# Automatic Detection and Correction for Chinese Misspelled Words Using Phonological and Orthographic Similarities

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

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

**Yuen-Hsien Tseng**
Information Technology Center
National Taiwan Normal University
samtseng@ntnu.edu.tw

**Jian-Liang Zheng**
Department of Computer Science
and Information Engineering, National
Kaohsiung University of Applied Sciences
chaosgodyi@gmail.com

## Abstract

How to detect and correct misspelled words in documents is a very important issue for Mandarin and Japanese. This paper uses phonological similarity and orthographic similarity co-occurrence to train linear regression model. Using ACL-SIGHAN 2013 Bake-off Dataset, experimental results indicate that the detection F-score, error location F-score of our proposed method for Subtask 1 is 0.70 and 0.43 respectively, and the correction accuracy of the proposed method for Subtask 1 is 0.39.

## 1 Introduction

How to automatically detect and correct misspelled words in documents is a very important issue. It is not an easy task for programs to spot misspelled words automatically. In English sentences, words are separated by space, thereby leading to the result that it is not difficult to distinguish if there are characters with non-existing orthography and unknown words. However, Chinese sentences are constructed by successive single-character, and a word could consist of one character or more. As a result, it is difficult to identify whether a character is a part of a misspelled word or not.

Based on our observation, misspelled words mainly occur as the following cases: phonological similarity and orthographic similarity. For example, word ‘已經’ is mistakenly written as ‘以經’ due to the fact that characters ‘已’ and ‘以’ are pronounced as ‘yi’. In addition, word ‘代表’ is mistakenly written as ‘伐表’ because the orthographic of characters ‘代’ and ‘伐’ are quite confusing. As a result, it may work to identify the possible misspelled words within sentences by phonological similarity and orthographic similarity between two characters.

The purpose of the study is to propose a method to detecting and correcting misspelled words in sentences. The proposed method does not rely on the collection of similar words, but based on the following assumption. Supposing there was no misspelled word in sentences, ideal word segmentation method could divide sentence into serial correct words. However, if there was a misspelled word, the segmentation could separate words containing misspelled character by serial characters. For instance, sentence ‘我們都喜歡學佼’ will be segmented into ‘我們 都 喜歡 學 佼’ due to the fact that ‘學佼’ cannot be found in the dictionary, thus segmenting ‘學’ and ‘佼’ respectively.

By the observation mention above, a sentence may include several character sequences consisting of two or more than two characters, denoted as sentential fragments. Each character in fragments may be the wrong part of a misspelling word while other characters are the correct part of the word. Hence, for each character treated as correct part of a misspelled word, the proposed method picks up the words containing the character. The words will be denoted as “candidate words”. On the other hand, all characters in the fragment may be single-character words. The sentential fragment referring to candidate words is called “original string” in this

*Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7)*, pages 97–101,
Nagoya, Japan, 14 October 2013.

paper. By calculating the probability of candidate words and original strings, the proposed method can determine whether the original strings contain misspelled words or not and correct the words. Next section will address the details.

## 2 Related Works

Chang (1995) proposed detecting technique of Mandarin misspelled word. Although it was able to find out misspelled words, there were some defects needed to be improved. For example, too much False Alert, long detection time, not able to refer to entire paragraph. Ren et al. (2001) utilized rule-based with linguistic model to detect mistakes. Although it was not very efficient, it was a new concept at the time. Lin et al. (2002) focused on misspelled words occurred in Cangjie input method and put forward a detecting system. Huang et al. (2008) designed a correcting system for wrong phonological words which built up similar phonological word collection for every single word. The correcting system also used bi-gram linguistic model to position the misspelled word, and replaced it with the most likely fit word.

Afterwards, there were many proposes under different circumstances. For example, Chen (2010) following previous studies, he amended detecting templates in order to automatically generate positive and negative knowledge corpus by Using Template and Translate modules to correct sentences. And final correction was conducted by part-of-speech Language Model to improve the accuracy probability of misspelled word correction.

## 3 Methods

Chang et al. (2012)’s approach is refined as the algorithm for automatically correcting misspelled words in this paper. They observed that there is a specific phenomenon when misspelled words occur. They envisaged that there was no misspelled words in sentences, ideal tokenization system would divide sentence into correct vocabulary combinations. However, if there is a misspelled word, the system would segment the majority of vocabulary contained misspelled words by means of single-character formation.

According to this property, existing misspelled words was assumed to appear in a string formed by successive single-character words in this paper. As a result, for a string including two or more than two single-characters, the words which contain some characters in a string from the dictionary can be listed. In this study, these words are called candidate word while the string is called original string.

Linear regression prediction model was used to determine whether an original string should be replaced with a candidate word or not. Three parameters between candidate word and original string are used in linear regression model as an input. The values of three parameters are respectively called similarity, the probability of character co-occurrence, and the probability of POS co-occurrence. The values are utilized as the input in linear regression formula, and then the probability of misspelled words in original string can be obtained. If several candidate words are predicted as the correct words of original string by prediction model at the same time, the word with highest score is treated as correct word.

The similarity between candidate word and original string is the average between phonological similarity and orthographic similarity. The Mandarin phonetic code was employed to compute phonological similarity. High similarity means that the two characters are easily represented as misspelled words for each other. On the other hand, radical structures are utilized to determine spatial structure of two fonts. Two characters with higher graphic resemblance easily represents as misspelled words. In the following sections, the detail of similarity will be addressed.

### 3.1 Candidate Words

For each sentence, it is segmented into words and the part-of-speeches of words are tagged by WeCAn system (Chang et al., 2012). Based on the assumption mentioned earlier in this paper, words which contain misspelled words can result in consecutive single-character string. Hence, the model will identify all words contained in consecutive single-character string from dictionary. As seen from Figure 1, a sentence ‘人生又何償不是如此’ is segmented into ‘人生_又_何_償_不是_如此’. ‘償’ is a misspelled words of ‘嘗’ in this sentence, ‘何’ and ‘償’ are thus segmented respectively. Followed by, for the string with successive single-character ‘又何償’, the system will select the candidate word from each character in the string. For the ‘何’ in string, the proposed method identifies ‘何’ as the second word and its length less than or equal to 3 in the dictionary, such as ‘任何’, ‘如何’ and so on. Additionally, the word ‘何’ identified as the first word and its length less than or equal to 2 is also the candidate word, such as '何必', '何嘗'and so on.

All candidate words will be compared to correspondent original string with their phonological similarity and orthographic similarity. In Figure 1, the phonological similarity and orthographic similarity between the character '任' in candidate word '任何' and character '又' in original string will be computed. The similarities determine whether '任何' is needed to be further analyzed.

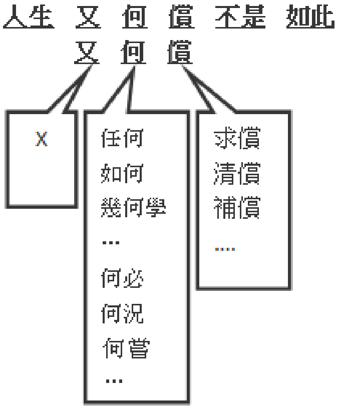

Figure 1. An Example of Candidate Words and an Original String.

### 3.2 Phonological Similarity

Mandarin phonetic symbols are used to evaluate phonological similarity. There are 37 symbols in Mandarin phonetic symbols dividing into initial (ㄅ/b/,ㄆ/p/,ㄇ/m/), medial (ㄧ /yi/,ㄨ /wu/,ㄩ/yu/), final (ㄚ/a/,ㄛ/o/,ㄜ/e/) and five tones. Chang et al.(2010) mentioned that some Chinese phonetic alphabets have identical articulation method and speech position, whereas articulation confusability is the causes of misspelled words. For example, symbols 'ㄅ' and 'ㄆ' have similar speech position ; symbols 'ㄕ' and 'ㄙ' have identical articulation; symbols 'ㄣ' and 'ㄥ' in final category belongs to a confusable articulation set.

This paper compares two characters with their Mandarin phonetic symbols of its initial, medial, final and tone respectively to measure phonetic similarity. The rules for comparison are as follows:

1. If there are identical initial, a similarity score will achieve one point. The score will achieve 0.5 point for initials of two characters which are 'ㄅ' and 'ㄆ', 'ㄉ' and 'ㄊ', 'ㄋ' and 'ㄌ', 'ㄍ' and 'ㄎ', 'ㄓ' and 'ㄗ', 'ㄔ' and 'ㄘ', or 'ㄕ' and 'ㄙ'.
2. If there are identical finals, the score will increase one point. The score will increase 0.5 point for finals of two characters which are 'ㄣ' and 'ㄥ'.
3. If there are identical medial in two characters, the score will increase one point.
4. If the tones are consistent in two characters, the score will increase one point.
5. Phonetic similarity between two characters can be obtained by dividing the similarity score by 4.

Phonological similarity between candidate word and original string is the average of phonological similarity of all characters in the candidate word. For instance, given candidate word '應該' corresponding to original string '因該', phonetic symbol of characters '因' and '應' is 'ㄧㄣ' and 'ㄧㄥ' respectively. Hence, the phonological similarity is (1+1+0.5+1)/4 = 0.875. The similarity between the same two characters '該' is one. Therefore, phonological similarity between candidate word '應該' and original string '因該'is (1+0.875)/2=0.9375.

### 3.3 Orthographic Similarity

Measurement of orthographic similarity in this paper is based on the method proposed by Chang et al. (2012). The measurement first disassembles two characters into a set of basic components and compare the differences between the two using Chinese Orthography Database proposed by Chen et al. (2011). There are 446 basic constituents in the database, and each unit of a character is linked by their spatial relations. There are 11 types of spatial relations, such as vertical combination and horizontal combination.

Through the database, a Chinese character can be converted into a series of branch-like structure consisting of parts and combination relations. The structure is called the constituent structure. Figure 2 shows the constituent structure of the Chinese character '查' in which '-' represents horizontal combination, and '木', '曰', '一' are constituents.

In the constituent structure of a character, nodes represent combination relations while leaves represent constituents. Every relation and constituent in the whole structure has a level representing the position as well as a weight representing the strokes for that constituent. The weight for each relation denotes the total number of strokes for all relevant constituents. Chang et al.(2012) used the level and weight of constituent structure of two characters to calculate the degree of orthographic similarity.

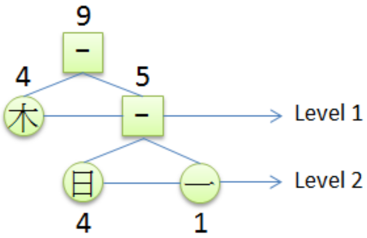

Figure. 2. Constituent Strucutre in the Chinese Character ‘查’.

This paper modifies measuring formula for similarity in previous study Chang et al. (2012). In previous formula, only the similarity between two identical constituents is scored. It is suggested that two similar constituents should increase the score. Therefore, this paper uses the stroke information in the database for constituents to calculate the similarity between two constituents. For example, stroke information of constituents ‘已’ and ‘己’ in the database is as follows:

‘已’ :[{冂 2},{一}~(1:9@9),{㇄}~(2:0@2)]
‘己’: [{冂 2},{一}~(1:9@9),{㇄}~(2:0@0)]

It is noted that most constituents for Chinese characters ‘已’and ‘己’ are very similar which would receive a score close to 1 in similarity measure.

### 3.4 Probability of Co-occurrence

In large corpuses, some specific characters may have high frequency to be adjacent to another character, and this is called co-occurrence. The probability of co-occurrence between two characters is called probability of character co-occurrence (PCC). If a sentence has misspelled words, the PCC among characters in the sentence should be lower than the sentence which has no misspelled words. Hence, if the PCC of character in the candidate word is great higher than that in original string, misspelled words may occur in the sentence. In addition, there exists co-occurrence in part-of-speeches. The probability of co-occurrence between two characters is called probability of character co-occurrence (PPC).

Bi-gram method is utilized in both this paper and the previous study to calculate the PCC of characters in candidate words as well as that in original string. Ratio of PCC (RPCC) can be obtained by dividing the PCC of characters in candidate word by that in original string. The ratio of the probability of part-of-speech co-occurrence (RPPC) can be obtained by the same methods. The higher the two values, the possible the misspelled words occur in original string. As a result, both ratios will be two inputs for prediction model.

### 3.5 Prediction Model

This paper adopts linear regression formula to be the prediction model. For a candidate word and correspondent original string, the values of three inputs can be obtained by using approaches in subsection 3.2 to 3.4. Candidate words and correspondent original string in training data are utilized in this paper to compute each regression coefficient in formula 1.

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \qquad (1)$$

For any set of candidate words and correspondent original string, three parameters are substituted into formula 1 to obtain *y*. The *y* value represents the probability of the original string within misspelled words. Based on the values of *y* obtained from training data, a threshold can be set. If *y* is higher above the threshold, there are misspelled words in original string. If *y* is lower than threshold value, there are no misspelled words in original string.

## 4 Experiments

This paper use the data provided by ACL-SIGHAN 2013 Bake-off to conduct performance evaluation. The data is divided into two sets called ‘dry run’ and ‘final test’ while the evaluation includes two tasks sub-task 1 and sub-task 2. Each set consist of two subsets which is employed to evaluate the performance of methods for two tasks respectively. In dry run, sub-task 1 and sub-task 2 each use 50 example sentences for testing. In final test, sub-task 1 and sub-task 2 each use 1000 example sentences for testing. The purposes of sub-task 1 and sub-task 2 are to respectively evaluate the performance of error detection and error correction of methods.

Table 1 presents the evaluation result of Sub-task 1 in our proposed method, denoted as KUAS-NTNU, and results of sub-task 2 are shown in Table 2. Since SIGHAN has not reported the F-score for sub-task 1 in dry run, Table 1 does not show the detection F-score and error location F-score of dry run.

| | Dry Run | Final Test |
|---|---|---|
| False-alarm Rate | 0.23 | 0.23 |
| Detection Accuracy | 0.80 | 0.79 |
| Detection Precision | 0.50 | 0.61 |
| Detection Recall | 0.90 | 0.82 |
| Detection F-score | - | 0.70 |
| Error Location Accuracy | 0.76 | 0.69 |
| Error Location Precision | 0.39 | 0.38 |
| Error Location Recall | 0.70 | 0.51 |
| Error Location F-score | - | 0.43 |

Table 1. The Performance of KUAS-NTNU System for Subtask 1.

| | Dry Run | Final Test |
|---|---|---|
| Location Accuracy | 0.30 | 0.44 |
| Correction Accuracy | 0.28 | 0.39 |
| Correction Precision | 0.36 | 0.51 |

Table 2. The Performance of KUAS-NTNU System for Subtask 2.

## 5 Discussion

Methods suggested by previous studies often rely on data collected from confusable character sets. Although corresponding characters in the set are high in similarity and can be easily confused, they could not be assessed correctly if they are not from the confusable sets. Our proposed methods calculate phonological similarity and orthographical similarity between misspelled words and original string, which are not restricted by confusable sets. The proposed method can still obtain a reliable estimation by other parameters with characters in low similarity.

Some issues and works could be explored and developed in the further. First, this study only examines characters with misspelled words. The detection and correction of single-character misspelled words only rely on simple rule-based approaches. It results in many single-character misspelled words cannot be extracted. Second, collections of unknown words in sentences are often considered having misspelled words which might cause a decrease in system accurate rate but an increase in false-alarm rate. Differences analysis for unknown words and misspelled words are issues that must be dealt with in future research.

### Acknowledgments

This research was partially supported by National Science Council (NSC), Taiwan, under the grant NSC 100-2511-S-003-053-MY2 and NSC 102-2511-S-151-002, and the “Aim for the Top University Project” of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan. We are also grateful to the support of International Research-Intensive Center of Excellence Program of NTNU and NSC, Taiwan, R.O.C., under the grant 101WFA0300229.

## 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 C.-H., Lin S.-Y., Li S.-Y., Tsai M.-F., Liao H.-M., Sun C.-W., and Huang N.-E. 2010. *Annotating Phonetic Component of Chinese Characters Using Constrained Optimization and Pronunciation Distribution.* International Journal of Computational Linguistics and Chinese Language Processing, 15(2):145-160.

Chang T.-H., Su S.-Y., and Chen H.-C. 2012. *Automatic Correction for Graphemic Chinese Misspelled Words.* Proceedings of ROCLING 2012, 125-139.

Chang T.-H., Sung Y.-T., & Lee Y.-T. 2012. *A Chinese word segmentation and POS tagging system for readability research.* Proceedings of 42nd SCiP.

Chen H.-C., Chang L.-Y., Chiou Y.-S., Sung Y.-T., and Chang K.-E. 2011. *Chinese Orthography Database and Its Application in Teaching Chinese Characters.* Bulletin of Educational Psychology, 43:66-86.

Chen, Y.-Z. 2010. *Improve the Detection of Improperly Used Chinese Characters with Noisy Channel Model and Detection Template.* Chaoyang University of Technology, Taiwan, R.O.C.

Huang C.-M, Wu M.-C., and Chang C.-C. 2008. *Error Detection and Correction Based on Chinese Phonemic Alphabet in Chinese Text.* World scientific publishing company, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 16(1):89-105.

Lin Y.-J., Huang F.-L., and Yu M.-S. 2002. *A Chinese Spelling Error Correction System.* Proceedings of the 7th Conference on Artificial Intelligence and Applications.

Ren F., Shi H., and Zhou Q. 2001. *A hybrid approach to automatic Chinese text checking and error correction.* Proceedings of IEEE SMC, 1693-1698.