer to make correct judgments and responses. Therefore, the study of automatic checking methods for grammatical errors of natural language sentences, as the key to improving the human-computer interaction experience, has naturally become an important direction in the field of natural language processing by computers. Based on this research background, this article proposes an automated grammatical correction method for typos and word order errors that may occur in Chinese writing by beginners. On this basis, the paper implements a simple Chinese grammatical error automatic detection system, which can provide help for the optimization of questions and answers in the question-and-answer system.

2. Model structure design

2.1. Self-attention

The core of Transformer lies in the use of self-attention network. This network has played an important role in machine translation, text generation, dialogue system and other fields due to its efficient parallelism and long-distance information dependence, and has received extensive attention from scholars [2]. It is composed of N identical neural modules, and the input and output of the modules are connected together. Each module contains two parts, a multi-head attention layer and a feedforward layer. Among them, the multi-head attention layer Multi-Head is composed of multiple attention layers spliced together, and each attention layer Attention uses a range dot product, as shown in equations (1) and (2).

\[
\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
\]

\[
\text{Multi-Head} = [\text{Att}_1, \text{Att}_2, \ldots, \text{Att}_h]
\]

Among them, Q, K, and V represent the query matrix, key-value matrix and real-valued matrix of the attention layer respectively, which are obtained from the input vector through three different linear layers, that is, the self-attention of the input vector is calculated. \(d_k\) is the size of the third dimension of the embedding layer of the model. This factor is to adjust the size of the inner product of the transpose of Q and K, so as to prevent the excessively large inner product from being unevenly distributed after SoftMax. Multi–Head Is the splicing of multiple attentions?

It is also composed of N identical neural modules. In addition to the two parts identical to those in the encoder, it also contains an additional codec multi-head attention layer. The calculation method is shown in equation (1). The difference from the multi-head attention layer lies in the values of Q, K and V. Among them, Q and K are both the output vector of the encoder, and V is the output of the multi-head attention layer of the decoder, that is, the calculation is the attention between the encoder and the decoder vector. In the encoder and decoder, each independent layer has a normalization layer and a residual structure. The normalization layer can map the past vector values to between 0 and 1, speeding up the convergence speed of the model; the function of the residual structure is to make the gradient not 0 when the model depth is too deep.

2.2. Motivation

The Transformer model uses the self-attention mechanism to do attention calculations on all texts to get the attention content. However, from the actual situation, the particularity of the error correction task is that the difference between the wrong text and the correct text is very small, and the model's focus is too broad, so the model's focus is on characters within a certain range of the error text., To extract the characteristics of error characters, thereby enhancing the model's adaptability to error correction tasks.
Based on the above ideas, this paper adds a quadratic function bias term similar to the Gaussian distribution on the basis of the self-attention score to reduce the attention to the non-error part. As shown in Figure 1, a Gaussian distribution centered on the wrong character "Bai" is added to the original distribution, so that the model can strengthen the attention of the character and surrounding characters [3]. The distribution is trainable. There are two distribution modes, Gaussian distribution and enhanced Gaussian distribution. Both distributions focus on the error part of the information. The difference is that the enhanced Gaussian distribution reduces the prediction of error points to increase the attention to surrounding information, thereby improving the recall rate of the model.

![Figure 1. Two different Gaussian distributions.](image)

### 2.3. Local Attention Model (LF)

After calculating the attention score, the bias G is added to the score to mask part of the self-attention score. The formula is as follows:

\[
ATT(Q,K) = \text{softmax} \big( \text{scores} + G \big)
\]  

(3)

Among them, scores are the result of the dot product calculation, and G is the added bias term. The calculation method of G is as follows:

\[
G_{i,j} = -\frac{(j - P_i)^2}{2\sigma_i^2}
\]  

(4)

\(\sigma_i\) Represents the standard deviation, \(\sigma_i = D_i / 2\) where \(D_i\) is the size of a window, \(P_i\) is the position of the prediction error, \(G_{i,j} \in [0, -\infty]\), \(P_i\) and \(D_i\) are calculated as follows:

\[
\frac{P_i}{D_i} = \text{Isigmoid} \left( \begin{bmatrix} P_i \\ z_i \end{bmatrix} \right)
\]  

(5)
The parameter $I$ projects the scalars $p_i$ and $z_i$ between 0 and the sentence length, and $P$ depends on $p_i$ and $z_i$ to be calculated separately. It is worth noting that we propose that the information encoded by the Transformer in different layers is different. At the lower level, the model is mainly used to encode the grammatical structure of the text, while at the higher level, the model is encoding semantic information.

### 2.4. Grammatical strengthening

The analysis found that the wrong text will lead to the destruction of the grammatical structure to a large extent, as shown in Figure 2. In the example "Children like to pat the ball", if the "pat" is incorrectly entered as "Bai" when inputting, it can be clearly seen that the grammatical structure of the wrong sentence has changed a lot. In order to extract the corrupted grammatical structure information, in the experiment we use the middle-level structure of LSTM, and call it ON_LSTM. This structure is different from the traditional LSTM structure. It learns the grammatical structure of sentences by controlling the update frequency of high-level information and low-level information in the encoding process. In this paper, the structure information generated by ON_LSTM is added to the original text information, and finally the obtained information is sent to the coding layer of Transformer for joint coding [4]. As shown in Figure 3, this structure is similar to the residual module of the encoding part of Transformer, which is intended to enhance the circulation of data and reduce the problem of the disappearance of the gradient in the deep model.

![Figure 2. Examples of grammatical structure changes caused by wrong words.](image)

![Figure 3. The overall structure of the model.](image)
3. Grammar check system design

3.1. System framework
The Chinese vocabulary and grammar checking system can be used by three categories: students, teachers and system administrators. By using the grammar check system, students can take self-tests and obtain corresponding learning feedback, thereby improving their learning [5]. Through the system, teachers can understand the vocabulary mastery of the taught students, create their own question bank, organize students to take exams, and improve their work efficiency. The administrator has the highest authority to manage and maintain students, teachers and question banks. The functional framework of the Chinese vocabulary grammar checking system is shown in Figure 4:

![System function framework](image)

**Figure 4.** System function framework.

3.2. Grammar check process
The process of computer-assisted vocabulary grammar checking is as follows:

1) Construction of vocabulary and grammar check resources (question bank). Use multimedia technologies such as text, sound, graphics, images, videos, and animations to construct vocabulary grammar check resources, present them to the grammar checkers in a more humane, closer to the learners’ physical and psychological needs, and build a good grammar checking environment close to the real language communication situation.

2) Identification of response data. Through natural language processing technology, including Chinese character, speech recognition and other data processing technology, the data input by the learner is recognized and converted into a form that the computer can understand.

3) Automatic grammar checking. Automated grammar checking is also called Computer Automated Scoring (CAS), which refers to the process of computer simulating manual scoring to score questions (including objective questions and open questions), skills, operations, and performance activities.

4) Construction of feedback content. Feedback content can guide students’ learning, help teachers improve teaching and adjust teaching goals. The grammar check result is processed through multimedia technology, and the grammar check result is presented to the user in a variety of targeted forms. The design and construction of feedback content is the research focus of computer-assisted vocabulary and grammar checking.
3.3. Contents and methods of Chinese vocabulary grammar check

3.3.1. Chinese vocabulary and grammar check content. The content of the vocabulary grammar check is the learner's vocabulary ability. Although the expressions of vocabulary ability are different from each scholar, the vocabulary ability includes at least two aspects: vocabulary knowledge and the ability to use vocabulary knowledge. Vocabulary knowledge is the core of vocabulary ability. Vocabulary knowledge can be divided into vocabulary breadth and vocabulary depth, namely quantity and quality. Vocabulary includes receptive vocabulary and productive vocabulary [6]. The grammar check content of this system is mainly based on in-depth vocabulary knowledge, that is, students' mastery of a single vocabulary, which evaluates students' vocabulary learning at a micro level, and does not involve grammatical checks on students' Chinese vocabulary.

3.3.2. Chinese vocabulary grammar check method. This system will select the "Chinese Course" edited by Yang Jizhou as the vocabulary source. This textbook is intended for college students and adult Chinese learners. Therefore, the selection of multimedia resources will focus on practicality, simplicity and beauty. In the grammar check of word sounds, two multimedia elements, text and sound, are used. Word sounds are presented in the form of pinyin. However, in order to prevent some learners from being unfamiliar with the Chinese phonetic scheme and unable to answer the questions, the system uses sound to present the test questions at the same time-Click the pinyin to hear the pronunciation and complete the answer smoothly. In the grammatical check of the word form, it is directly presented in the form of Chinese character text. In the grammatical check of the meaning of a word, it is mainly presented in interlanguage (English), and is assisted by multimedia such as images, animations and videos. In the grammar check of words, taking into account the principle of economy, the test is mainly carried out with text [7]. There are currently three types of questions in the system: unequal item matching questions, set library-style fill-in-the-blank questions and multiple-choice questions. The purpose of this design is to separate test question types. The knowledge or ability points of the assessment are very clear, which can provide teachers and students with clear feedback and diagnostic information; the use of "minimum context" design questions is to reduce other knowledge the interference affects the validity of the question.

3.4. System construction mode
The B/S structure is divided into three parts: browser (Browser, server (Webserver) and database, also known as presentation layer, function layer and data layer. The browser is a user-oriented interface and is a window for information interaction. The layer is composed of Web pages. The server layer is the core part of the realization of system functions. Data processing and statistical analysis functions are concentrated in this part. The database mainly stores user information, topic information, data generated in the test, etc., the construction mode of the system as shown in Figure 5:
4. System test analysis

4.1. Experimental parameter settings
The Transformer model used in the experiment has 6 layers of encoding and decoding, the number of "heads" is 8, the word vector dimension is 512, the maximum sequence length is 36, the number of fully connected layer nodes is 4096, and the initial learning rate is 0.001. When there is no improvement in the performance of the 2nd generation, the linear attenuation is performed at a rate of 0.95. If there is no improvement in the 6 consecutive generations, the training is terminated early, and the best model on the verification set is retained for testing [8]. Under different experimental configurations, as the training progresses, the verification set the change curve of the loss function is shown in Figure 6. The data set used in the experiment is expanded on the basis of the public data set of the NLPCC competition. The statistical attributes of the original data are shown in Table 1.
### Table 1. Experimental data statistics.

| Data category | Quantity | The maximum length | Minimum length | Average length | Different word count |
|---------------|----------|---------------------|----------------|----------------|---------------------|
| **Training set** |          |                     |                |                |                    |
| Original S    | 65133    | 212                 | 1              | 17.8           | 7782               |
| Original T    | 6         | 213                 | 1              | 18.2           | 7721               |
| **Test set**  |          |                     |                |                |                    |
| Generate 1S   | 16087    | 36                  | 2              | 16.26          | 7782               |
| Generate 1T   | 54        | 36                  | 2              | 16.67          | 7721               |
| Generate 2S   | 15526    | 36                  | 2              | 16.26          | 8506               |
| Generate 2T   | 87        | 36                  | 2              | 16.67          | 7721               |
| S             | 2000      | 247                 | 7              | 29.66          | 2214               |

#### 4.2. Performance test

The test data set used in the experiment is randomly selected from the above learning corpus, a total of 10 test subsets, which contain 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50 in sequence. For the same test corpus, the language technology platform (LTP) provided by the laboratory is also used for testing, and the experimental results are compared. The test determines the success rate according to the correctness of the sentence component analysis. The results are shown in Figure 7 and Table 2.

![Figure 7. Performance test chart.](image-url)
Table 2. Comparison table of experimental results.

| Model     | Data          | P   | R   | F0.5 |
|-----------|---------------|-----|-----|------|
| AliGM     | Original      | 41  | 13.75 | 29.36 |
| YouDao    | Original      | 35.24 | 18.64 | 29.91 |
| Transformer | Original     | 32.27 | 18.58 | 27.65 |
|           | Original + Generate 1 | 34.76 | 21.55 | 30.96 |
|           | Original + Generate 2 | 34.92 | 21.88 | 31.2 |
|           | All           | 35.03 | 23.87 | 32.03 |
| C-Transformer | Original   | 36.49 | 19.15 | 30.89 |
|           | Original + Generate 1 | 37.09 | 21.57 | 32.42 |
|           | Original + Generate 2 | 37.34 | 22.74 | 33.09 |
|           | All           | 38.22 | 23.72 | 34.05 |

5. Conclusion

The thesis constructs an effective set of rules for Chinese grammatical error detection by analyzing the characteristics of different error types in the corpus. First, we preprocess the sentences in the corpus; then, compare each type of grammatically wrong sentence with its corresponding correct sentence to obtain the characteristics between adjacent words and words, part of speech and part of speech in the sentence; finally, according to these the relationship between features and sentence error types constructs a rule set to detect and recognize Chinese grammatical errors. Aiming at the current computer-assisted Chinese grammar checking system's insufficient feedback and lack of statistical analysis functions, this research designed and developed a computer-assisted Chinese vocabulary grammar checking system for Chinese teaching. It is used in Chinese vocabulary grammar checking and computer-aided Chinese grammar checking.

References

[1] Manchanda, B. Athavale, V. A. & kumar Sharma, S. Various techniques used for grammar checking. International Journal of Computer Applications & Information Technology, vol. 9, pp.177-189, January 2016.

[2] Qi, W. Zhang, H. & Sorokina, N. Linguistic Landscape for Korean Learning: A Survey of Perception, Attitude, and Practice of Korean Beginners at a Korean University. Journal of Language Teaching and Research, vol. 11, pp. 956-961, June 2020.

[3] Zabolotskikh, A. Zabolotskikh, A. Dugina, T. & Taveridze, D. Creating individual learning paths in the Moodle plugin for undergraduate students to study English grammar. Education and Information Technologies, vol. 26, pp. 617-637, January 2021.

[4] Gill, M. S. Lehal, G. S. & Joshi, S. S. Part of speech tagging for grammar checking of punjabi. The Linguistic Journal, vol. 4, pp. 6-21, January 2009.

[5] Frajzyngier, Z. Liu, M. & Ye, Y. Reference system in modern Mandarin Chinese. Australian Journal of Linguistics, vol. 40, pp. 45-73, January 2020.

[6] Bolt, P. & Yazdani, M. The evolution of a grammar-checking program: LINGER to ISCA. Computer Assisted Language Learning, vol. 11, pp. 55-112, January 1998.

[7] Zhang, W. & Vesnina, L. E. Online teaching Russian practical grammar at a Chinese university during the epidemic. Russian Language Studies, vol. 18, pp. 383-408, April 2020.

[8] Sharma, S. K. Rule Based Grammar Checking Systems (A Survey). International Journal of Computer Applications & Information Technology, vol. 10, pp. 217-220, January 2016.