Improving the Template Generation for Chinese Character Error Detection with Confusion Sets

Yong-Zhi Chen*, Shih-Hung Wu*, Ping-che Yang†, and Tsun Ku†

Abstract

In this paper, we propose a system that automatically generates templates for detecting Chinese character errors. We first collect the confusion sets for each high-frequency Chinese character. Error types include pronunciation-related errors and radical-related errors. With the help of the confusion sets, our system generates possible error patterns in context, which will be used as detection templates. Combined with a word segmentation module, our system generates more accurate templates. The experimental results show the precision of performance approaches 95%. Such a system should not only help teachers grade and check student essays, but also effectively help students learn how to write.

Keywords: Template Generation, Template Mining, Chinese Character Error.

1. Introduction

In essays written in Chinese by students, incorrect Chinese characters are quite common. Since incorrect characters are a negative factor in essay scoring, students should avoid such errors in their essays. Our research goal is to build a computer tool that can detect incorrect Chinese characters in student essays and correct them, so that teachers and students can learn faster with help from the computer system.

Compared with the detection of spelling errors in English, the detection of incorrect Chinese characters is much more difficult. In English, a word consists of a series of letters while a meaningful Chinese word usually consists of 2 to 4 Chinese characters. The difficulty lies partly in the fact that there are more than 5,000 high-frequency characters.

In previous works on Chinese character error detection systems (Zhang, Huang, Zhou, &
Pan, 2000) (Ren, Shi, & Zhou, 1994), a confusion set for each character is built and is used to
detect the character error with the help of a language model. The confusion set is based on a
Chinese input method. The characters that have similar input sequences probably belong to the
same confusion set. For example, the Wubizixing input method (Wubi), which is a Chinese
character input method primarily for inputting both simplified and traditional Chinese text in a
computer, is used in (Zhang, Huang, Zhou, & Pan, 2000). The Wubi method is based on the
structure of the characters rather than on the pronunciation. It encodes every character in four
keystrokes at the most. Therefore, if one keystroke is changed, another character similar to the
correct one will show up. Once a student chooses the similar character instead of the accurate
one, a character error is established, and a confusion set is automatically generated by the
character error. Another approach is to manually edit the confusion set. Common Errors in
Chinese Writings gives 1477 common errors (National Languages Committee, 1996).
Nevertheless, this amount is not sufficient to build a system. Hung manually compiled 6701
common errors from different sources (Hung & Wu, 2008). These common errors were
compiled from essays of junior high school students and were used in Chinese character error
detection and correction.

Since the cost of manual compilation is high, Chen et al. proposed an automatic method
that can collect these common errors from a corpus (Chen, Wu, Lu, & Ku, 2009). The idea is
similar to template generation, which builds a question-answer system (Ravichandran & Hovy,
2001) (Sung, Lee, Yen, & Hsu, 2008). The template generation method investigates a large
corpus and mines possible question-answer pairs. Templates for Chinese character error
detection can be generated and tested by the chi-square test on the basis of a large corpus. In
this paper, we will further improve the methods for building confusion sets and automatically
generating a template.

According to recent studies (Liu, Tien, Lai, Chuang, & Wu, 2009a; 2009b), character
errors in student essays are of four major types: errors in which characters have similar shapes
(30.7%), errors in which characters have similar pronunciation (79.9%), errors in which the
two previous types are combined (20.9%), and other errors (2.4%). Therefore, an ideal system
should be able to deal with these errors, especially those resulting from similar pronunciation
and similar character shapes. The confusion set for similar pronunciation is relatively easy to
build, whereas the confusion set for similar shapes is more difficult. In addition to the Wubi
input method, the Cangjie input method is also used to compile confusion sets (Liu & Lin,
2008).

The paper is organized as follows. In Section 2, we introduce the system design and
related works. In Section 3, we describe a new process of template generation. Section 4
describes the experimental procedure and the data. Finally, in Section 5, we give the
conclusion and propose our future research.
2. System Design

2.1 Chinese Character Error Detection and Correction System

The system that can detect and correct Chinese character errors works as follows. First, it needs a student to input an essay. The system then reports the errors in the essay and gives suggestions on correction, as shown in Figure 1. Such a system uses templates that can detect whether common errors have occurred. A template consists of a pair of words, a correct one and an error one, such as “辯論會”-“辨論會”. For example, if the error template “辨論會” is matched in an essay, our system can conclude that there is an error and make a suggestion on correction to “辯論會”.

Figure 1. System function of Chinese character error detection in an essay

In previous works, these templates were compiled manually (Liu, Tien, Lai, Chuang, & Wu, 2009b). The quality of the manually-edited templates is high. Nevertheless, the method is time-consuming and costs too much manpower. Therefore, an automatic template generation method based on the context of errors was proposed in 2009 (Chen, Wu, Lu, & Ku, 2009), several examples of automatically generated tri-gram and four-gram templates are shown in Figure 2. The automatic template generation method is less costly; however, it does not accommodate conventional vocabulary. The template generation method has a serious drawback. In Figure 2, we find that several templates contain unrecognizable words, such as “辯護律”,”“視辯論,” and “電視辯,” which are trigrams of Chinese characters that do not have
any meaning. These templates can be used to detect character errors, but are not suitable for suggesting corrections.

In the following subsections, we will propose a new method to avoid this drawback.

### Table: Templates for Error Detection and Correction

| Correct | Error |
|---------|-------|
| 會議長 | 會首常 |
| 會給予 | 會給予 |
| 辯論會 | 辯論會 |
| 辯護律 | 辯護律 |
| 的辯論 | 的辨論 |
| 視辯論 | 視辨論 |
| 電視辦 | 電視辦 |
| 半世紀 | 辦世紀 |
| 半以上 | 辦以上 |
| 半個小 | 辦個小 |

| Correct | Error |
|---------|-------|
| 清潔隊長 | 清潔隊常 |
| 交通隊長 | 交通隊常 |
| 辯護律師 | 辯護律師 |
| 視辯論會 | 視辯論會 |
| 公開辯論 | 公開辯論 |
| 半個世紀 | 辦個世紀 |
| 一年半的 | 一年辦的 |
| 的另一半 | 的另一辦 |

**Figure 2. The templates for error detection and correction in (Chen, Wu, Lu, & Ku, 2009)**

### 2.2 Confusion Set

The first step in template generation is to replace one character in a word with a character in the corresponding confusion set. For example, by replacing one character in the correct word “芭蕉,” we get a wrong word “笆蕉”. Such a correct-wrong word pair is used as the template for error detection and correction suggestion.

According to Liu et al. (Liu, Tien, Lai, Chuang, & Wu, 2009a; 2009b), the most common error types are characters with similar shapes and characters with similar pronunciation. The percentage of these two types of errors combined is 89.7% of all errors. Therefore, the confusion set should deal with characters with similar pronunciation and shapes.

We first compile all of the characters that have the same pronunciation from a dictionary and make them the elements of a confusion set. For example, “八(ba1)” and “巴(ba1)” have the same pronunciation. Therefore, they belong to the same confusion set. To reduce the size of the confusion set, we treat characters with different tones as belonging to different sets, even though they sound similar. For example, “罷(ba4)” is not in the confusion set of “八(ba1)”. We formed 1,351 sets with a total of 15,160 characters, as shown in Figure 3.

In this paper, we use a simple rule to compile characters with similar shapes. In the first book on Chinese characters, known as Shuowen Jiezi (說文解字) (Xu, 2009), in the second
century, radicals (部首) were used to categorize characters. We use the key component of a character, its radical, as the basic shape of the character to find the characters with the same radicals. There are 214 radicals in Chinese, according to the Kangxi Dictionary (康熙字典) (Zhang, 1999). Therefore, we compile 214 confusion sets with a total of 9,752 different characters. Figure 4 shows some examples.

After constructing the confusion sets, our system can find characters with the same pronunciation and characters with similar shapes for any character that is input. For example, given a character “兇,” the system can find characters with the same pronunciation “凶兄匈洶恟胸,” and characters with similar shapes “兄兆兄兊兄兊,” as shown in Figure 5. This is a crucial step of our new template generation.

| Zhuyin | Pinyin | Characters |
|--------|--------|------------|
|ㄅㄚ | ba1 | 蝦讨八巴佭呎亜殞讨砳砳砳砳砳 |
|ㄅㄚˊ | ba2 | 銖錐拔駅敲敲敲敲敲敲 |
|ㄅㄚˇ | ba3 | 笆把把 |
|ㄅㄚˋ | ba4 | 伯罷粑粑爸爛爛把把 |
|ㄅㄛ | bo1 | 剝嚗波袚玻柭砵缽啵菠碆播播鱍岥播襏 |
|ㄅㄛˊ | bo2 | 爆伯犮箋柏賭賭賭賭賭 |
|ㄅㄛˇ | bo3 | 簆皚 |
|ㄅㄛˋ | bo4 | 播蝻華毫賬賬北振葉葉葉 |

Figure 3. Examples of characters in confusion sets

| Radicals | Characters |
|----------|------------|
| 一 | 一丐丁之三下万万瓦丑丐丐丐丐丐丐丐 |
|丶 | 丸凡丹主 |
|丿 | 丿又乃么么之乍乍乎乒乓丰乖乖 |
|乙 | 乙九乜也乞乱乱乱乱乱 |
|丅 | 了子事 |
|二 | 二于云井互五亓瓦瓦瓦瓦瓦 |
|丷 | 亡亢交亦亥亭亭亭亭亭亭 |
|人 | 人仁什仃仆仇仍今介灰伤伤以付仔仕他他 |
|儿 | 儿元允充兄光兊兄兄兄兄兊兊兊兊兊兊兊 |
|入 | 入内全全 |

Figure 4. Examples of characters in confusion sets
2.3 Automatic Template Generation

Figure 6 shows the flowchart of our automatic template generation process. The basic assumption is that the corpus might contain more correct words than wrong ones. Therefore, our system first replaces one character in the correct words to form the corresponding wrong words. Then, our system checks the frequency of the words in the corpus. If the replacement creates a word with a relatively high frequency, we do not treat it as a wrong word.

As we mentioned in Section 2.1, the automatically-generated templates might not be suitable for suggesting corrections. To overcome this drawback, we use existing vocabulary, instead of n-gram character sequences, as the candidate for a template. There are 145,608 words in the MOE dictionary (Ministry of Education, 2007). We treat them as the seeds of the templates. In our experiment, we focus on 4,998 high-frequency characters that were compiled on the basis of a 1998 survey (National Languages Committee, 1998).

Our system generates templates by checking each high-frequency character and finding all of the words that contain the character. Then, the system replaces the character in each
word with a character in the corresponding confusion set. The correct-wrong word pair undergoes a simple statistical test. If it passes the test, it will be kept as a template; otherwise, it will be discarded. The statistical test is based on the frequency of each word in the pairs appearing in a large corpus. To prevent the process from generating controversial templates, our system also conducts a close test. The close test checks whether the new template will cause a false alarm on our old test data. The template that generates conflicting templates will also be discarded. The close test threshold is set to 0, which means any template that might cause a false alarm will not be used. A template generation example is shown in Figure 7.

The statistical test in our system is not a rigid test. We tune the threshold of relatively high frequency based on two formulae. One is adopted from the chi-square test, and the other one is from our observation. The first test is a simplified \((n=1)\) chi-square test used in a previous work (Hung & Wu, 2008):

\[
X^2 = \frac{(O - E)^2}{E},
\]

where \(E\) is the frequency of a correct word and \(O\) is the frequency of a wrong word. To avoid further disputation, we assume that \(E > O\) in our study. The chi-square test provides a threshold mechanism to decide whether a correct-wrong pair is a proper template or not.

In this study, we suggest the test should be like Equations (2) and (3).

\[
\sqrt{C_{freq}} > W_{freq}, \quad C_{freq} > AveragFreq
\]

\[
Threshold = \frac{\sum_{i=1}^{n} C_{vocabulary}(i)}{n},
\]
where $C_{feq}$ is the frequency of the correct word, $W_{feq}$ is the frequency of the wrong word, and $AverageFreq$ is the average of the frequencies of all correct words.

If the frequency of the correct word is higher than the threshold and if the square root of the frequency of the correct word is higher than the frequency of the wrong word, then the pair passes the test.

We have found that the templates that do not pass the test are also the ones that will cause false alarms; for example, the pairs “未來”-“為來,” “已經”-“以經,” and “但是”-“但事”. When the context is different, these templates do not always give correct detection results and cause false alarms.

### 2.4 Word Segmentation

As in the examples above, short templates with only two characters could cause false alarms. The reason is that, when we treat words as bi-gram character sequences, many word boundaries may be unclear. For example, as shown in Figure 8, the template “擁有”-“雍有” can be used to detect and correct the first sentence, “一個人可雍有很多快樂”, in which one of the word pair appears, but the template “擁有”-“以有” cause a false alarm in the second sentence, “一個人可以有很多快樂”. We find that this failure can be avoided by using correct word segmentation. The character “以” should be a part of the previous word “可以”. If we have enough confidence in the word segmentation, then the characters in a segmented word should not be candidates for character error detection.

![Figure 8](http://ckipsvr.iis.sinica.edu.tw/)

**Figure 8. A false alarm in the second sentence for a short template “擁有”-“雍有” and “擁有”-“以有”**

We assume that a word segmentation tool can give the correct results for normal input sentences and does not segment sentences with wrong character sequences into words. Figure 9 shows the segmentation results of the two sentences shown in Figure 8. In our experiment, we used the segmentation tool provided by CKIP, Academia Sinica. With the help of this segmentation tool, our system can compile more accurate short templates. Some short templates are shown in Figure 10.

---

1 http://ckipsvr.iis.sinica.edu.tw/
Improving the Template Generation for Chinese Character Error Detection with Confusion Sets

Figure 9. Segmentation tool can help prevent false alarms

| Correct templates | Incorrect templates | Correct templates | Incorrect templates | Correct templates | Incorrect templates |
|-------------------|---------------------|-------------------|---------------------|-------------------|---------------------|
| 衝擊              | 衝急                | 經腳石            | 伴腳石              | 遼不得已          | 遼不得已           |
| 檢視              | 機視                | 大部分            | 大不分              | 情非得已          | 情非得已           |
| 經濟              | 經紀                | 手電筒            | 手電桶              | 遼不得已          | 遼不得已           |
| 循環              | 循還                | 不經意            | 不經易              | 大勢已去          | 大勢以去           |
| 成績              | 成績                | 不願意            | 不願易              | 不能自己          | 不能自以           |
| 薪水              | 新水                | 董事長            | 懂事長              | 迫不得已          | 迫不得以           |
| 賺錢              | 賺錢                | 三輪車            | 三軸車              | 情非得已          | 情非得已           |
| 關鍵              | 關鍵                | 腦震盪            | 腦振盪              | 萬不得已          | 萬不得以           |
| 老闆              | 老版                | 辦公室            | 辦公室              | 遼不得已          | 遼不得已           |
| 雖然              | 隨然                | 成績單            | 成績單              | 巡弋飛彈          | 巡弋飛彈           |

Figure 10. Some short templates generated by our system

3. Experimental Settings, Results and Analysis

3.1 Training Corpus and Student Essays

Our method requires a large corpus to compile templates. Therefore, we used the largest available news corpus as our training set. The corpus is described in Table 1.

| Year       | News sources       | # of Docs | File size |
|------------|--------------------|-----------|-----------|
| 1998-1999  | China Times        | 38,163    |           |
|            | China Times Commercial | 25,812    |           |
|            | China Times Express | 5,747     |           |
|            | Central Daily News  | 27,770    |           |
|            | China Daily News    | 34,728    | 209MB     |
Student essays were collected from one junior high school in Taipei. We used some of the essays for the close test and the rest as the open test, keeping them unseen to the system. The students were 7th or 8th graders. The essays were reviewed by their teachers, and the character errors were highlighted. These 3264 essays were written by hand and were digitized later. See Figure 11 for an example. This is part of our experimental setting that tries to avoid the influence of different input methods. We deleted some symbols and characters that could not be represented by Unicode.

![Figure 11. The file format of our test corpus](image)

Table 2 shows the analysis of the student essays. Most of the characters (94%) in use fell into the frequent characters set. Character errors were not very serious for most of the students, with less than 2 character errors per essay.

Table 3 shows our analysis of the character error types. We find that even in written essays, students tend to write characters having the same pronunciation (66–70%). There is also a high percentage of wrong written characters with the same radical (13–16%).
shows the templates most used for the student essays. These templates are quite common and are too simple for teachers to teach at the 7th and 8th grade levels. A system that can correct these errors may reduce the work of teachers.

**Table 2. Analysis of the student essays**

|                | # of Essays | Average score | Average # of characters | Average # of character errors | % of frequent characters |
|----------------|-------------|---------------|-------------------------|-------------------------------|--------------------------|
| Close test essay | 2241        | 3.62          | 367.12                  | 1.74                          | 94.23%                   |
| Open test essay  | 1023        | 3.61          | 420.02                  | 1.94                          | 94.33%                   |

**Table 3. Analysis of the character error types in student essays**

|                | % with the same radical | % with the same pronunciation | % of both | % out of the two main types |
|----------------|-------------------------|-------------------------------|-----------|----------------------------|
| Close test essay | 13.82%                  | 70.27%                        | 4.92%     | 20.81%                     |
| Open test essay  | 16.96%                  | 66.31%                        | 2.85%     | 19.58%                     |

**Table 4. The most used templates in the test corpus**

|                | Close essay | Correct | Wrong |
|----------------|-------------|---------|-------|
|                | Close essay | 已經 變得 自己 景象 一旦 寄託 已經 畢竟 而已 根本 | 己經 變的 自已 景像 一但 寄托 以經 必竟 而己 依本 |
|                | Open essay  | 己經 一旦 已經 選擇 煩惱 應該 已經 而已 選擇 後悔 | 自己 一但 己經 選擇 煩腦 因該 以經 而己 選擇 後悔 |

**3.2 System Evaluation**

In this study, we compare the quality of characters manually compiled from books and students with that of automatically generated ones. Since the frequencies of 2-character words, 3-character words, and 4-character words are very different, our system uses different thresholds - 2300, 500, and 100 for 2-character words, 3-character words, and 4-character words, respectively, in the experiment.

The precision and recall are defined as follows:

\[
\text{Macro Recall} = \frac{\sum (\frac{dr}{r})}{N} \quad \text{(4)}
\]

\[
\text{Macro Precision} = \frac{\sum (\frac{dr}{sd})}{N} \quad \text{(5)}
\]

\[
\text{Micro Recall} = \frac{\sum (dr)}{\sum (r)} \quad \text{(6)}
\]

\[
\text{Micro Precision} = \frac{\sum (dr)}{\sum (sd)} \quad \text{(7)}
\]

where \(dr\) is the number of correct characters, \(r\) is the number of character errors, \(sd\) is the number of character errors that our system detects, and \(N\) is the number of all of the essays.
Macro Precision and Macro Recall are focused on the performance of correction per essay. This is what real world students might encounter with the system. As Micro Recall and Micro Precision treat the whole data set as one essay, they are suitable for evaluating the average performance of the system. We prefer high precision while maintaining a relatively high recall because we do not want the users to see too many false alarms.

3.3 Experimental Results

We conducted a series of experiments to determine how to improve our system. First, we used confusion sets and the chi-square test to generate templates and compared the performance with the previous work, which did not use confusion sets. Second, we tested whether the square root test is more suitable for our system than the chi-square test. Third, we tested the influence of the segmentation added to our system. We report the best performance of the experimental results by combining the automatically generated templates with the manually edited templates.

3.3.1 The Comparison of Experimental Results of Four Automatic Template Generation Settings

Figure 12 shows the experimental results of using the chi-square test in template generation. Setting A used the automatically generated 19,402 templates in the previous work. Setting B used the confusion sets during the process of automatic template generation. The total number of generated templates was 54,253. The performance of the method proposed in this paper is better than the previous work for both precision and recall. Setting C was the automatically generated templates using the confusion set and the square root test. The total number of templates was 50,467. This new setting results in much higher precision. The Macro Precision value is even better than the manually edited Macro Precision value. This result shows that, when we reduce the automatically generated templates with the square root test, we also reduce noise. For Setting D, our system used confusion sets and a word segmentation tool before the square root test, which generated 9,013 templates. We find that the number of templates is reduced while the performance is improved in terms of both Macro Precision and Micro Precision. The trade off is the performance of recall.
3.3.2 Combining Automatically Generated Templates with Manually Edited Templates

Figure 13 shows the comparison of the performance of our system combining automatically generated templates with manually edited templates. Setting E used the 6,701 manually edited templates. Setting F used the combination of Setting E and Setting C, which had a total of 57,167 templates. Setting G used the combination of Setting E and Setting D, totaling 15,713 templates. The performance of the combinations declines a little bit in terms of both Macro Precision and Micro Precision. Nevertheless, there is an increase in both Macro Recall and Micro Recall. Compared with the results in the previous experiment, the combination helps the overall performance. This means that our system can incorporate more templates and attain better performance in the future.
Figure 13. A comparison of experimental results of combining manually edited with automatically generated templates

Based on the analysis of the confusion sets, our system should have a 70% to 80% recall rate because we compile all of the characters with the same pronunciation and some similar characters in the confusion sets. Nevertheless, the recall remains low, even though we are able to control the high-precision performance. Therefore, we will need to conduct further analysis of our system.

3.4 Analysis of the Mistakes in the Experiment

In this subsection, we discuss the 90,135 templates in Setting I of the third experiment, which were generated by using confusion sets, word segmentation, and the square root test. This setting was designed to maintain high precision and to increase recall.

3.4.1 Regarding the Precision

Theoretically, our system can get 100% precision using templates. In practice, however, there are still many exceptions. In Table 5, we list some false alarms in the open tests. According to an online dictionary (Ministry of Education, 2007), some templates that we compiled are interchangeable, such as: “垃圾桶”-“垃圾筒,” “奇跡” - “奇蹟,” “電線桿” - “電線杆,” and “銷聲匿跡” - “消聲匿跡”. This is not consistent with the judgment of some teachers. Some templates are just too short and cannot include the necessary context in order for a correct decision to be made, such as “一再”.”一在”. The necessary context should include more semantic rather than surface syntax. There were some bad templates that our system should
have not generated, such as “放聲大哭”-“放聲大叫”, “不用說”-“不用講,” and “讀書人”-“讀書做,” which can be attributed to the size of the corpus. Nevertheless, no corpus is large enough to be perfect for all applications. We find that these are the major causes of false alarms.

Table 5. Some templates that caused false alarms

| Correct word | 垃圾桶 | 奇蹟 | 電線桿 | 銷聲匿跡 | 一再 | 放聲大哭 | 不用說 | 讀書人 |
|--------------|-------|-------|--------|---------|------|---------|-------|--------|
| Wrong word   | 垃圾筒 | 奇跡 | 電線杆 | 銷聲匿跡 | 一在 | 放聲大叫 | 不用講 | 讀書做 |

3.4.2 Regarding the Recall

We treated the errors that the teachers provided from the student essays as templates and compared them to the automatically generated templates, as shown in Table 6. The first column shows the percentage of “not in the automatically generated template”. The second column shows the percentage of an error occurring in a word that is not in the dictionary. The third column shows the percentage of an error occurring in a word that is not in the corpus. The last column shows the percentage of an error occurring in a word that is neither in the dictionary nor in the corpus.

We find that most student errors were not mined from the news corpus, although our system has mined many useful error templates. From the union set of those not in a dictionary and not in a corpus, we find that 53.17% of the necessary templates in the close test set cannot be generated by our system, while 32.97% of the necessary templates in the open test cannot be generated by our system. This is a mismatch of the corpus and student essays. The assumption of our system is that the corpus contains the correct and wrong usages. Nevertheless, since news reporters and junior high school students make character errors for different words, we need to have a more suitable corpus to improve our system. If we have a more contemporary dictionary that includes the words in Table 7, our system can perform better.

Table 6. Comparison of real world errors to system generated templates

|                      | Not matching template | Not in dictionary | Not in corpus | Neither in dictionary nor in corpus |
|----------------------|-----------------------|-------------------|---------------|-------------------------------------|
| Close test essay     | 91.53%                | 37.73%            | 35.64%        | 20.20%                              |
| Open test essay      | 93.15%                | 16.27%            | 23.94%        | 7.24%                               |

Table 7. New words not in dictionary

| 傳告欄 | 蒸飯機 | 值日生 | 作業本 | 辦派對 | 睡午覺 | 全班齊心 | 勤加練習 | 警恥心 | 無厘頭 |
|--------|--------|--------|--------|--------|--------|-----------|-----------|--------|--------|
| 重拾信心 | 萬事興 | 經濟人 | 維持 | 公正 | 與善 | 窮有 | 勤勞 | 幸福 | 無養 |
| 應對 | 勞工 | 生活 | 象徵 | 積極 | 融入 | 親密 | 芳香 | 燒炭 | 拉筋 |
4. Conclusion and future works

Based on the confusion sets of Chinese characters, word segmentation, and the square root test, our system can generate a large number of templates from a corpus. These templates can detect and correct Chinese character errors in essays. The templates are more readable and have better performance in both precision and recall performance compared to that of previous system.

To improve the system, we will work in two areas. In the knowledge part, we will enlarge the confusion sets to include more seeds for template generation. We will compile a more suitable corpus for detection and correction of errors in student essays. For the dictionary, we will collect more contemporary terms via the Internet, such as from Wikipedia and Wiktionary. For the language model part, we will use the student essays that we collected in this study to generate an error model, and use that error model to help determine character errors.

Acknowledgement

This study was conducted under the "Project Digital Convergence Service Open Platform" of the Institute for Information Industry, which is subsidized by the Ministry of Economic Affairs of the Republic of China.

Reference

Chen, Y.-Z., Wu, S.-H., Lu, C.-C., & Ku, T. (2009). Automatic Template Generation for Chinese Essay Spelling Error Detecting System. The 13th Global Chinese Conference on Computer in Education, 402-408.

Hung, T.-H., & Wu, S.-H. (2008). Chinese Essay Error Detection and Suggestion System. Taiwan E-Learning Forum.

Liu, C.-L., & Lin, J.-H. (2008). Using structural information for identifying similar Chinese characters. Proceedings of the Forty Sixth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL'08), 93-96.

Liu, C.-L., Tien, K.-W., Lai, M.-H., Chuang, Y.-H., & Wu, S.-H. (2009). Capturing errors in written Chinese words. Proceedings of the Seventh Workshop on Asian Language Resources (ALR7), the Forty Seventh Annual Meeting of the Association for Computational Linguistics (ACL'09), 25-28.

Liu, C.-L., Tien, K.-W., Lai, M.-H., Chuang, Y.-H., & Wu, S.-H. (2009). Phonological and logographic influences on errors in written Chinese words. Proceedings of the Seventh Workshop on Asian Language Resources (ALR7), the Forty Seventh Annual Meeting of the Association for Computational Linguistics (ACL'09), 84-91.

Ministry of Education. (2007). MOE Chinese Dictionary. Taiwan: Ministry of Education.
National Languages Committee. (1996). *Common Errors in Chinese Writings (常用國字辨似)*. Taiwan: Ministry of Education.

National Languages Committee. (1998). *The Investigation on Common Used Word(八十七年常用語詞調查報告書)*. Taiwan: Ministry of Education.

Ravichandran, D., & Hovy, E. (2001). Learning surface text patterns for a Question Answering system. *in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 41-47.

Ren, F., Shi, H., & Zhou, Q. (1994). A hybrid approach to automatic Chinese text checking and error correction. *In Proceedings of the ARPA Workshop on Human Language Technology*, 76-81.

Sung, C.-L., Lee, C.-W., Yen, H.-C., & Hsu, W.-L. (2008). An Alignment-based Surface Pattern for a Question Answering System. *the IEEE International Conference on Information Reuse and Integration*, 172-177.

Xu, S. (2009). *Shuowen Jiezi (說文解字)*. Volumes publishing company(萬卷出版公司).

Zhang, Y.-s. (1999). *Kangxi Dictionary (康熙字典)*. Chung Hwa Book company(中華書局).

Zhang, L., Huang, C., Zhou, M., & Pan, H. (2000). Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 248-254.
