Research and Implementation of Chinese Text Automatic Proofreading System

Yonggang Gong, Junying Fu, Xiaoqin Lian and Yuying Li

Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing, 100048, China.
Email:765076648@qq.com

Abstract. The news media platform has a huge amount of original news releases every day, it is impractical to use manual review of text typos. This paper designed and implemented a Chinese text automatic proofreading system for large-scale text content and high-speed processing. The proofreading content is first analyzed and classified: typos and sensitive information. Firstly, the system used the n-gram model to statistically analyze the corpus after segmentation to form a 2-gram model library and a contextual context library; secondly, builded a typo confusion set, and then calculated the probability of the target word in the knowledge base to realize automatic error detection and correction of Chinese text. The system has been successfully applied to the error of the content of many government news media platforms, each server can handle one million articles every day. The results show that the recall rate of the article is 78.9% and the accuracy rate is 85.1%. It meets the demand of high speed and accurate processing of massive text error, and has important practical significance and application fields.

1. Introduction

With the advent and rapid development of the "Internet +" era, new media (microblog, WeChat and blog etc.) has become an inseparable part of people's lives. Many news media platforms have a huge amount of original news releases every day, and the timeliness of news makes it widely reposted by major media in a short period of time, and is read by hundreds of millions of netizens. Therefore, if there are bad information such as typos and political sensitive words in the news content, it may have extremely bad effects, so it is necessary to use the context-based automatic identification technology to find problems, locate problems, and solve problems in a timely and accurate manner before the manuscript is issued.

In addition, the research on the automatic proofreading system of Chinese text is still a research problem in the field of natural language[1], with far-reaching theoretical significance and research significance. The English text automatic proofreading system began in the 1960s. So far, the English text automatic proofreading system has matured and has high accuracy and recall rate[2][3]. The research on Chinese text automatic proofreading system is relatively late. Researchers began to research and explore Chinese automatic proofreading in the 1990s[4]. The methods of research mainly include: approach based on machine learning[5], semantic information[6] and probability statistics[7]. The machine learning-based method must rely on the pre-defined confusion set of typos, in fact, the error correction of Chinese text is regarded as the ambiguity resolution problem of the target word and its corresponding confusion set. There is no way to check the correctness of words that are not in the pre-defined confusion set. The method based on semantic information do not

require a pre-defined set of confusion, the essence of which is to determine the right or wrong of the target word by satisfying some semantic relationship between the target word and its context. However, this semantic association does not apply to spelling errors, and the method has a small scope of use and a high false negative rate. The method of probability and statistics is to find the typos by the n-gram model, which is based on the n-gram sequence of the large-scale corpus statistic, and calculates the probability of the word in the n-gram sequence. Low probability sequences are generally considered to be erroneous, and high probability sequences are candidate lists for error correction suggestions[8].This method is better for typos, but requires a large corpus for training. The method of n-gram statistical language model has the advantages of intelligent self-learning, fast processing speed, strong error checking ability and wide adaptability. Therefore, this paper used the method of probability statistics and rules based on n-gram model to realize automatic proofreading of Chinese text.

2. The Overall Design of the Chinese Text Automatic Proofreading System

The Chinese text proofreading system is mainly composed of five parts, namely the acquisition module, the training module, the thesaurus, the proofreading module and the API interface, as shown in Figure 1. The following mainly introduces the preprocessing module (composed of the acquisition module, the training module and the thesaurus) and the proofreading module.

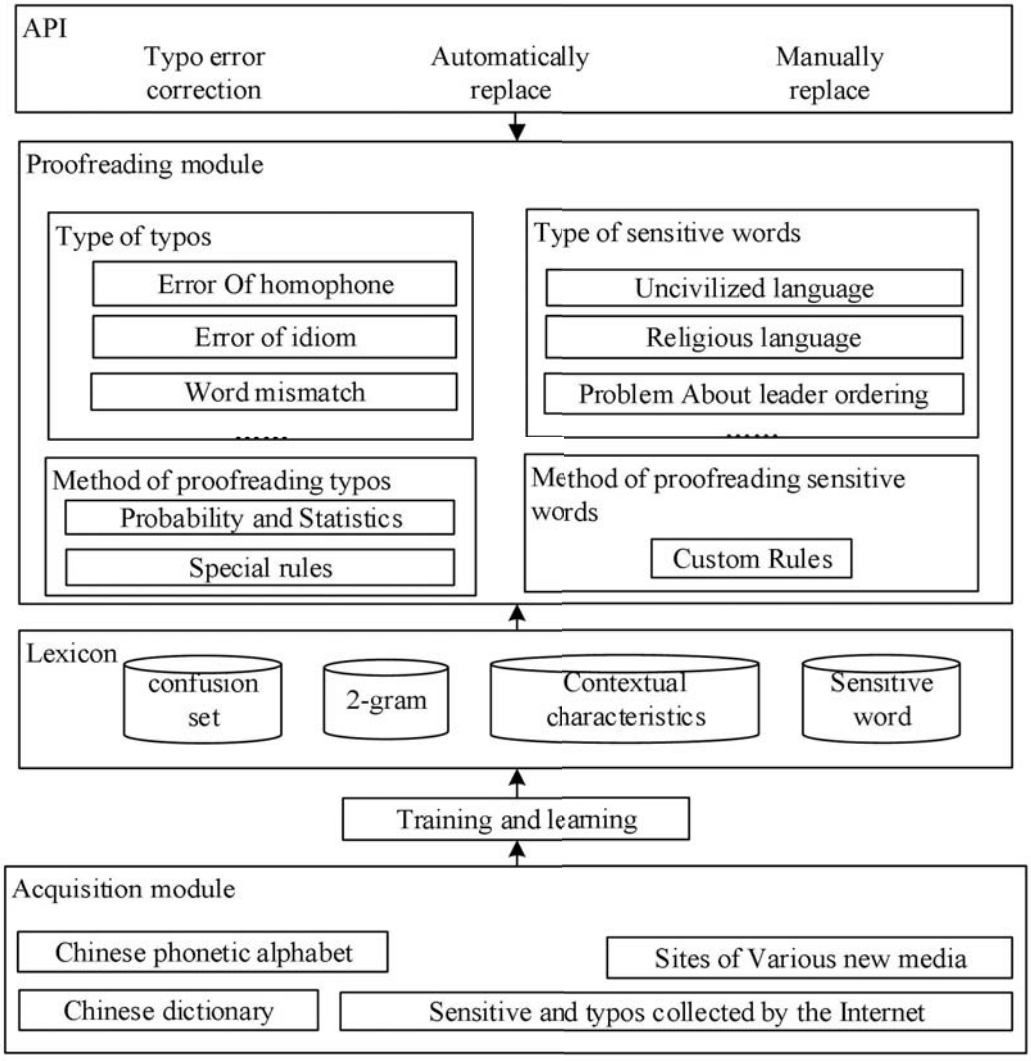

Figure 1. The overall framework of Chinese text automatic proofreading system

3. Preprocessing Module

3.1. Construction of the Confusion Set

With the development of Chinese proofreading work and the improvement of people's requirements, the typos confusion set becomes indispensable in typos recognition [9]. The automatic error detection and error correction of Chinese text must be based on the typo confusion set, so the typos confusion set is of great significance in the automatic proofreading of Chinese text. This paper constructed a typo confusion set for homophone errors and phonetic word errors that often appear in Chinese characters.

3.1.1. Construction of homonymous confusion set. Pinyin input method is currently the most used input method, because there are many words that can be selected by the same or similar pinyin, so users need to choose the correct word according to their own situation[10] . Sometimes because of the frequent input of a wrong word, the word will be ranked first in the Pinyin input method, such as : "一心一意" the word is correct and ranked at the top of the input method. However, for some reasons, it is often necessary to input "一新一意", which causes the word to be ranked at the top of the input method. When the next word is selected, it is easy to select "一新一意" due to inertia, resulting in homophone errors. It can be seen that the homophone error is an important thesaurus of the typo module. The set of a word and its corresponding homophone is a homonymous set (Homonym Dic), recorded as: $\mathrm{HD}(w)=\{w_1, w_2, w_3, \ldots, w_n\}$. E.g. The confusion set of "中国特色社会主义" is : HD(中国特色社会主义)={中国特色社会注意，中国特设社会主义，中国特色社会主仪……}. The flow chart of the homonym confusion set construction is shown in Figure 2 below:

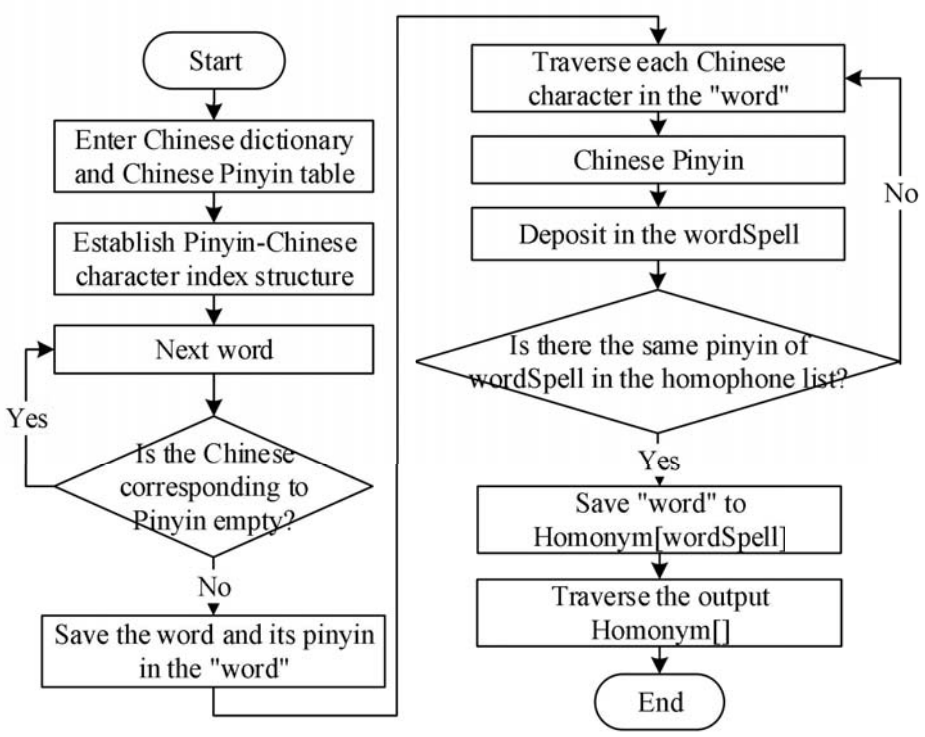

Figure 2. The flow chart of the confluence set

3.1.2. Construction of a confidence set of sound like words. The phonetic word means that the mothers of the two words are similar or identical and the finals are similar or identical[11]. For example, At the same time, the sound of "zhou" is "洲周州轴粥舟咒皱"; the sound of "zou" is "揍奏

走邹驺". By summarizing the similar phonetic words, we construct the sound like table. The construction of the confusing set of sound-like words is consistent with the construction steps of the confusing set of homophones, except that the Chinese phonetic alphabet of homophones is replaced by a sound-like table.

3.2. N-gram Module

The language model is actually a probability distribution whose formula is: $P(w_1, w_2, \ldots, w_n)$, where w_1 means a word in the language model[12].

For a sentence consisting of n words Sentence = $w_1, w_2, \ldots, w_n$, The probability is:

$$P(s) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)\ldots P(w_n \mid w_1 \ldots w_{n-1}) = \prod_i^n P(w_1 \mid w_1 \ldots w_{i-1})$$

In the formula, the probability that the i-th word w_i appears is determined by the i-1 words $w_1 w_2 \ldots w_i$ before it, and the former i-1 words are referred to as the pre-words of the i-th word. As the length of the pre-words grows, the amount of different pre-words a word may have will increase exponentially[13].

Suppose a vocabulary has a set size of L, that is, there are L different words. When the length of pre-words is i-1, the words at position i will have L^{i-1} kinds of different pre-words. At this time, in order to give the probability of the i-th word in the sentence, it is necessary to consider all the L^{i-1} kinds of different pre-words mentioned above, and the number of parameters in the statistical model will reach L^i[13]. With the difference of i, the number of parameters will also change exponentially, and the amount of calculation will increase, so this article only considers the case of i=2 or 3.

3.3. Construction of Binary For 2-gram and Contextual Context

The binary for 2-gram model is the number of simultaneous occurrences of all two adjacent words in the statistical corpus on the basis of word segmentation. i.e. in a sentence $s = w_1 w_2 \ldots w_n$, the frequency at which two consecutive words $< w_1, w_2 >, < w_2, w_3 > \ldots < w_{n-1}, w_n >$ appear at the same time. For example, in the sentense "我/rr, 的/ude1, 账号/n, 为什么/ryv, 被/pbei, 封/q, 了/ule。", counting the number of occurrences of <我, 的><的, 账号>, <账号, 为什么>…… in large-scale training corpus. This library is the most basic resource for Chinese text automatic proofreading based on n-gram model. The training corpus in this article uses the textual content of Headlines today, and use HanLP word segmentation system to segmentation words.

The construction of the contextual context's binary is the similar as that of the binary of 2-gram. The difference is that the context statistics is the probability that the word appears at the same time as the left and right words.

4. Proofreading System

The proofreading module is mainly divided into two types: typos and sensitive words. Typos consists of homophone, tone-similar words errors, idiom errors, word mismatch, more or less word's errors, double-word errors etc. The above errors can also be divided into absolute error and non-absolute error. Sensitive words consists of uncivilized words, religious words, issues of Hong Kong, Macao and Taiwan , leaders sorting problem and other political errors, etc. The identification method of the typos in this system is mainly based on the automatic proofreading method of probability statistics combining rules of the n-gram model. The identification of sensitive words needs political norms to write corresponding rules for proofreading. Due to article length restrictions, the article mainly introduces the proofreading of typos.

4.1. Absolute Error

Absolute error mainly include idiom errors, double word errors. It means that the word does not exist, that is, it must be wrong. For example, "一枝之长" is absolute error, it should be "一技之长". For this kind of error, we simply replace the typo by writing simple rules.

This system uses a double Array Trie to achieve fast query of data in mongDb[14]. Reduces the space occupied by the Trie tree, and effectively reduce the waste space of the Trie tree and ensures the efficiency of the query[15]. The following non-absolutely error's database query part also uses a double Array Trie.

4.2. Non-Absolute Error

Non-absolute error means that it is necessary to judge whether a word in the sentence is correct based on the context. For example, "'天仙宝宝'称自己因为家庭问题心理不舒服，想出来喝酒解闷", the word "心理" itself is correct, but it is wrong to put it in this sentence, so we need to judge whether the word is suitable for use here in combination with the context.

Let's take the homonym in non-absolute error as an example to introduce the proofreading process of the Chinese text proofreading system, as shown in figure 3 below.

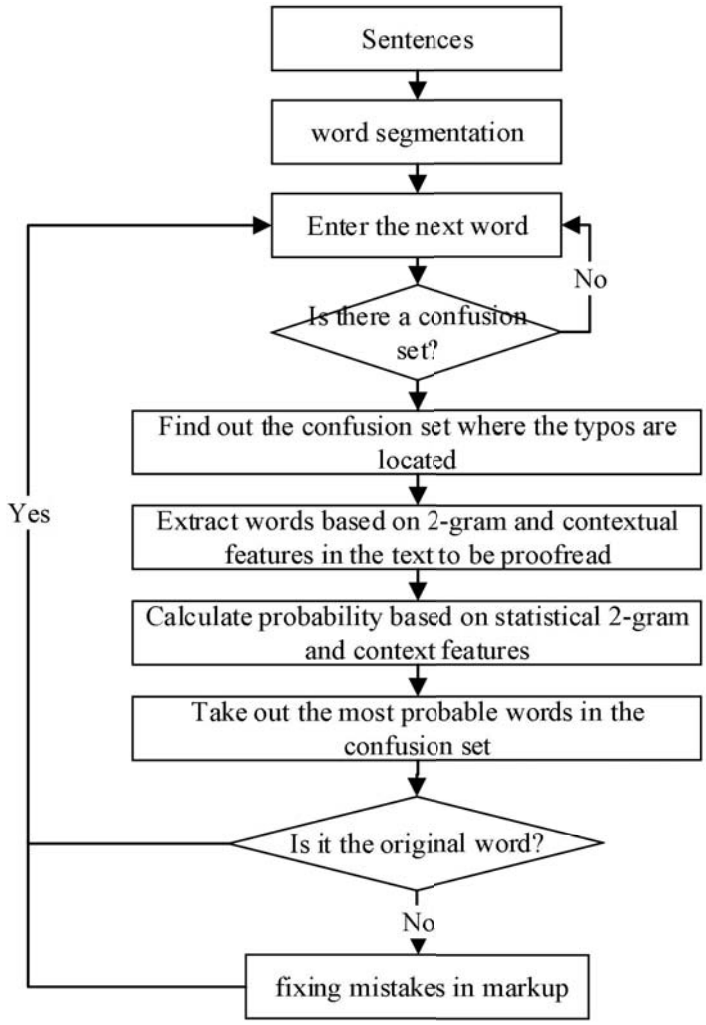

Figure 3. Chinese text automatic proofreading flow char

The formula for calculating the text support degree is as follows:

$$\sup(w_i, \theta_j) = \ln \frac{p(w_i, \theta_j)}{\sum_{k \neq i} p(w_k \mid \theta_j)}$$

$P(w_i,\theta_j)$ is the probability that w_i,θ_j occur simultaneously. θ_j is the 2-gram feature and context of w_i in the current text. In formula $P(w_i \mid \theta_j) = \frac{fre(w_i,\theta_j)}{\sum_k fre(w_k,\theta_j)+\alpha}$, the $fre(w_i,\theta_j)$ is the probability that w_i,θ_j occur simultaneously. because $fre(w_i,\theta_j)$ can be 0, we take $\alpha = 0.1^2$.

The principle of probability statistics based on n-gram model is mainly based on this phenomenon to judge whether the text to be proofread is correct or not. The idea is to obtain the confusion set of the current word and calculate the support degree of the confusion set of the current word in the context. According to the experimental results, when the first three words with higher support degree are selected as the indicators for judging the typo, the accuracy is higher. Therefore, this article will use the three words with higher support as the recommended words. If one of the first three words in the recommendation word is the same as the word in the original sentence, the word is judged to be correct, otherwise, the word is wrong, and then the word is corrected according to the recommendation word.

5. Analysis of Results

The system has the functions of typos correction, political sensitive word warning and normalization detection. It has been developed and has deployed 5 servers, each capable of processing more than 1 million articles per day. The system features massive, high-speed, and efficient processing of text.In order to evaluate the effectiveness of the system, this paper takes a certain amount of data to calculate the accuracy, recall rate and false positive rate of the system.

$$\text{the accuracy rate: } P = \frac{CorrectlyFoundTheTotalNumberOfErrors}{TheTotalNumberOfFoundErrors} \times 100\%$$

$$\text{recall rate: } R = \frac{CorrectlyFoundTheTotalNumberOfErrors}{TheTotalNumberOfErrorWordsInTheText} \times 100\%$$

$$\text{false positive rate: } F = \frac{TheTotalNumberOfErrorWordsFoundIncorrectly}{TheTotalNumberOfErrorWordInTheText} \times 100\%$$

First build a test set, build 1000 correct data and 1000 wrong data into a test set test, and there is only one error in each data. The results after statistics are shown in Table 1:

Table 1. Result analysis

	total	Found typos	Actual typos	Correctly found typos	recall rate(/%)	the accuracy rate(/%)	false positive rate(/%0
correct word	1000	150	0	0			15
typos	1000	789	1000	789	78.9		
Total	2000	939	100	789		85.1	

As can be seen from Table 1, the recall rate of the system is 78.9%, the accuracy rate is 85.1%, and the false positive rate is 15%, which has a high recall rate and accuracy. In addition, in the existing version, we will continue to improve the Chinese text automatic proofreading system through user feedback. For typos with high false positive rate and false negative rate, we will use the combination of n-gram model and special rules to reduce the false positive rate and false negative rate, thus improving the recall rate and accuracy of the whole system.

6. References

[1] Min Shi 2015 *Proc*. Chinese text automatic proofreading system *Jiangsu University of Science and Technology*
[2] Kukich K 1992 Techniques for automatically correcting words in text *Acm Computing Surveys,* **24(4)** pp 377-439.
[3] Rachele De Felice and Stephen G.Pulman A classifier-based approach to preposition and determiner error correction in L2 English. Proceeding of the 22nd International Conference on Computational Linguistics, Coling 2008, pp 167-176
[4] Jianqiang Yan and Xinbo Gao 2014 A new method of proofreading based on Google's ORC *Chinese Journal of Computers* **37(6)** pp 1261-1267
[5] Feng Hou 2010 Proc. Research on text quality intelligent assistant control technology published by Chinese newspaper industry *National University of Defense Technology*
[6] Chong Guo and Yang-sen Zhang 2010 Study on the automatic error detection of Chinese text semantic level based on the combination of Yiyuan and Yiyuan *Computer Engineering and Design* **31(17)** pp 3924-3928.
[7] Zhipeng Chen, Yuqin Lu, Huasheng Liu, Gang Liu and Hui Tu 2009 Search Error Correction of Chinese Based on N-gram Statistical Model *Journal of China Academy of Electronics and Information Technology* **4(03)** pp 323-326
[8] Liangliang Liu and Cungen Cao 2016 Research on Automatic Correction of Chinese True Word Errors Based on Combination of Local Context Features *Computer Science* **43(12)** pp 30-35.
[9] Hengli Shi, Liangliang Liu, Shi Wang, Jianhui Fu, Zaiyue Zhang and Cungen Cao 2014 Study on the Construction Method of Chinese Character Seed Confusion Set *Computer Science* **41(08)** pp 229-253.
[10] Hengli Shi 2014 Proc. Research on the construction method of Chinese character seed confusion set. *Jiangsu University of Science and Technology*
[11] Qiangze Feng and Cungen Cao 2004 Proc. The method of distinguishing voice in voice query: *China, CN1514387*.
[12] Xin Zhang 2015 Proc. Research and implementation of Chinese text proofreading method for social media. *Heilongjiang University*
[13] Tao Shen 2017 Proc. Combining N-gram model with grammatical error correction of syntactic analysis. *Southeast University*
[14] Huan Zhao and Hongquan Zhu 2009 Research on Chinese Word Segmentation Based on Double Array Trie Tree *Journal of Hunan University: Natural Science Edition* **36(5)** pp 77-80.
[15] Sili Wang, Huaping Zhang and Bin Wang 2006 Optimization of Dual Array Trie Tree Algorithm and Its Application *Journal of Chinese Information Processing* **20(5)** pp 24-30.