The Research of Chinese Proofreading Based on Knowledge Graph

Ligang Dong[a], Mengying Wu, Hong Shao and Xian Jiang

School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou, China

[a]donglg@zjgsu.edu.cn

Keywords: Chinese automatic text proofreading, Knowledge graph, Word proofreading, Semantic proofreading

Abstract: Automatic proofreading refers to the use of computers to identify and correct writing or grammatical errors in text. Today's automatic proofreading studies use large-scale lexicon to proofread words, making it difficult to syntactically proofread and not support large-scale free text processing. Therefore, this paper uses the knowledge map to achieve syntactic and semantic proofreading of the text, in which the proofreading semantic error types are wrong word, missing components and defining contradictions. Compared with the widely used Chinese automatic proofreading system, the semantic proofreading method has a high recall rate.

1. Introduction

The wide application of computer spawned Chinese text automatic proofreading Tools, to substitute the time-consuming traditional artificial proofreading, is one of the most common Office bring Chinese automatic proofreading tool Office Proofing Tools, and other widely used tool of proofreading assistant, small red pen, hermas proofreading system, etc. However, the existing proofreading tools can only realize word proofreading based on large-scale thesaurus, which is difficult to achieve syntactic and semantic proofreading. Moreover, these proofreading tools are paid software, and users need to pay relatively high fees to be used [1]. If the lexicon is not updated in time, the proofreading effect will be affected. Moreover, due to the excessive reliance on lexicon, we can only proofread the word errors in the text, and fail to recognize the syntactic and semantic errors in the sentence, such as the incomplete composition of the sentence and the definition contradiction between the sentences.

Chinese text automatic proofreading started relatively late in China, and the existing text automatic proofreading technologies mainly include context-based local language features, rule-based and statistics-based proofreading methods [2].

(1) Proofreading method based on local language features of context

The features of local language refer to the features of part of speech, character, and so on. Microsoft research China first used the Winnow method to learn the word-related local language features and long-distance language features in the text, and then selected the words in the target word confusion set according to the context features [3]. The difficulties of this multi-feature-based proofreading method lie in feature extraction and obfuscation set construction [4]. Harbin Institute of Technology based on the candidate words of all the words in the sentence to be proofread to obtain the candidate matrix of the corresponding sentence. According to the statistical and structural features of the sentence, the best word sequence is selected from the candidate matrix and compared with the original sentence to find the wrong words. The difficulty of this method lies in the construction of word candidate matrix [5]. Although the accuracy of the method based on local features is high, the algorithm complexity is high and it is limited in practical application.

(2) Proofreading method based on rule

Beijing normal university USES the correction grammar rule to proofread the text. When the sentence meets the rule, it only needs to mark the corresponding word wrong according to the rule. However, the university has limited error-correcting ability to the method [6]. Harbin Institute of Technology USES the phrase rule to combine a word with a participle to form a phrase, then

gradually binds the correct string and marks the remaining single characters as errors. The limitation of this method is that it cannot detect the substitution error of multiple strings, and the phrase rule constructed by it covers a narrow range [7]. Wu yan et al. [8] used the reverse maximum matching method and the local corpus statistical algorithm to obtain the scattered strings in the text, and then obtained the candidate error strings through word matching and grammar analysis. Finally, they corrected the error strings through the interactive method. The method of this school is simple to implement, which only needs to proofread according to the established rules. However, due to the inability to exhaustive all the rules, the proofreading accuracy is affected by the rules.

(3) Statistical proofreading method

Shen maobang [9] and ma jinshan [10] both proposed to use the n-gram model of Chinese characters and dependency syntactic analysis to obtain the structural information of a sentence, and then realize the proofreading of text misspellings. Duan liangtao et al. [11] proposed a word-based language model and corp-based n-gram error checking strategy to realize automatic proofreading of Chinese texts. Sun et al. [12] proposed the method of "strapping", using the n-gram language model to proofread and correct text. In the limited field, the method has a high error rate, but it is still affected by the size of the training corpus and the type of corpus field.

Based on the research of Chinese text proofreading system, this paper proposes a semantic proofreading method based on knowledge graph. Firstly, entity extraction technology is used to extract the entities in the text statement. Then, based on the matching results of the entities and relational rules, syntactic and semantic error types are found in the knowledge graph. By comparing with the existing automatic Chinese proofreading system, we find that the method has a high recall rate for all kinds of semantic errors in the limited field.

2. Design of Chinese text proofreading system

Because semantic errors are hard to find, the existing proofreading technology of Chinese text is basically based on large-scale thesaurus, such as black horse proofreading system and proofreading assistant. If the thesaurus is not updated in time, the proofreading result will be affected. In addition, it is difficult to accomplish syntactic and semantic proofreading by comparing the words in the proofreading text with the words in the lexicon. Therefore, based on the knowledge map, this section USES the entity extraction function to find syntactic and semantic errors in the sentence, whose proofreading error types include wrong words, missing components and definition contradictions. Therefore, the Chinese automatic proofreading system designed in this section mainly includes proofreading modules for wrong words, defining contradiction proofreading modules, proofreading modules for missing components and comprehensive proofreading modules. The specific structure is shown in the figure 1.

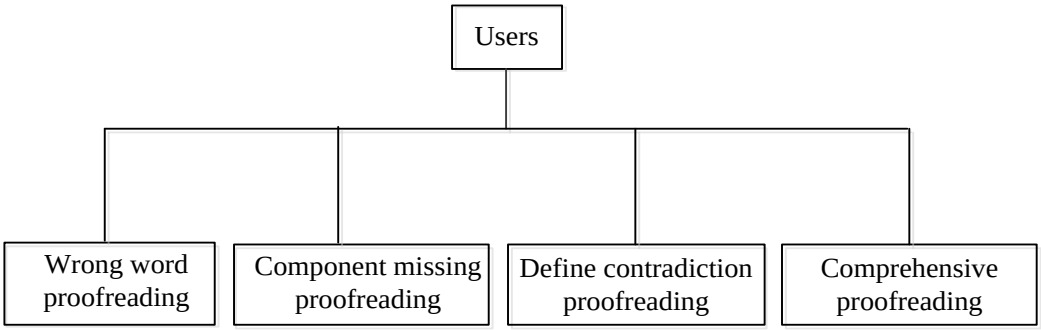

Figure 1. Structure of Chinese text proofreading system

2.1 Wrong word proofreading module

The wrong word is usually the sound of the right word or shape near the word, is usually in the use of different input methods to input the error, artificial not easy to find this kind of error. The wrong word module can only be used to proofread the word as a solid word in the text, and the

wrong word results given by the system can only reflect that the word has the tendency of misspelling, but cannot completely determine that it is wrong.

The process of proofreading in this module is as follows: firstly, entities in the text to be proofread are extracted by using entity extraction function, and then these entities are matched with triples in the knowledge graph. If an entity is matched and the results are consistent, the entity word is accurate. If it matches the related entity but does not match, it means that the entity word may be an error, then return the related entity as the correct word. If the relevant entity cannot be completely matched, it means that it is impossible to judge whether the entity word is an error, as shown in the figure 2.

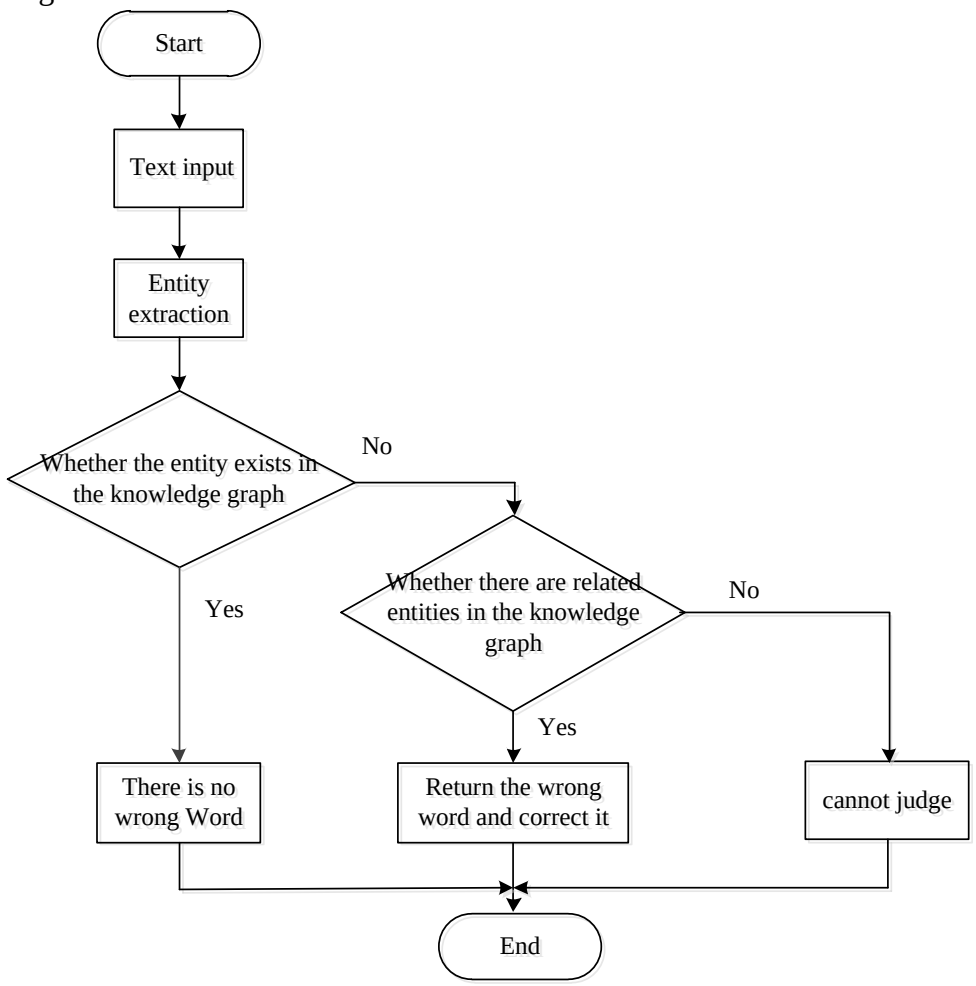

Figure 2. Structure of wrong word Proofreading Module

2.2 Component missing proofreading module

The component missing proofreading module is used to proofread whether there are missing components in the sentence. This section mainly proofreads the three components of subject-verb-object. Also, the entity extraction function is used first to extract the entities in the statement. If the extracted result is the entity pair, then the relationship of the entity pair is obtained by using the relationship rule matching again. If the entity pair and the relationship label can match successfully in the knowledge graph, then the statement is complete. If the relation of entity pair cannot be obtained by using relationship rule matching, the predicate component may be missing. When the extracted result only has a single entity, and the corresponding relationship label can be found by using relationship rule matching, if the matching is successful in the knowledge graph, it is necessary to determine whether the entity is entity 1 or entity 2 of triples. If the result is entity 1, the statement may be missing the object and be completed with entity 2, whereas it may be missing the subject and completed. Among them, relationship rule matching refers to scanning the statement according to the keywords of each relationship label. If there are keywords of corresponding

relationship labels in the statement, the matching will be successful. Some relationship rules are shown in table 1.

Table 1. Partial relationship rules

Key words of questions	Relationship label
Inclusion, include, coverage, division, divide into...	Hyponymy relationship
Equal to, also known as, that is, approximately equal, also called...	Synonymy relationship
concept, refers to, what is it, meaning, Called, explanation...	Interpretive relationship
Belongs to, characteristic, attribute, advantage...	Part-whole relationship
It is for, because, so...	Causality

If the knowledge graph cannot be used to determine whether the sentence has missing components, the dependency parsing can be used to analyze the sentence components, but this method cannot be used to correct errors. As shown in the figure 3.

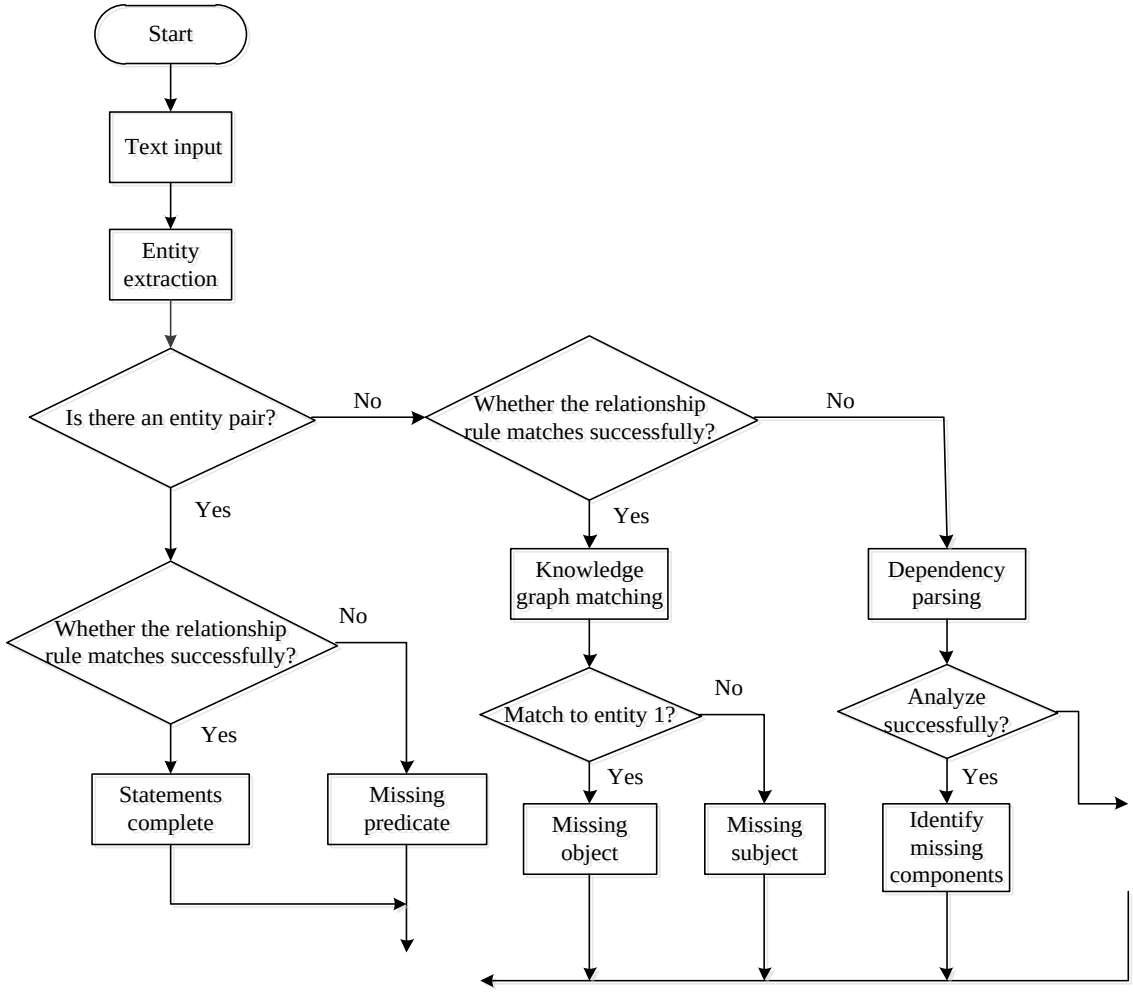

Figure 3. Structure of component missing proofreading module

2.3 Define contradiction proofreading module

The proofreading module that define contradiction is used to check whether there is definition contradiction between statements, that is, whether there is ambiguity in the content described by different statements under the same topic. Similarly, firstly, the entity extraction function is used to extract the entity pairs in each statement, then the relationship rule matching is used to obtain the entity pairs in the statement, and finally, the search result based on knowledge graph is used to analyze whether the contradiction is defined. If the entity pairs of different statements are the same, but the relationship labels are different, the statement may define a contradiction. If the entity 1 and

relationship labels of different statements are the same, but the entity 2 is different, the statement may define a contradiction. If the entity 2 and relationship labels of different statements are the same, but the entity 1 is different, the statement may also define a contradiction. The specific process is shown in the figure 4.

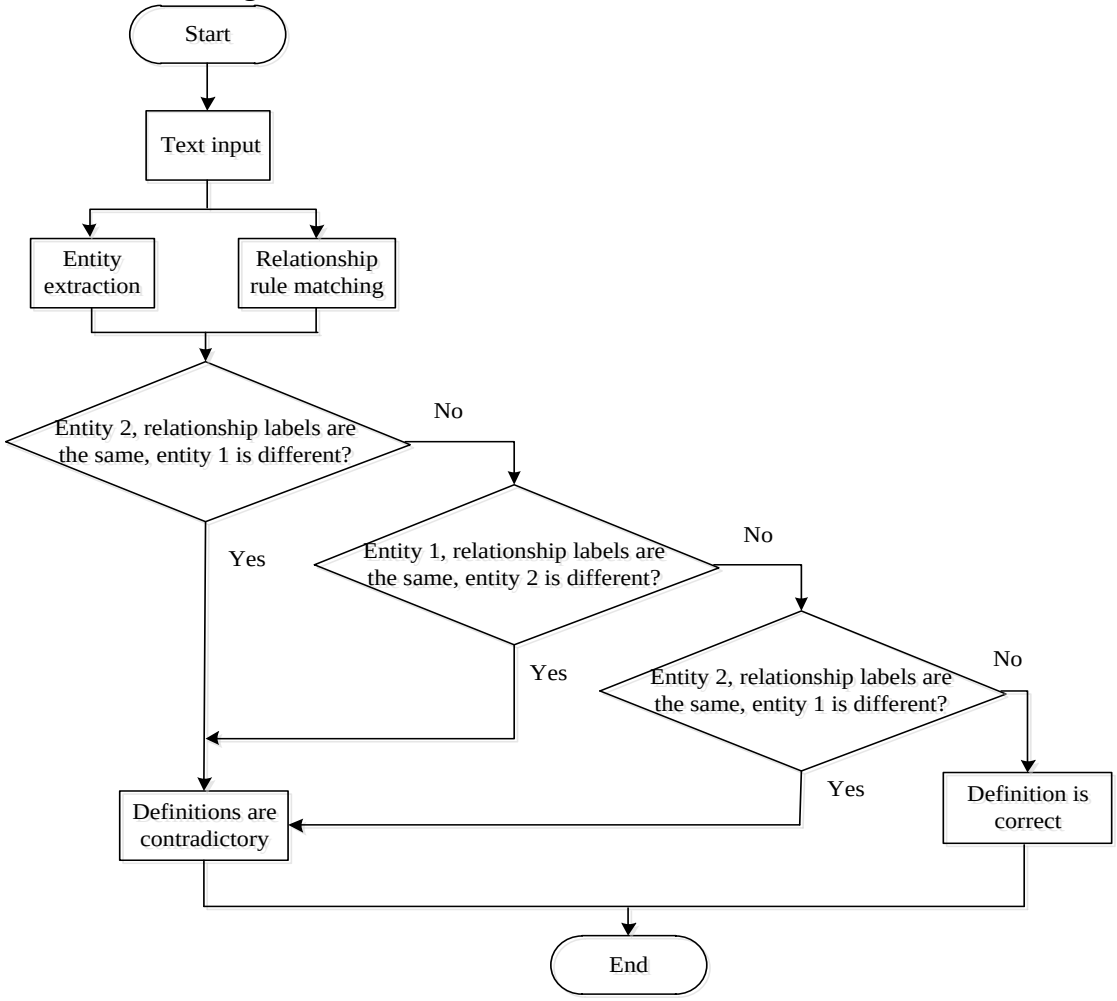

Figure 4. Structure of define contradiction proofreading module

3. Comparison of proofreading performance

Existing Chinese text automatic proofreading tools are mainly black horse proofreading system, proofreading assistant, proofreading expert and small red pen. The performance of the text proofreading method based on entity extraction function and knowledge graph is compared with the four proofreading tools mentioned above. Based on four types of errors, 50 statements related to the Data Structure course that contain corresponding errors and 50 statements that do not contain corresponding errors were selected, and five proofreading methods were used to proofread the statements and calculate the accuracy of proofreading. The accuracy is calculated by dividing the number of errors found by the total number of errors, as shown in the table 2.

Table 2. Performance comparison of proofreading systems

	Wrong word	Component missing	Define contradiction	Speed(S)
Dark horse proofreading system	71.42%	34.1%	28.6%	0.58
Proofreading experts	28.6%	N/R	N/R	0.63
Proofreading assistant	29.7%	N/R	N/R	1.03
Small red pen	80.95%	N/R	N/R	0.67

Proofreading method based on knowledge graph	66.7%	64.8%	72.43%	3.1

4. Conclusion

This paper studies Chinese automatic proofreading. Because the existing widely used proofreading system can only proofread the wrong words based on the large-scale thesaurus, unable to proofread the syntax and semantics. Therefore, this paper uses the entity extraction function and knowledge graph in question and answer research to realize semantic proofreading, and designs a Chinese automatic proofreading system. Among them, the types of semantic errors that can be proofread are wrong word, missing components, defining contradictions and missing contents. By comparing the performance of the proofreading system with that of the widely used proofreading system, it is found that in the limited field, the proofreading function of this paper has a high recall rate in the recognition of wrong words and syntactic and semantic errors.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61871468), the Key Research and Development Program of Zhejiang under Grant No. 2017C03058, the Key Research and Development Program of Zhejiang under Grant No. 2020C01079, Zhejiang Provincial Key Laboratory of New Network Standards and Technologies(No.2013E10012), Education Science Planning Project of Zhejiang No. 2019SCG222, Subsidized Projects for the Promotion of Scientific and Technological Achievements for Postgraduates of Zhejiang Gongshang University No. Js2018472059, School-level Teaching Project of Zhejiang Gongshang University No. PX-1918884 and School-level Project of Zhejiang Gongshang University No. 1120KU219017.

References

[1] T. Zhang, “Design and implementation of Chinese text automatic proofreading system,” Southwest Jiaotong University, 2017.

[2] Y. Zhang and S. Yu, “A review of research on automatic text proofreading,” Computer application research, 2006, 23 (6): 8 - 12.

[3] Golding A. R. and Roth D. “Applying Winnow to Context-Sensitive Spelling Correction,” Computer Science, 1996, 34 (1-3): 182 - 190.

[4] L. Zhang, M. Zhou, C. Huang and H. Pan, “Winnow-based approach in automatic error detection and correction of Chinese text,” Microsoft Research Paper Collection, 2000, 193 - 197.

[5] J. Li, X. Wang and P. Wang, “Research on multi-feature Chinese text proofreading algorithm,” Computer engineering and science, 2001, 23 (3): 93 - 96.

[6] R. Yi and K. He, “Computer Chinese proofreading system,” Computer research and development, 1997 (5): 346 - 350.

[7] T. Liu, H. Shi and Y. Shao, “Principles of Chinese computer-aided proofreading system,” Chinese information, 1997 (2): 21 - 22.

[8] Y. Wu, X. Li and T. Liu, “Research and Implementation of Chinese Automatic Proofreading System,” Journal of Harbin Institute of Technology, 2001, 33 (1): 60 - 64.

[9] M. Shen, “Chinese Automatic Proofreading Technology Based on Statistical Model and Dependency Analysis,” Harbin Institute of Technology, 2003.

[10] J. Ma, “Research on Chinese Automatic Proofreading Based on N-gram and Dependency Analysis,” Changchun University of Technology, 2003.

[11] L. Duan and S. Guo, “Research on Chinese text proofreading technology,” Computer Knowledge and Technology, 2014 (19): 4602 - 4604.

[12] Y. Sun, Y. Zhang and Y. Zhang, “Chinese Text Proofreading Model of Integration of Error Detection and Error Correction,” The Workshop on Chinese Lexical Semantics, Springer International Publishing, 2016: 376 - 386.