International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 4, No. 1, 2022

Analysis of Chinese Text Automatic Proofreading Technology

Yang Zhang

College of Liberal Arts, Jiangxi Normal University, Nanchang Jiangxi 330022, China

Abstract: The development of computer technology promotes the research on automatic text proofreading technology. The research of Chinese text automatic proofreading started late but developed rapidly. This paper analyzes the difficulties of Chinese text proofreading and reviews the research methods of Chinese text automatic proofreading technology, including statistical-based approach, rule-based approach, deep learning-based approach and hybrid approach. Although some new progress has been made in Chinese text automatic proofreading in recent years, there are many problems remain to be settled due to the complexity of Chinese and the scale of the universal data set of Chinese proofreading corpus.

Keywords: Text proofreading, Chinese information processing, Deep learning, Real-word error.

1. Introduction

As the carrier of language, writing is of great significance to the inheritance and development of human civilization. The appearance of writing marks the entry of human beings into a civilized society. Text proofreading is to check whether there are errors in a piece of text and correct them. Text proofreading, which is an important task of text editing, is applicable to the media, government, academic papers and other fields, and has a wide range of application values. The traditional way of text proofreading is manual proofreading, which has high cost and low efficiency. The development of computer technology and the explosive growth of electronic text promote the development of automatic text proofreading technology.

As an important research content of natural language processing, the research on automatic text proofreading originated in the 1960s. In 1960, IBM Thomas J. Watson Research Center implemented the TYPO English spelling checker on IBM /360 and IBM /370 with UNIX. In 1971, researchers at Stanford University implemented the Spell spelling checker [1]. Subsequently, with the emergence of new input technologies such as speech recognition and OCR recognition, the research on automatic text proofreading has also continued to deepen. At present, some word processing software such as Microsoft Word and WordPerfect have embedded the English spelling check function. The research on automatic proofreading of Chinese text started late but developed rapidly. Since the 1990s, some scientific research institutions and companies have begun to conduct theoretical and technical research on Chinese text proofreading.

2. Difficulties in Automatic Proofreading of Chinese Text

The most common errors in English text are spelling errors. According to statistics (Kukich K., 1992), spelling errors of typed texts in English range from about 0.05% to 30%. 0.05% are carefully edited news texts, and 38% are spelling-difficult applications such as phone book queries. Kukich K. classifies English text errors into two categories, one is isolated-word error, also known as context-free error, that is, the misspelled word does not exist in the dictionary, such as spelling "word" into "wird"; the other type is real-word error, also known as context-dependent word error, that is, the wrong spelling happens to be another word that actually exists in English, such as spelling "again" as "gain"[2].

Chinese text is different from English. Chinese characters are input by means of Chinese character coding, so it is impossible to produce words that are not in the coding during the input process, and thus isolated-word error will not be generated. However, huge character set also brings difficulties to the parameter calculation of the model. In addition to the huge character set, Chinese text proofreading also faces other difficulties.

2.1. Word Segmentation

There is no isolated-word error in Chinese text but real-word error, so automatic proofreading must rely on context dependency, which is inseparable from Chinese natural language processing technology. In order to realize the automatic proofreading of Chinese, the text must be segmented. However, there is no space between words in Chinese writing. Word segmentation is the first step in Chinese natural language processing. The difficulties of Chinese word segmentation mainly focus on ambiguity segmentation and recognition of unknown words. Ambiguity segmentation means that some sentences can be segmented and interpreted in two or more ways, and the word segmentation needs to deal with the ambiguity phenomenon in combination with the context. The recognition of unregistered words mainly includes new words and proper nouns. Although these words account for a small proportion, if not processed, many word segmentation errors will appear, which will affect the syntax and semantic analysis and lead to inaccurate automatic text proofreading.

2.2. Lack of Morphological Changes

The grammatical means of Chinese are mainly word order and function words. Chinese words lack morphological changes, and there is no simple one-to-one correspondence between parts of speech and syntactic components. In text processing, the scale of the corpus and the learning ability of the model are required to be higher.

2.3. Various Chinese Input Methods

Chinese characters are indirectly input into the computer system through manual coding, and different input methods may lead to different types of text errors. For example, the Five-stroke input method generally leads to the error of similar form words, while the Pinyin input method generally leads to errors in homophones or similar sound characters. These different types of errors should also be taken into account when proofreading the text.

2.4. Syntax and Semantic of Long Distance

Most of the real-word errors in Chinese text conform to local language rules, but will affect the syntax and semantics of the entire sentence, and interfere with the analysis of the entire sentence. Automatic text proofreading needs to consider whether the word matches the long-distance words, and also requires the computer to have certain background knowledge and understanding of the context. Of course, this problem is common in automatic text proofreading in all languages, how to enrich the knowledge base of computer and improve its ability to call related knowledge is an important research content of natural language processing.

The first commercial Chinese text proofreading system is the Woodpecker Chinese Spelling Checker. Although the system has low detection performance and lacks error correction function, its appearance has pioneering value for the Chinese proofreading system. Since then, the research on Chinese text proofreading has continued to develop, and more and more universities and research institutions have begun to carry out basic research and software development of text proofreading.

3. Chinese Text Automatic Proofreading Technology

Text automatic proofreading methods can be summarized into four categories, which are statistical-based approach, rule-based approach, deep learning-based approach and hybrid approach. For automatic proofreading of Chinese texts, scholars have also conducted research using these methods.

3.1. Statistical-Based Approach

The essence of Chinese text automatic proofreading based on statistics is statistical probability information, that is, the probability of text occurrence is inferred by calculating the probability of the occurrence of the word, including the mutual occurrence probability between characters and characters, words and words. The statistics on this probability information is usually implemented through language model. The current research mainly focuses on text error detection, which is achieved by comparing with the set threshold, and there are relatively few researches on error correction.

The Woodpecker System (1992) is an early statistical-based typo detection system for Chinese text [3]. This method believes that most text errors will lead to single-character strings after word segmentation. After word segmentation, the system first locates the single-character strings, and then scores the single-word strings according to its word frequency and the strength of the connection between the two Chinese characters forward and backward. After the score is compared with the set threshold, it determines whether it is a wrong word. Zhaohuang Z. (1994) collated the words with similar pronunciation, shape, meaning or input code into an approximate character set, and replaced each Chinese character in the sentence to be proofread with the Chinese character in the approximate character set to form multiple candidate hash strings, and then used bi-gram model scores each hash string, and the parts of the sentence to be proofread that do not correspond to the hash string with the highest score are judged as errors and be corrected [4]. This method can only check the errors of typos, and it is difficult to find text errors such as missing words, addition errors and translocation errors. Yangsen Z. et al. (2001) proposed an automatic text proofreading method based on the bi-neighborship, which mainly used the connection relationship of words. The method firstly located the suspected errors through the word level, and finally determined the errors through the connection relationship between parts of speech and semantic classes [5]. Rong L. (2009) proposed a spelling check system for OCR output of Chinese text. An error-pattern database was established, in which every error pattern can be considered as a rule for correcting errors. The matching algorithm is used to extract the matched corpora for comparison. If the two sentences are different, the wrong words and corresponding correction words are extracted, and the error pattern set is finally obtained. The error pattern is used as a correction rule to find and correct errors in the text to be corrected [6].

Due to the large scale of combinations of words and words, it is necessary to obtain large-scale corpus using statistical methods, and at the same time, it faces the problem of sparse data. Therefore, the effect of automatic text proofreading based solely on statistical methods is not ideal.

3.2. Rule-Based Approach

Rule-based automatic proofreading of Chinese text is to realize automatic proofreading of text by using the knowledge of syntax and semantic rules of linguistics. By analyzing the phenomenon of language and text errors, many scholars have found some rules that can be used for automatic text proofreading. For example, researchers found that after word segmentation in Chinese sentences, in general, characters and characters, words and words are adjacent or alternate, and it is rare that three or more single-character words appear consecutively. If this happens, it is likely that there is an error in this text. The utilization of Chinese text rules can be divided into word level and syntax level.

Some Chinese text automatic proofreading methods mainly use word-level rules. For example, Chaohuang C. (1995) proposed a text proofreading method based on a construction of word confusion sets to form candidate hash strings [7]. This method provided a certain reference value for automatic proofreading of Chinese text, and better solved the problems of low accuracy and the lack of error correction ability of the proofreading system at that time. Jianhua L. et al. (2001) designed and implemented a set of Chinese spelling proofreading system based on the language rules between Chinese words and words [8].

Some Chinese text automatic proofreading methods mainly use the rules of syntax and semantic level. For example, Ting L. et al. (1997) took Chinese clauses as units, after automatic word segmentation, used phrase rules for single-word hash strings to synthesize phrases, scanned three times, and gradually bound the correct hash strings. Single characters that cannot be bundled are flagged as errors [9]. The Chinese manuscript proofreading system developed by Rongxiang Y. et al. (1997) used correction grammar rules to mark the corresponding words if the sentences meet the rules [10].

Xiaojin G. et al. (2003) proposed a method for checking Chinese text syntactic errors combining pattern matching and sentence component analysis by mining the characteristics of Chinese syntactic errors and compiling error pattern rules [11]. Yangsen Z.et al. (2017) constructed a semantic error detection model based on semantic knowledge base and D-S theory [12].

The rule-based text proofreading method improves the accuracy of proofreading to a certain extent, but the limited rules can not cover all language phenomena, and scholars cannot summarize all types of errors. At the same time, language is a dynamic system, and there may be conflict between rules. These factors restrict the proofreading results, resulting in a low recall rate of the method and an unsatisfactory error detection effect.

3.3. Deep Learning-Based Approach

Both statistics-based and rule-based methods belong to traditional proofreading methods. These methods cannot effectively extract text information and lack long-distance dependencies, which affect the effect of text proofreading. In recent years, with the rapid development of deep learning technology, some scholars have applied deep learning technology to Chinese text automatic proofreading, and achieved good results.

Deep learning is a learning method that starts from the original data and transforms each layer of representation or feature into a higher-level and more abstract representation, thereby discovering the intricate structure in high-dimensional data [13] .The advantages of deep learning are long-distance dependence, stronger constraint, which can greatly reduce the problems caused by data sparseness, and stronger generalization ability. These advantages can provide a good modeling method for Chinese text proofreading.

Baiqing R. (2017) trained a multi-layer model for automatic text proofreading based on the method of deep learning using the news corpus of Xinhua news agency. The model continuously extracts the features of the combined text through multi-layer nonlinear operation combination, and finally outputs the abstract representation of the text and the high-level semantic information of the data [14]. Xin X. (2017) used the statistical machine translation and neural network machine translation methods to select the best orthographic method in candidate sentences by using RNN in the N-gram model and proofread the wrong words. The experimental effect was good [15]. Zhou J. et al. (2018) constructed a neural translation model based on double-layer LSTM and attention, and achieved higher text proofreading accuracy in the 2018 NLPCC Chinese Grammar Error Correction Competition [16]. The text proofreading model designed by Yongcai T. (2020) based on ensemble algorithm and Long Short-Term Memory (LSTM) expands the extraction range of semantic information [17]. This model uses a trained neural network to convert the text into sememe sequences, and then predicts and sorts the sememe sequences through LSTM, and finally extracts the words corresponding to the semaphores with high scores as proofreading suggestions.

The convolutional neural network and recurrent neural network, which are traditional deep learning model, have certain limitations. They both use the preceding hash strings to predict the current hash string, and cannot effectively learn complex contextual relationships. In 2018, Google researchers released the BERT (Bidirectional Encoder Representation from Transformers) model, which promoted the development of natural language processing and was also applied to the field of automatic text proofreading. The Chinese spelling checker FASPell [18] designed by iQIYI is simple, fast and adaptable. The SpellGCN Chinese text automatic proofreading model designed by Ant Financial [19] can generate predictions with reasonable semantics and similar pronunciation and font to the original text. In 2020, the Soft-Masked BERT model [20] designed by Fudan University divides text error correction into two parts: a detection network and a correction network, and uses the output of the detection network as the weight to achieve better proofreading results.

It can be seen that the automatic proofreading of Chinese text based on deep learning has a good effect, but there may be some strange semantics or grammatical collocation errors during proofreading. Since deep learning is a "black box model" with poor interpretability, it is difficult to explain the proofreading errors of the algorithm. Therefore, it cannot be modified in a targeted manner. At the same time, Chinese text proofreading currently lacks a common data set and has few labeled data, so there is still a lot of research space in this field.

3.4. Hybrid Approach

Some researchers have adopted the method of combining rules and statistics, and achieved a high error detection rate.

Wu Yan et al. (2001) proposed a proofreading method combining word matching and syntax analysis. This method finds character strings through the algorithm of reverse maximum matching and corpus statistics, and uses word matching and grammatical analysis to process character strings, and corrects candidate error hash string through human-computer interaction [21]. Yangsen Z. et al. (2014) proposed a two-level detection method, which combines statistics and rules to achieve semantic proofreading of texts in the field of political news [22]. However, the rules extracted by this method are limited in scope and cannot be applied to proofreading of Chinese Texts in various fields. Hai Z. et al. (2017) proposed a joint error detection model that combines CRF model, graph model and rule model. The model finds errors within a preset window, and then replaces suspicious words with confusion sets, and then constructs a fuzzy word graph, and finally determines the most reasonable sentence with reference to the shortest path word segmentation algorithm. The rule model is used to proofread specific types of errors [23]. Dezhi G. et al. (2017) proposed an automatic proofreading method based on statistical methods (N-gram, Bayesian model, mutual information, etc.), contextual feature generalization and collocation based on a large-scale corpus, which can automatically proofread global errors and is no longer limited to local errors [24]. Junpei Z. et al. (2018) used a rule-based statistical model, a statistical machine translation-based error correction model, and the LSTM-based translation model for joint proofreading, and finally used a conflict algorithm to merge the models to output error correction results [25]. Yongmei T. et al. (2018) proposed an automatic grammatical error correction method based on LSTM and N-gram to automatically proofread sensitive information in massive texts [26]. Yongcai T. et al. (2018) proposed component analysis of the text to be proofread, and constructed a two-layer Chinese grammar-word collocation knowledge base. On this basis, it combined with the Markov Chain model to construct a Chinese text proofreading system [27]. However, due to the complex

structure of Chinese, it is difficult to extract language rules, and the system is still limited by the refinement of rules.

4. Conclusion

In general, the theory and technology of Chinese text automatic proofreading have developed rapidly in recent years. With the rapid development of deep learning technology, automatic text proofreading has also made some new progress. However, due to the complexity of Chinese and the scale of the universal data set of Chinese proofreading corpus, Chinese text automatic proofreading has many problems remain to be settled. At the same time, the current related research mainly focuses on the texts of Chinese native speakers, and there are few researches on the proofreading of Chinese texts for second language learners, and scholars should pay more attention to this field.

Acknowledgment

This work was supported by grants from National Social Science Foundation of China (No.19CYY007) and Social Science Planning Project of Jiangxi Province (No. 17BJ21).

References

[1] Yangsen Z. and Shiwen Y., Summary of Text Automatic Proofreading Technology, Application Research of Computers, vol.6, 2006, pp.8-12.

[2] Karen K, Techniques for Automatically Correcting Words in Text, ACM Computing Surveys, vol.4, 1992, pp.377-439.

[3] Deshen S. and Liangzhi W. et al, Chinese Spelling Error Detection Based on Statistics, Computer and Communication, vol.8, 1992, pp.19-26.

[4] Chaohuang C., A Pilot Study on Automatic Chinese Spelling Error Correction, Communication of COLIPS, vol.4, 1994, pp.143-149.

[5] Yangsen Z. and Bingqing D., Automatic Errors Detecting of Chinese Texts Based on the Bi-neighborship, Journal of Chinese Information Processing, vol.3, 2001, pp.36-43.

[6] Rong L., A Chinese Spelling Check System for the OCR Output, Journal of Chinese Information Processing, vol.5, 2009, pp.92-97.

[7] Chang Chaohuang C., A New Approach for Automatic Chinese Spelling Correction, In Proceedings of the Natural Language Processing Pacific Rim Symposium, 1995, pp.278-283.

[8] Jianhua L. and Xiaolong W. The Research of Multi-Feature Chinese Text Proofreading Algorithms, Computer Engineering & Science, vol.3, 2001, pp.93.

[9] Ting L. and Hongbin S. et al, Principle of Chinese Computer Aided Proofreading System, Chinese Information, vol.2, 1997, pp.21-23.

[10] Rongxiang Y. and Kekang H., Proofreading Chinese Manuscript with Computer, Journal of Computer Research and Development, vol.5, 1997, pp.346-350.

[11] Xiaojin G. and Zhensheng L., Automatically Detecting Syntactic Errors in Chinese Texts, Computer Engineering and Applications, vol.8, 2003, pp.98-100.

[12] Yangsen Z. and Jia Z., Study of Semantic Error Detecting Method for Chinese Text, Chinese Journal of Computers, vol.4, 2017, pp.911-924.

[13] Shilong M. et al, Deep Learning with Big Data: State of the Art and Development, CAAI Transactions on Intelligent Systems, vol.6, 2016, pp.728-742.

[14] Boqing R., An Intelligent Chinese Text Proofreading Method Based on Deep Learning, vol.4, 2017, pp.55-58.

[15] Xin X., Research on the Method of Correcting Chinese Homophony Misspelling Based on Machine Translation Model, Heilongjiang University, 2017.

[16] Zhou J.and Li C., et al, Chinese Grammatical Error Correction Using Statistical and Neural Models, NLPCC-2018,cham, pp.117-128.

[17] Yongcai T. and Wenle W., et al, Text Proofreading Model with LSTM and Integrated Algorithm, Journal of Chinese Computer Systems, vol.5, 2020, pp.967-971.

[18] Yuzhong H. and Xianguo Y. et al, FASPell: A Fast, Adaptable, Simple, Powerful Chinese Spell Checker Based On DAE-Decoder Paradigm, Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), 2019.

[19] Xingyi C. and Weidi X. et al, SpellGCN: Incorporating Phonological and Visual Similarities into Language Models for Chinese Spelling Check, ACL 2020. arXivpreprint arXiv:2004.14166.

[20] Shaohua Z. and Haoran H. et al, Spelling Error Correction with Soft-Masked BERT, ACL 2020. arXiv preprint arXiv:2005.07421.

[21] Yan W. and Xiukun L.et al, Research on and Implementation of Chinese Text Proof-reading System, Journal of Harbin Institute of Technology, vol.2, 2001, pp.60-64.

[22] Yangsen Z. and Anjie T. et al, Chinese Text Proofreading for Political News Field, Journal of Chinese Information Processing, vol.6, 2014, pp.79-84+128.

[23] Zhao H. and Cai D. et al, A Hybrid Model for Chinese Spelling Check, ACM Transactions on Asian and Low-Resource Language Information Processing, vol.3, 2017, pp.1-22.

[24] Dezhi G., Research on Chinese Real-word Error Automatic Detection and Correction, Jiangsu University of Science and Technology, 2017.

[25] Junpei Z. and Chen L. et al, Chinese Grammatical Error Correction Using Statistical and Neural Models, CCF International Conference on Natural Language Processing and Chinese Computing. Berlin: Springer, 2018: 117-128.

[26] Yongmei T. and Yixiao Y. et al, Grammatical Error Correction Using LSTM and N-gram, Journal of Chinese Information Processing, vol.6, 2018, pp.19-27.

[27] Yongcai T. and Zhaoyang H. et al, Study of Chinese Word Collocation Feature Extraction and Text Proofreading, Journal of Chinese Computer Systems, vol.11, 2018, pp.2485-2490.