Introduction to a Proofreading Tool for Chinese Spelling Check Task of SIGHAN-8

Tao-Hsing Chang
Department of Computer Science and Information Engineering
National Kaohsiung University of Applied Sciences
changth@gm.kuas.edu.tw

Hsueh-Chih Chen
Department of Educational Psychology and Counseling
National Taiwan Normal University
chcjyh@ntnu.edu.tw

Cheng-Han Yang
Department of Computer Science and Information Engineering
National Kaohsiung University of Applied Sciences
1101108129@kuas.edu.tw

Abstract

The detection and correction of erroneous Chinese characters is an important problem in many applications. This paper proposed an automatic method for correcting erroneous Chinese characters. The method is divided into two parts, which separately handle two types of erroneous character: the occurrence of an erroneous character in a word length of one, and the occurrence in a word length of two or more. The first primarily makes use of a rules-based method, while the second integrates parameters of similarity and syntax rationality using a linear regression model to predict erroneous characters. Experimental results shown that the F1 and FPR of the proposed method are 0.34 and 0.18 respectively.

1 Introduction

The detection and correction of erroneous characters is a key problem in many applications. For example, approaches for information retrieval need to analyze a document's lexicon, syntax, and semantics, but the analysis of documents containing erroneous characters is likely to result in errors in the results of such analysis. Furthermore, with regard to language teaching, tools that can automatically correct erroneous characters can be of considerable assistance to a student's independent learning. To detect misspelled words within an alphabetic writing system, a dictionary method can generally be employed: if a word is not found in the dictionary and is not a newly created word, then it is incorrect. Moreover, proofreading for misspelled words can use a similarity comparison with currently available vocabulary to seek words that can correct the misspelled words.

There are great differences between the problems encountered in the automatic correction of erroneous characters in Chinese and the problems in alphabetic writing systems. Because there are no spaces between Chinese words, which would allow for their identification, it is quite difficult to use the dictionary method. Furthermore, Chinese words are composed of at least one character, so that an erroneous character may make up an existing word in combination with its adjacent characters. This results in difficulties in terms of identification. Additionally, a Chinese character may constitute a word in itself, and thus it is difficult to distinguish between a single-character word and an erroneous character. These characteristics of pictographs mean that different methods must be developed to resolve problems related to the correction of Chinese script from those used with alphabetic writing systems.

Since Chang (1995) proposed research into the automatic detection and correction of erroneous Chinese words, many methods have been advanced successively to do this. In the early stages, the most method used was that of correcting commonly confused character sets. There are three ways to establish commonly confused character sets: the first is using manually established confused character sets; the second is based on the

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 50–55,
Beijing, China, July 30-31, 2015.

statistical occurrence of biased error text corpus words composed of erroneous characters and their frequency; and the third is the method of calculating the degree of similarity so as to enter characters with similar phonetic values and forms in a list of confused character sets. The main problem with the confused character set method lies in the presence of erroneous characters that are not in confused character sets and are therefore undetectable.

The objective of this paper is to propose an automatic method for correcting erroneous Chinese characters. The method is divided into two parts, which separately handle two types of erroneous character: the occurrence of an erroneous character in a word length of one, and the occurrence in a word length of two or more. The first primarily makes use of a rules-based method, while the second integrates parameters of similarity and syntax rationality using a linear regression model to predict erroneous characters. The other sections of this paper are organized as follows: Section 2 introduces the progress made and methods used in related research in recent years. Section 3 gives a detailed explanation of the method proposed by this paper. Section 4 shows the experimental results achieved by this method in a test text corpus. Section 5 discusses the characteristics, limitations, and future research directions of this method.

2 Related Works

Proposed automated detection and correction methods for Chinese erroneous characters can be traced back to the detection and correction method put forward by Chang (1995). This method used the four commonly occurring forms of erroneous characters–"characters with similar pronunciation," "characters with similar form," "characters with similar connotation," and "characters with similar input code value"–to establish relationships of confusion between the characters. Using such databases of computer characters that may produce erroneous character relationships, it is possible to provide a list of corrections for use in attempting to detect erroneous characters and correct sentences. The input sentences use confused character sets one by one as substitutes for the Chinese computer characters in the sentence, producing a variety of possible combination sentences as candidate sentences. By calculating sentence probability based on a bi-gram model, the system seeks to obtain the optimum solution in relation to the candidate sentences that have been produced. If the optimum solution differs from the original sentence, it then compares the differing computer character and serves as the corrected result. In recent years, since some competitions have been held to correct Chinese erroneous characters, many studies have proposed a wide variety of methods to resolve this problem.

These methods can be divided essentially into three categories. The first consists of initially processing the sentence using a Chinese word segmentation tool, then detecting whether erroneous characters occur among serial single Chinese character sequences (abbreviated to SSCS below). Chang, Chen, Tseng, & Zheng (2013) searched for possible correct words among each character in an SSCS, and using the three parameters of "similarity of phonetic value," "similarity of form," and "probability of co-occurrence of adjacent characters" established a linear regression prediction model. Wang and Liao (2014) used the Chinese word segmentation system to analyze a sentence's word segments, and then, if there was a suspected occurrence of an erroneous character in a two-character word or single-character word, used a character with a high degree of similarity of phonetic value and form to replace the possible erroneous character. Finally, they used a tri-gram model to assess whether to conduct a replacement.

The second category is the direct utilization of a probability model to detect an erroneous character. Han and Chang (2013) proposed using maximum entropy in relation to 5311 characters and the seven-grams trained model to correct erroneous characters. The fundamental hypothesis of this study was: if there was a possible erroneous character in the sentence, then the matched pairs that the character and the characters preceding and following it produced may not exist in the text corpus. Conversely, if the matched pair made by the character and the character preceding it or following it is commonly seen in the text corpus, then that character's degree of erroneousness is very low here. Xiong et al. (2014) proposed using the Hidden Markov Model (HMM) as the basis for a model to detect and correct erroneous characters. This method presupposes that unknown erroneous characters exist in the sentence, and seeks out each character's substitute character by means of phonetic writing (pinyin) and the Cangjie input code using Bayes' rule as its basis. Because there are many substitute characters, this method then uses methods such as n-gram and statistics from internet search results to determine substitute words. Gu, Wang, & Liang (2014) use SSCS as their target in the same way but use character

blocks within SSCS. Exploiting the statistical method of serial computer characters forming character blocks, it is possible to detect and correct erroneous characters while not utilizing a word segmentation system.

The third method uses multiple prediction models to predict different categories of erroneous character. For example, Xin, Zhao, Wang, & Jia (2014) converted the problem of erroneous characters into the problem of seeking the shortest pathway in a graph. Because the graph model can only identify erroneous characters in long words, for erroneous single-character words it additionally uses rule-based methods and a CRF model to make corrections.

3 Methods

There are two patterns for the formation of Chinese words. One pattern is that of a character itself as a word, such as "我" (meaning 'I'), which is termed a single-character word; the other is a long word of two or more characters combined, such as "工作" (meaning 'work.') If we suppose that an erroneous character appears in a certain long word, word segmentation will break up the word into a series of single characters. Therefore, detecting whether an SSCS appears in a sentence after it has been segmented is an effective method for detecting an erroneous character. Section 3.1 of this paper is based on research by Chang et al. (2013), which proposed a method for correcting erroneous characters in long words. In section 3.2, this paper also uses the characteristics of erroneous single-character words to put forward a rules-based correction method based on syntactic structure.

3.1 Correcting erroneous characters in long words

With regard to each character of an SCSS, we hypothesize that it is not an erroneous character, and also that it may be a character in a long word. Hence, we use the dictionary method to seek out all long words containing this character. Using as an example the Chinese sentence "因_偽_他_必須_工作" (because he must work,) long words that contain the character include "因為" (because) and '因素" (factor,) etc. If we determine that "因素" is the correct word in this sentence, then "偽" is an erroneous character for "素". This paper refers to these long words as "candidate words," and refers to the candidate words' corresponding original sentence character sequence as "suspected word blocks." For example, the candidate words for the suspected word block "因_偽" include "因素".

Because there are numerous candidates for each suspected word block, it is necessary to go through a filtering process to verify whether there are words among the candidate words suitable for substituting for the suspect word block. Chang et al. (2013) noted that the majority of erroneous characters were caused by a similarity of character form or phonetic value, and thus only gave consideration to suspected word blocks where candidate words were similar in character form or phonetic value. In addition, some suspected word blocks are commonly encountered SSCSs and are not erroneous characters. Furthermore, in terms of syntactical structure, the sequence of parts of speech in some suspected word blocks sometimes makes more sense than candidate words' parts of speech within the structure of the entire sentence. Hence, the method proposed by this paper envisages four parameters: similarity of phonetic value, similarity of character form, frequency ratio, and probability ratio for parts of speech, to determine whether candidate words should be used in the correction of suspected word blocks. If a suspected word block has no candidate word within the parameters for deciding that it qualifies for correcting the word group, then it is determined that the suspected word block does not contain an erroneous character.

The first parameter is similarity of phonetic value, and the method proposed by this paper is to seek out pronunciations from all of the 37 phonetic notation symbols that are both similar and easily confused, and then to state in advance a defined degree of similarity, for example, the initial consonants "ㄅ" and "ㄆ," "ㄕ" and "ㄙ" and the vowels "ㄣ" and "ㄥ", etc. By separately calculating the difference between two characters' initial consonants, medials, vowels, and tones, it is possible to derive the degree of similarity of phonetic value between two characters. For example, the medials, vowels, and tones of the characters "讀" (to read) and "圖" (picture) are identical, but the degree of similarity of their initial consonants is 0.5; thus, the degree of phonetic similarity between the two characters is

$$(0.5+1.0+1.0+1.0)/4=0.875.$$

The second parameter is degree of similarity in terms of form. This paper proposes using the 439

basic Chinese script components and 11 types of structural relation put forward by Chen et al. (2011) and disassembling Chinese characters into a composite stroke structure. Taking the character ”大” (big) as an example, its composite stroke structure is

[{一},{月 1}+(1:5@3),{[尺 /]}~(1:5@0)~(2:3@0)].

Subsequently, the LCS-based calculation algorithm put forward by Chang et al. (2014) is utilized to calculate the degree of similarity of form between the two characters.

If the suspected word block is indeed a correct serial single-character word combination and does not contain an erroneous character, then these words should have appeared together in the broad scale text corpus. On the other hand, if there is an erroneous character within the word block, then other single-character words should appear together very rarely between the erroneous character and word block in the broad scale text corpus. Thus, if it is assumed that the suspected word block frequency of co-occurrence is FS, and the corresponding candidate word's frequency of occurrence is FT, we can use the frequency ratio of the two FT/FS to assess whether the frequency of the suspected word block is sufficiently greater than the candidate word's frequency of occurrence. If so, then the suspected word block may not contain any erroneous characters. Hence, this ratio can act as a third parameter for determining the possibility of erroneous characters occurring.

Furthermore, after a sentence undergoes a process of tagging parts of speech, the parts of speech of each word will be tagged. Generally speaking, the most common method of tagging parts of speech is that of using such probability model as HMM to seek out the various possible parts of speech sequences with the highest probability within an entire sentence. When comparing a sentence containing an erroneous character with a corrected sentence, the latter should have a higher probability value. Since sentences containing an erroneous character and corrected sentences may differ in terms of the number of words, the probability values of the two must undergo standardization before they can be compared. If we suppose that, following the probability standardization of the original sentence's parts of speech tagging, its value is PS, and the sentence following the use of candidate word correction is PT, we can use the parts of speech sequence probability ratio of the two, PT/PS, to evaluate whether the original sentence's parts of speech sequence probability is sufficiently greater than the probability for the corrected sentence. If it is, then the original sentence may not contain an erroneous character. Hence, this ratio can act as a fourth parameter for determining the possibility of occurrence of an erroneous character.

Using the above four parameter values as regression coefficients for each sentence within training materials, this paper established a linear regression model to act as a prediction model to detect and correct erroneous characters occurring in long words. If an original sentence containing a suspected word block and a corresponding candidate word's corrected sentence undergoes predictive model calculation, and the predicted value exceeds the threshold value, then it is determined that the suspected word block should be corrected using the candidate word. If the same suspected word block's multiple candidates' prediction values all exceed the threshold value, then the word with the highest predicted value is used as the corrective word.

3.2 Correction of single-character erroneous words

Unlike erroneous characters in long words, two single-character words frequently stand as a correct word and erroneous word in relation to each other, and we term this a single-character word confusion set. Words in a single-character confusion set frequently must be examined in the context of the whole sentence or even the preceding and following sentences, before it is possible to determine whether an erroneous character has occurred. Hence, it is very difficult to use a partial statistical model to correct an erroneous character. Furthermore, single-character erroneous characters may occur in any word, but erroneous characters are particularly likely to appear in some words. Thus, in light of these characteristics, this paper has adopted a rules-based method to differentiate between six types of erroneous words common in single-character word confusion sets. The six confusion sets are respectively {的、地、得, *de*}, {再、在, *zai*}, {子、字, *zi*}, {阿、啊, *a*}, {者、著, *zhe*}, {座、坐, *zuo*}, and {他、她, *ta*}.

The establishment of rules is mainly based on knowledge of grammar. For example, the character ”的” should be used between adjectives and nouns, as in for instance, “快樂的小孩” (happy child), while ”地” should be used between adverbs and verbs, as in ”飛快地奔跑” (run like lightning). Based on the characteristic usage of

these single-character words, this paper has established rules for identification of syntax in these confusion sets. The generation of these rules was summarized as possibilities following manual observation of training materials, followed by the correctness of its rules, and the state of the exceptions was verified from an extensive text corpus, before the rules were further amended. This process was repeated until the correctness of the rules reached an acceptable level. This paper established a total of 33 rules of this kind.

In addition, with regard to confusion sets {“她” (she) and “他” (he)}, we employed semantic identification rules. The basic concept that gave rise to the rules was first to seek an object referred to by a pronoun, and then decide on the correct single-character word based on the object’s gender. For example, in the text ”媽媽工作很辛苦、但是他從來不抱怨” (Mother works very hard but he never complains), the character ”他” (he) is the pronoun used for Mother, but because Mother is female it is determined that ”她” (she) should be used in order for the usage to be correct. This paper listed manually the gender of every personal noun in the dictionary as the basis for corrections.

4 Experimental Results

This method employs test data released by the Chinese Spelling Check competition held by SIGHAN-8 as its basis for evaluation. The data set is made up of 1100 sentences, of which half are completely correct sentences, and the other half are incorrect sentences containing erroneous characters. In some of the incorrect sentences, there is more than one erroneous character. Evaluation items are divided into items for detection and correction, and each item uses Accuracy, Recall, Precision, and F1-measure to evaluate the method’s effectiveness. In addition, False Positive Rate was used to calculate the proportion of correct sentences and misjudged incorrect sentences. Since the proportion of erroneous characters is not high in ordinary documents, a low false positive rate would not puzzle users. Table 1 shows the test results of this method.

	Accuracy	Precision	Recall	F1	False Positive Rate
Detection Level	0.5318	0.5745	0.2455	0.3439	0.1818
Correction Level	0.5145	0.537	0.2109	0.3029	

Table 1 Effectiveness evaluation of the method proposed in this paper

5 Discussion And Future Work

After analysis of the reasons for this method’s misjudgments, it is possible to summarize three factors.

1) This method employs rules-based handling of erroneous single-character words and it is unable to detect non-rule based erroneous characters. However, for many erroneous single-character words, it is also very difficult to use only syntactic rules detection. For example, in the wrong sentence ”我每天六天起床” (every day I get up at six days,” the character “六天” (six days) should corrected by ”六點” (six o’clock). In terms of syntax, the erroneous word does not cause a problem, and it is necessary to rely on semantic rules to handle this type of problem. However, given the results of the experiment, the formulation of semantic rules is far more difficult than that of syntactic rules.

2) Erroneous characters do not exist in an SSCS form, but rather have become constituent characters in another vocabulary. For example, in the sentence ”我聽說這個禮拜六你要開一個誤會” (I hear that you will hold a misunderstanding on Saturday,) the two-character word ”誤會” (misunderstanding) should be ”舞會” (dance party). However, ”舞會” and ”誤會” are both vocabulary words and this method cannot handle such erroneous characters that are not in SSCS.

3) Serially-occurring erroneous characters. For example, in the sentence ”可是福物生對我們很客氣” (but the *fuwusheng* [untranslateable] is very polite to us), the word ”福物生” (*fuwusheng*) is an erroneous version of ”服務生” (waiter). However, because this method’s way of defining candidate words is based on an assumption that an erroneous character is paired with a correct character, it will not classify the word ”服務” as a candidate word.

It follows that there will be two major directions for primary work to follow in the future. The first is aimed at further improving the limitations of the aforementioned three methods, and increasing the accuracy of identification. The second is exploring a single prediction model that can integrate different categories, long words, and single-character erroneous characters. Such a model would bring effective training and prediction even closer and be more stable in terms of its application.

Acknowledgements

This work is supported in part by the Ministry of Science and Technology, Taiwan, R.O.C. under the Grants MOST 103-2511-S-151-001. It is also partially supported by the "Aim for the Top University Project" and "Center of Learning Technology for Chinese" of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, R.O.C. and the "International Research-Intensive Center of Excellence Program" of NTNU and Ministry of Science and Technology, Taiwan, R.O.C. under Grant MOST 104-2911-I-003-301.

References

Chang, C. H. 1995. A new approach for automatic Chinese spelling correction. *Proceedings of Natural Language Processing Pacific Rim Symposium*, 95:278-283.

Chang, T. H., Chen, H. C., Tseng, Y. H., & Zheng, J. L. 2013. Automatic detection and correction for Chinese misspelled words using phonological and orthographic similarities. *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing* (SIGHAN-7), 97-101.

Chen, H. C., Chang, L. Y., Chiou, Y. S., Sung, Y. T., & Chang, K. E. 2011. Construction of Chinese Orthographic Database for Chinese Character Instruction. *Bulletin of Educational Psychology,* 43:269-290.

Gu, L., Wang, Y., & Liang, X. 2014. Introduction to NJUPT Chinese Spelling Check Systems in CLP-2014 Bakeoff. *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing* (CLP-2014), 167-172.

Han, D., & Chang, B. 2013. A Maximum Entropy Approach to Chinese Spelling Check. *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing* (SIGHAN-7), 74-78.

Wang, Y. R., & Liao, Y. F. 2014. NCTU and NTUT's Entry to CLP-2014 Chinese Spelling Check Evaluation. *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing* (CLP-2014), 216-219.

Xin, Y., Zhao, H., Wang, Y., & Jia, Z. 2014. An Improved Graph Model for Chinese Spell Checking. *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing* (CLP-2014), 157-166.

Xiong, J., Zhao, Q., Hou, J., Wang, Q., Wang, Y., & Cheng, X. (2014). Extended HMM and Ranking models for Chinese Spelling Correction. *In Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing* (CLP-2014), 133-138.