# A Hybrid Approach to Automatic Corpus Generation for Chinese Spelling Check

**Dingmin Wang**[♣,*], **Yan Song**[♠], **Jing Li**[♠], **Jialong Han**[♠], **Haisong Zhang**[♠]
[♣]Tencent Inc.
dimmywang@tencent.com
[♠]Tencent AI Lab
{clksong, ameliajli, jialonghan, hansonzhang}@tencent.com

## Abstract

Chinese spelling check (CSC) is a challenging yet meaningful task, which not only serves as a preprocessing in many natural language processing (NLP) applications, but also facilitates reading and understanding of running texts in peoples' daily lives. However, to utilize data-driven approaches for CSC, there is one major limitation that annotated corpora are not enough in applying algorithms and building models. In this paper, we propose a novel approach of constructing CSC corpus with automatically generated spelling errors, which are either visually or phonologically resembled characters, corresponding to the OCR- and ASR-based methods, respectively. Upon the constructed corpus, different models are trained and evaluated for CSC with respect to three standard test sets. Experimental results demonstrate the effectiveness of the corpus, therefore confirm the validity of our approach.

## 1 Introduction

Spelling check is a crucial task to detect and correct human spelling errors in running texts (Yu and Li, 2014). This task is vital for NLP applications such as search engine (Martins and Silva, 2004; Gao et al., 2010) and automatic essay scoring (Burstein and Chodorow, 1999; Lonsdale and Strong-Krause, 2003), for the reason that spelling errors not only affect reading but also sometimes completely alter the meaning delivered in a text fragment. Especially, in Chinese language processing, spelling errors can be more serious since they may affect fundamental tasks such as word segmentation (Xue, 2003; Song and Xia, 2012) and part-of-speech tagging (Chang et al., 1993; Jiang et al., 2008; Sun, 2011), etc. Of all causes lead to spelling errors, a major one comes from the misuse of Chinese input methods on daily texts, e.g., emails and social media posts. Table 1 illustrates two examples of such Chinese spelling errors. The first incorrect sentence contains a misused character, 己 (ji2)[1], which has a similar shape to its corresponding correct character, i.e., 已 (yi3). In the second incorrect sentence, the boxed spelling error 她 (ta1) is phonetically identical to its corresponding correct one 他 (ta1).

| Sentence | Correction |
|---|---|
| 我们应该认真对待这些 己 (ji2) 经发生的事 | 已 (yi3) |
| 在我们班上， 她 (ta1)是一个很聪明的男孩 | 他 (ta1) |

Table 1: Two examples of Chinese sentences containing spelling errors. Spelling errors are marked in red. The first sentence contains a visually similar spelling error, i.e., 已 (yi3) is misspelled as 己 (ji2). The second sentence contains a phonologically similar spelling error, i.e., 他 (ta1) is misspelled as 她 (ta1).

Owing to the limited number of available datasets, many state-of-the-art supervised models are seldom employed in this field, which hinders the development of CSC. Currently, some mainstream approaches still focus on using unsupervised methods, i.e., language model based ones (Chen et al., 2013; Yu et al., 2014; Tseng et al., 2015; Lee et al., 2016). As a result, the development of CSC techniques are restricted and thus CSC performance is not satisfied so far (Fung et al., 2017). To enhance CSC performance, the biggest challenge is the unavailability of large scale corpora with labeled spelling errors, which is of high value for training and applying supervised models. Such issue of data absence is mainly caused by the fact that annotating spelling errors is an expensive and challenging task.

To address the data unavailability issue so that

[*] This work was conducted during Dingmin Wang's internship in Tencent AI Lab.

[1] We use Chinese pinyin to identify the pronunciation for each character, where the last number represents different tones (ranging from 1 to 4) in Pinyin, same below.



*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2517–2527
Brussels, Belgium, October 31 - November 4, 2018. 

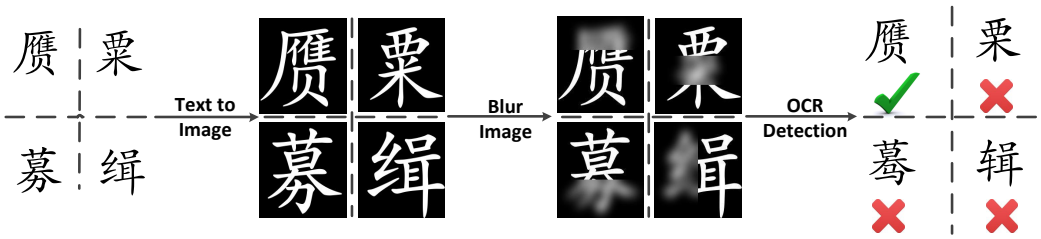


Figure 1: An example process of generating V-style errors by the OCR-based method. In the OCR detection results, except for 赝 (yan4), the other three characters, i.e., 粟 (li4), 慕 (mu4), and 缉 (ji1), are incorrectly recognized as 栗 (su4), 蓦 (mo4), and 辑 (ji2), respectively. All the three incorrect characters have similar shapes with their corresponding correct references.

to facilitate CSC with data-driven approaches, in this paper, we propose a novel approach that automatically constructs Chinese corpora with labeled spelling errors. Specifically, given that Chinese spelling errors mainly result from the misuse of visually and phonologically similar characters (Chang, 1995; Liu et al., 2011; Yu and Li, 2014), we propose OCR-based and ASR-based methods to produce the aforementioned two types of misused characters. Note that, different from detecting spelling errors from incorrect sentences, our proposed approach aims at automatically generating texts with annotated spelling errors like those in Table 1. With the help of OCR- and ASR-based methods, CSC corpora are constructed with annotated visual and phonological spelling errors.

In our experiments, qualitative analysis illustrates that incorrectly recognized Chinese characters by OCR or ASR toolkits are not trivial for human to detect, while interestingly, human are likely to make such spelling errors in their daily writing. In the quantitative comparison, we cast Chinese spelling check into a sequence labeling problem and implement a supervised benchmark model, i.e., bidirectional LSTM (BiLSTM), to evaluate the performance of CSC on three standard testing datasets. Experimental results show that the BiLSTM models trained on our generated corpus yield better performance than their counterparts trained on the training dataset provided in the standard testing datasets. To further facilitating the CSC task, we construct confusion sets by collecting all incorrect variants for each character and their corresponding correct references. The effectiveness of the confusion set is confirmed in the error correction task, indicating that the constructed confusion sets are highly useful in many existing Chinese spelling check schemes (Chang, 1995; Wu et al., 2010; Dong et al., 2016).

## 2 Automatic Data Generation

Spelling errors in Chinese are mainly caused by the misuse of visually or phonologically similar characters (Chang, 1995; Liu et al., 2011; Yu and Li, 2014). Errors of visually similar characters (henceforth V-style errors) are due to the prominence of character pairs visually similar to each other. The reason is that, Chinese, as a hieroglyph language, consists of more than sixty thousand characters[2]. They are constructed by a limited number of radicals and components[3]. As for errors caused by the misuse of phonologically similar characters (henceforth P-style errors), we note that pronunciations of Chinese characters are usually defined by Pinyin, which consists of initials, finals, and tones[4]. According to Yang et al. (2012), there are only 398 syllables for thousands of characters in modern Chinese. As a result, there are many Chinese characters sharing similar pronunciation, which further leads to the prominence of P-style errors. In the rest of this section, we describe how we generate these two types of errors in Section 2.1 and 2.2, respectively.

### 2.1 OCR-based Generation

Inspired by the observation that optical character recognition (OCR) tools are likely to misidentify characters with those visually similar ones (Tong and Evans, 1996), we intentionally blur images with correct characters, and apply OCR tools on them to produce V-style spelling errors.

In detail, we use Google Tesseract (Smith, 2007) as the OCR toolkit and the generation process is illustrated in Figure 1. Given a sentence,

[2] http://www.hanzidian.com.

[3] There are less than three hundred radicals in total. https://en.wikipedia.org/wiki/Radical_(Chinese_characters)

[4] https://en.wikipedia.org/wiki/Pinyin

as the first step, we randomly select $1 \sim 2$ character(s) from it as our target characters to be detected by Tesseract, denoted as $C_{targets}$. Specifically, except for Chinese characters, other characters like punctuations and foreign alphabets are excluded and we also filter those Chinese characters of low frequency[5] based on the statistics of the Chinese Wikipedia Corpus[6]. Second, we transfer $C_{targets}$ from text to image with $100 \times 100$ pixels, namely, each generated image has a same size. Third, we randomly blur a region[7] in the produced images using Gaussian blurring (Bradski, 2000), which aims at leading the OCR toolkit to make mistakes. Finally, we use Google Tesseract to recognize the blurred images. Once the recognized result does not match to the original one, a V-style error is generated, which is used to replace the original character in the sentence, resulting in a sentence with V-style spelling error(s). After the aforementioned steps, we obtain the spelling errors for each sentence with their correct references.

The raw texts used for OCR-based method are mainly from newspaper articles, which are crawled from People's Daily, an official newspaper website[8], of which articles are reported to undergo a strict edition process and assumed to be all correct. We divide these texts into sentences using clause-ending punctuations such as periods (。), question mark (？), and exclamation mark (！) (Chang et al., 1993). In total, we obtain $50,000$ sentences, each of which contains $8$ to $85$ characters, including punctuations. These sentences are then handled by the OCR-based method as we describe before, resulting in an annotated corpus containing about $40,000$ annotated sentences with $56,857$ spelling errors. Note that in our experiment, we find that the OCR toolkit can still correctly detect the characters in the produced images even we blur part of the images. This explains

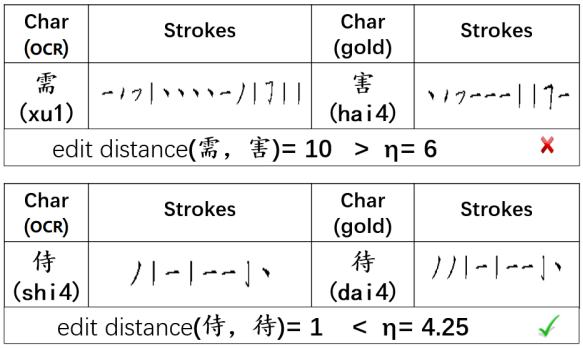


Figure 2: Two cases from the OCR-based method. Char (OCR): the character recognized by the OCR-based method; Char (gold): the corresponding correct character.

why the size of generated annotated sentences is smaller than that of the original sentences. We denote the dataset by D-ocr, and show its statistics in the D-ocr column in Table 3.

**Bad Cases and the Solution** Although the OCR-based method work smoothly, there are still some cases worth further investigation. By analyzing the generated spelling errors by this method, we find that in terms of shape, there exist some incorrectly recognized characters by the OCR toolkit that greatly differ from their corresponding correct characters. For example, for the blurred image containing the character 领 (ling3), the Tesseract incorrectly recognizes it as 铈 (shi4), which is totally different from 领 (ling3) in shape. Therefore, these cases should be excluded because human are less likely to make such mistakes. In solving this problem, we propose a novel approach to judge whether two characters are visually similar by calculating the edit distance based on the strokes of Chinese characters. Similar to English words consisting of alphabets, a Chinese character can be split into different strokes[9]. To this end, we obtain the strokes of Chinese character from the Online dictionary[10]. Empirically, given two Chinese characters, $c_1$ and $c_2$, we set $0.25 * (len(c_1) + len(c_2))$ as the threshold, $\eta$, where $len(c)$ denotes the number of strokes for the Chinese character $c$. If the edit distance of two characters is more than a threshold $\eta$, we consider them not to be similar in shape. To better clarify it, a bad case and a good case are shown in Figure 2.

[5] In our setting, Chinese characters occurring less than five times in the corpus are considered as low-frequency ones.

[6] https://dumps.wikimedia.org/zhwiki/

[7] A Chinese character may be misspelled to other different characters, resulting in different spelling errors. For example, the Chinese character 缉 (ji1) could be misspelled as 揖 (ji1), 辑 (ji2), or 楫 (ji1), such spelling errors could be obtained by blurring different locations of the given character, which makes it possible for the OCR toolkit to give different recognized results for a same character. According to our experiment, if we blur the entire character, we found that the incorrectly recognized results are almost the same. As a result, we could not obtain different spelling errors for a same character. Specially, the blurred location may be different, but the size of the region to be blurred is the same for all images.

[8] http://www.people.com.cn/

[9] https://en.wikipedia.org/wiki/Stroke_(CJKV_character)

[10] https://bihua.51240.com/


Figure 3: The full pipeline of generating P-style errors by ASR-based method. In this example, 仅 (jin3) is incorrectly recognized as 尽 (jin4) marked in red, both of which have a similar pronunciation.

| Case | Gold Sentences | Incorrectly-recognized results | Type |
|---|---|---|---|
| 1 | 而对楼市成交抑制作用最大的限购 | 而对面楼市成交抑制作用最大的限购 | D |
| 2 | 围绕亚运会进行的城市资金投入 | 没让亚运会进行的城市改造的资金投入 | N |
| 3 | 与院方协商赔偿问题 | 与岳风学生赔偿问题 | T |
| 4 | 但是不幸最终还是发生了 | 但是不行最终还是发生了 | C |

Table 2: Four incorrectly-recognized cases with different types of errors. D: Different lengths with that of the original one. N: Not a P-style error. T: Too many errors. C: Correct case to collect.

### 2.2 ASR-based Generation

Similar to OCR tools, automatic speech recognition (ASR) tools may also mistake characters for others with similar pronunciations (Hartley and Reich, 2005). To build an annotated corpus of P-style errors, we follow the similar inspiration with those for V-style errors and OCR tools, and adopted a pipeline as shown in Figure 3. However, given the availability of various speech recognition datasets, we employ a simpler approach. We exploit a publicly available Mandarin speech corpus, AIShell (Bu et al., 2017), which contains around 140,000 sentences with utterances[11]. We use Kaldi (Povey et al., 2011), a speech recognition toolkit, to transcribe the utterances into recognized sentences. Finally, by comparing the recognized sentences with the original ones, we can identify whether the recognition results are correct. If not, they can serve as incorrectly-recognized results and be used to build a corpus with P-style spelling errors.

**Bad Cases and the Solution** For generated P-style errors, we also identify some bad cases, which potentially introduce much noise. To improve the quality of the generated corpus, a solution is thus needed to remove them. Table 2 gives three types of bad cases with a good one. We describe the solution to deal with them as follows.

First, we discard all incorrectly recognized results similar to Case 1, which has different lengths comparing to the corresponding reference sentence. Second, the incorrect characters in Case 2 have totally different pronunciations with their corresponding characters in the gold sentence. Such cases do not satisfy our requirement in generating P-style errors. To this end, we obtain the pronunciation by pinyin[12] of Chinese characters from an online Chinese lexicon[13]. Then it is easy to identify whether the incorrectly-recognized characters have similar or same pronunciation with their corresponding characters in the gold sentence. Specifically, in terms of Pinyin, two characters have similar pronunciation when they have the same initials and finals but different tones, i.e., *da2* and *da1*. Third, according to Chen et al. (2011), there may have two errors per student essay on average, which reflects the fact that that each sentence will not contain more than two spelling errors on average. Therefore, we remove those incorrectly-recognized results that contains more than two incorrect characters as shown in Case 3. After the aforementioned steps, we generate a corpus with more than 7K P-style spelling errors in total. We denote it D-asr and show its statistics in the D-asr column in Table 3.

## 3 Evaluation

### 3.1 Benchmark Data

To evaluate the quality of the generated corpora, we adopt three benchmark datasets from 2013–2015 shared tasks (Wu et al., 2013; Yu et al., 2014; Tseng et al., 2015)[14] on CSC. Table 4 reports the

[11]Collected from 400 people from different dialect areas in China. http://www.openslr.org/resources/33/data_aishell.tgz

[12]Pinyin is the standard method to define the pronunciation of a Chinese character. https://en.wikipedia.org/wiki/Pinyin

[13]http://www.zdic.net

[14]These datasets are written in traditional Chinese, so to keep the consistency with our generated corpus of simplified

| | D-ocr | D-asr | D |
|---|---|---|---|
| Sentences # | 40,000 | 40,000 | 80,000 |
| Characters # | 915,949 | 716,509 | 1,632,458 |
| Errors # | 56,857 | 75,667 | 132,524 |

Table 3: Statistics of our generated corpus. D-ocr denotes the corpus generated by OCR-based method, D-asr denotes the corpus generated by ASR-based method and D is the combination of D-ocr and D-asr.

statistics of the three standard datasets, including training and test parts. Consider that the quality of the training dataset has a significant impact on models' performance on the test datasets. Thus in this paper, we propose a metric, $\mathrm{C}_{\text{train:test}}$, to measure the correlation degree between the training and test dataset by calculating the number of spelling errors that occur in them:

$$C_{train:test} = \frac{\left|E_{train} \cap E_{test}\right|}{\left|E_{test}\right|} \quad (1)$$

where $\mathrm{E}_{\text{train}}$ and $\mathrm{E}_{\text{test}}$ refer to the set of spelling errors in the training dataset and the test dataset, respectively. Each element in the set is a pair containing a correct character and a misspelled character, e.g., (撼 (han4), 憾 (han4)). Table 5 illustrates that, for three different testing datasets, the entire generated corpus D achieves $74.1\%$, $80.6\%$ and $84.2\%$ on $\mathrm{C}_{\text{train:test}}$, respectively, which are much higher than that of $\mathrm{Trn}_{13}$, $\mathrm{Trn}_{14}$ and $\mathrm{Trn}_{15}$. This difference may denote the validity of the generated corpus, with adequate spelling errors.

### 3.2 Qualitative Analysis

**Setup** To evaluate whether the generated corpus contains errors that are easily made by human, we randomly select 300 sentences from it for manual evaluation, with 150 from D-ocr and 150 from D-asr. Three native Chinese speakers, who are college students, are invited to read and annotate errors in these sentences. Then, we analyze the annotated results by three college students from two levels: sentence-level and error-level. On the sentence-level, we consider a sentence to be correctly annotated only if all errors in the sentence are recognized. On the error-level, we calculate the percentage of the number of correctly annotated errors out of the total number of errors.

Chinese, we convert these datasets from tradition Chinese to simplified Chinese using the OpenCC, an Open-source Chinese Converter. https://github.com/BYVoid/OpenCC

| | | Error # | Char # | Sent # |
|---|---|---|---|---|
| 2013[15] | $\mathrm{Trn}_{13}$ | 324 | 17,611 | 350 |
| | $\mathrm{Tst}_{13}$ | 966 | 75,328 | 1,000 |
| 2014[16] | $\mathrm{Trn}_{14}$ | 5,224 | 330,656 | 6,527 |
| | $\mathrm{Tst}_{14}$ | 510 | 54,176 | 1,063 |
| 2015[17] | $\mathrm{Trn}_{15}$ | 3,101 | 95,114 | 3,174 |
| | $\mathrm{Tst}_{15}$ | 531 | 34811 | 1,100 |

Table 4: Statistics of three standard datasets. Error # denotes the number of spelling errors, Char # represents the number of Chinese character and Sent # refers to the number of sentences.

| Train:Test | $C$ (%) | Train:Test | $C$ (%) |
|---|---|---|---|
| $\mathrm{Trn}_{13}$ : $\mathrm{Tst}_{13}$ | 16.2 | D : $\mathrm{Tst}_{13}$ | 74.1 |
| $\mathrm{Trn}_{14}$ : $\mathrm{Tst}_{14}$ | 53.9 | D : $\mathrm{Tst}_{14}$ | 80.6 |
| $\mathrm{Trn}_{15}$ : $\mathrm{Tst}_{15}$ | 46.7 | D : $\mathrm{Tst}_{15}$ | 84.2 |

Table 5: Correlation results of $\mathrm{Trn}_{13}$, $\mathrm{Trn}_{14}$, $\mathrm{Trn}_{15}$ and D with $\mathrm{Tst}_{13}$, $\mathrm{Tst}_{14}$, $\mathrm{Tst}_{15}$.

**Results** Table 6 shows the information of 300 sentences and the annotation results on them. The average recall in the table illustrates that three students have a higher recognition rate for errors from D-asr than that from D-ocr, which, to some extent, indicates that P-style errors are easier to be detected than V-style ones. Besides, we observe that three volunteers fail to identify around 36.9% errors on average, which may indicate that our generated sentences include some challenging errors which are likely to be made by human. Such errors are valuable for CSC since they are potential real cases in people's writing or typing.

**Case Study** To qualitatively analyze why some spelling errors are not detected by human, we conduct a case study on an example sentence containing some spelling errors that are not found by the three students. The sentence is 政企部分是一种痼疾 (translation: politics and industry parts are a kind of a chronic disease), in which the third character 部 (bu4) is a spelling error and should be corrected as 不 (bu4). In this example, 部分 (translation: a part of) and 不分 (translation: equally treat) are two highly common Chinese words and

[15] http://ir.itc.ntnu.edu.tw/lre/sighan7csc.html

[16] http://ir.itc.ntnu.edu.tw/lre/clp14csc.html

[17] http://ir.itc.ntnu.edu.tw/lre/sighan8csc.html

| | Stu1 | Stu2 | Stu3 | R |
|---|---|---|---|---|
| S-ocr | **84**/150 | **100**/150 | **75**/150 | 57.3 |
| E-ocr | **104**/170 | **121**/170 | **100**/170 | 72.0 |
| S-asr | **95**/150 | **79**/150 | **106**/150 | 62.0 |
| E-asr | **341**/393 | **179**/393 | **356**/393 | 74.3 |

Table 6: Human evaluation results. S-ocr refer to 150 sentences from the D-ocr and E-ocr represents the errors in S-ocr, similar for S-asr and E-asr. R denotes the average recall of three students. Numbers in bold denotes the correctly-annotated results by students.

easy to be considered as correct. However, considering the preceding word 政企 (translation: politics and industry) and the subsequent words 是一种痼疾 (translation: a kind of chronic disease), we can see that 部分 does not fit the current context and should be corrected as 不分. This case study confirms that our generated corpus contains some spelling errors like human made ones in their writing or typing, which further demonstrates its quality and the effectiveness of our approach.

### 3.3 Quantitative Comparison

#### 3.3.1 Chinese Spelling Error Detection

In this section, we evaluate the quality of our generated corpus through Chinese spelling detection. We firstly explore how different proportions of P-style and V-style errors affect the quality of the corpus. Then we compare the detection performance of the generated corpus with that of training datasets provided in the three shared tasks.

**Setup** We cast Chinese spelling error detection into a sequence tagging problem on characters, in which the correct and incorrect characters are tagged as 1 and 0, respectively[18]. We then implement a supervised sequence tagging model, i.e., bidirectional LSTM (BiLSTM) (Graves and Schmidhuber, 2005), as our baseline to evaluate the quality of different corpus. The hidden size of the BiLSTM is set to 150 and the other hyper-parameters are tuned on a development set consisting of 10% randomly selected sentences from the training data. We minimize categorical cross-entropy loss for the model, with RMSprop (Graves, 2013) as the optimizer.

[18]The detection process can be considered as a problem of binary classification.

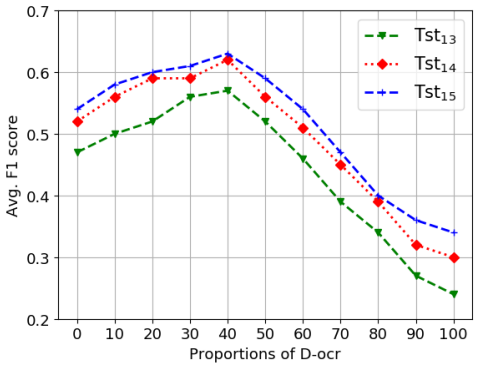


Figure 4: F1 scores achieved by different proportions of D-ocr and D-asr on three testing dataset ($\text{Tst}_{13}$,$\text{Tst}_{14}$ and $\text{Tst}_{15}$). The total size of the corpus is 40k, the percentage of which the x-axis represents from D-ocr.

**Results** The performance of BiLSTM trained on different proportions of D-ocr and D-asr aims at exploring the quality of the generated corpus influenced by the distribution of P-style and V-style spelling errors. Figure 4 shows that when the size of training dataset is fixed (=40k), different proportions of P-style and V-style errors achieve different F1 scores, denoting that the proportion of P-style and V-style spelling errors affects the quality of the generated corpus. Specifically, it is observed that a 4:6 proportion of D-ocr and D-asr achieves the best performance on three testing datasets when compared with other proportion ratios. In addition, for the two special proportions (0% and 100%), it is seen that with the same size of corpus, the BiLSTM model trained on D-asr achieves better performance than that on D-ocr, which indicates that P-style spelling error contributes more to the quality of the corpus. This experimental result complies with the previous conclusion (Liu et al., 2009, 2011) that most of spelling errors are related to the pronunciations.

Furthermore, to better illustrate the quality of the generated corpus, we compare it with some training datasets, which are manually annotated (Wu et al., 2013; Yu et al., 2014; Tseng et al., 2015). According to previous analyses on the experimental results shown in Figure 4, we choose the 4:6 proportion of P-style and V-style spelling errors to construct the training data in the following experiments. Specifically, we build five different sizes of datasets: D-10k[19], D-20k, D-30k, D-40k and D-50k, which are extracted from

[19]D-10k denotes the corpus contains 10k sentences, similar for D-20k, D-30k, D-40k and D-50k.

| | $\text{Tst}_{13}$ | | | $\text{Tst}_{14}$ | | | $\text{Tst}_{15}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| Trn | 24.4 | 27.3 | 25.8 | 49.8 | 51.5 | 50.6 | 40.1 | 43.2 | 41.6 |
| D-10k | 33.3 | 39.6 | 36.1 | 31.1 | 35.1 | 32.9 | 31.0 | 37.0 | 33.7 |
| D-20k | 41.1 | 50.2 | 45.2 | 41.1 | 50.2 | 45.2 | 43.0 | 54.9 | 48.2 |
| D-30k | 47.2 | 59.1 | 52.5 | 40.9 | 48.0 | 44.2 | 50.3 | 62.3 | 55.7 |
| D-40k | 53.4 | 65.0 | 58.6 | **52.3** | 64.3 | 57.7 | **56.6** | 66.5 | 61.2 |
| D-50k | **54.0** | **69.3** | **60.7** | 51.9 | **66.2** | **58.2** | **56.6** | **69.4** | **62.3** |

Table 7: The performance of Chinese spelling error detection with BiLSTM on $\text{Tst}_{13}$,$\text{Tst}_{14}$,$\text{Tst}_{15}$ (%). Best results are in **bold**. Trn represents the training dataset provided in the corresponding shared task, e.g., Trn denotes $\text{Trn}_{13}$ in $\text{Tst}_{13}$.

D-ocr and D-asr following the proportion of 4:6. Then, we train the BiLSTM model on these training datasets and evaluate error detection performance on $\text{Tst}_{13}$, $\text{Tst}_{14}$ and $\text{Tst}_{15}$. Table 7 shows the detection performance on three different testing datasets. We have the following observations.

***The size of training dataset is important for the model training.*** For $\text{Tst}_{13}$, D-10k achieves a better F1 score than $\text{Trn}_{13}$. A major reason may be the size of $\text{Trn}_{13}$ (=350, see in Table 3), which is much smaller than the testing dataset. In this situation, the model can not learn enough information, resulting in being unable to detect unseen spelling errors. Besides, we can see that the detection performance shows a stable improvement as the size of our generated corpus is continuously enlarged. Therefore, for data-driven approaches, it is of great importance to train our model with enough instances having different spelling errors.

***The precision may be compromised if the training dataset contains too many "noisy" spelling errors*** From Table 7, although the overall performance (F1 score) keeps improving as the size of our generated corpus increases, the precision and the recall demonstrate different changing trends. It is observed that as the size of training dataset increases, the model achieves a better performance in terms of the recall. An possible reason is that with more instances containing different spelling error including in the training dataset, the number of unseen spelling error in the testing dataset is reduced, thus facilitating the model to detect more spelling errors. However, the improvement of the precision is not so obvious as that of the recall. Specifically, in $\text{Tst}_{14}$ and $\text{Tst}_{15}$, D-50k does not achieve a higher precision than D-40k. A possible explanation is that with a larger training dataset containing more spelling error instances, it may lead the model to misidentify some more correct characters, resulting in a lower precision.

***Compared with the limited training dataset manually annotated by human, our generated large-scale corpus can achieves a better performance*** From Table 7, we can see that with a certain size of our generated corpus, it can train a model that achieve better detection performance than with the manually annotated datasets provided in the corresponding shared tasks. To some extent, this demonstrates the effectiveness of our generated corpus, thus confirms the validity of our approach.

### 3.3.2 Chinese Spelling Error Correction

Once the Chinese spelling errors are detected, we test on correcting them.[20] from Section 3.3.1.

Following previous studies (Chang, 1995; Huang et al., 2007; Wu et al., 2010; Chen et al., 2013; Dong et al., 2016), we adopt confusion sets[21] and a language model to handle the taks for Chinese spelling error correction. In particular, by collecting all incorrect variants for each correct character, we construct a confusion set for all involved correct characters, denoted as Ours. In addition, to illustrate its effectiveness, we compare it with two publicly available confusion sets (Liu et al., 2009), Con1 and Con2. Specifically, Con1 consists of SC (similar Cangjie), SSST (same-sound-same-tone) and SSDT (same-sound-different-tone), while Con2 consists of all sets, SC, SSST, SSDT, MSST (similar-sound-

[20] In particular, we choose the best detected results achieved by D-50k as shown in Table 7.

[21] A confusion set refers to a set that contains confusion characters for a given character. For example, a Chinese character 快 (kuai4) may have a confusion set, {筷 (kuai4),块 (kuai4),快 (yang1),· · · }, each of which has a similar pronunciation or a similar shape with 快 (kuai4).

| Metrics | $\text{Tst}_{13}$ | | | $\text{Tst}_{14}$ | | | $\text{Tst}_{15}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Con1 | Con2 | Ours | Con1 | Con2 | Ours | Con1 | Con2 | Ours |
| F1 (%) | 47.6 | **52.1** | 50.3 | 52.6 | **56.1** | 53.0 | 55.6 | **57.1** | 56.3 |
| Time(s) | 290.3 | 679.9 | **101.2** | 310.2 | 792.1 | **132.4** | 267.6 | 622.5 | **119.2** |

Table 8: Error correction results on $\text{Tst}_{13}$, $\text{Tst}_{14}$, $\text{Tst}_{15}$, using Con1, Con2 and Ours, respectively.

| Name | # of Char | Min. | Max. | Avg. |
|---|---|---|---|---|
| Ours | 4,676 | 1 | 53 | 5.6 |
| Con1 | 5,207 | 8 | 196 | 50.6 |
| Con2 | 5,207 | 21 | 323 | 86.0 |

Table 9: Statistics of different confusions sets.

same-tone) and MSDT (similar-sound-different-tone). Table 9 shows the statistics information of the three confusion sets.

**Setup** Similar to Dong et al. (2016), we adopt a tri-gram language model to calculate the probability of a given sentence. Based on the detected results by the sequence labeling models, we choose as the error correction the character from the corresponding confusion set that has the highest probability. For a given sentence containing $n$ words, $W = w_1, w_2, \cdots, w_n$, where $w_i$ represents the $i^{th}$ character in the sentence, $E$ is a set containing the indexes of detected incorrect characters, $W(w_i, c)$ denotes the new generated sentence after the $i^{th}$ character is replaced with c. The process of error correction can be formulated as follows:

$$\forall i \in E: \quad \underset{c \in C(w_i)}{\arg\max} \; P(W(w_i, c)) \tag{2}$$

where $P(W)$ is the probability of a sentence $W$ approximated by the product of a series of conditional probabilities as described in (Jelinek, 1997) and $C(w_i)$ refers to the confusion set of $w_i$, one of which with the maximum conditional probability is selected as the correction character.[22]

**Results** Table 9 shows the running time[23] and the F1 scores achieved by different confusion sets based on a tri-gram language model. We can observe that our constructed confusion sets achieve a better correction performance than that of Con1 while a little lower than Con2. However, from Table 9, we can see that Con2 has a much larger size of confusion characters than Ours; and for the same testing, Con2 needs much more time to finish the task while Ours always uses the least time. These observations indicate that our constructed confusions sets are more efficient in containing fewer redundant confusion characters that seldom serve as correction characters.

**Error Analysis** We conduct an error analysis on two types of incorrect cases, namely, the false positive and the false negative case, which affect the precision and recall in CSC, respectively.

For the false positive case, we find that one common issue is that for some fixed usages, such as idioms, phrases, and poems, our model often gives incorrect results. For example, in 风雨送春归 (translation: wind and rain escort Spring's departure), a line of a Chinese poem, 送 is incorrectly recognized as an irrelevant character. By checking the annotated corpus generated by the proposed methods, we observe that in most cases, 迎春 is a more common match, and 送 is annotated as an spelling error when it co-occurs with 春 in the generated corpus. To improve the precision, a possible approach to handle such cases is utilizing some external knowledge, such as building a collection of special Chinese usages.

For the false negative case, taking 想想健康，你就会知道应该要禁烟了 (translation: you will realize that you should give up smoking when your consider of your health) as an example, in which 禁 should be corrected as 戒. However, since 戒 and 禁 are neither visually nor phonologically similar, the proposed corpus generation approach is unable to construct such spelling errors, so it is understandable that the trained model can not detect such spelling errors. In addition, on the word-level, 禁烟 (translation: forbid smoking) and 戒烟 (translation: give up smoking) are two related common Chinese words; it needs to incorporate more context in order to improve the recall performance. Similar to our study on character-based corpus generation, one potential solution is to construct a word-level annotated corpus in order to better detect such spelling errors.

[22] If more than one character from the confusion set have the same maximum probability, we randomly select one of them as the correction character.

[23] We run the experiments on a computer with Intel Core i5-7500 CPU.

## 4 Related Work

In the line of research on spelling error detection and correction, most previous efforts focus on designing different models to improve the performance of CSC (Chang, 1995; Huang et al., 2007, 2008; Chang et al., 2013). Different from them, this work contributes to the generation of training datasets, which are important resources and can be used for improving many existing CSC models. Currently, the limited training datasets have set a high barrier for many data-driven approaches (Wang et al., 2013; Wang and Liao, 2015; Zheng et al., 2016). To the best of our knowledge, up to date, there is no large quantities of annotated data sets commonly available for CSC.

Some previous work (Liu et al., 2009; Chang et al., 2013) pointed out that visually and phonologically similar characters are major contributing factors for errors in Chinese texts, where the number of phonologically similar spelling errors is about two times than that of visually similar spelling errors. Vision- and speech-related technologies are then adopted in our approach. As a technology to extract text information from images, optical character recognition recognizes the shapes and assigns characters. According to Nagy (1988); McBride-Chang et al. (2003), incorrectly recognized results are mainly due to the visual similarities among some different Chinese characters. On the other side, automatic speech recognition is an acoustics-based recognition process for handling audio stream, where phonologically similar characters are usually confused (Kaki et al., 1998; Voll et al., 2008; Braho et al., 2014).

## 5 Conclusion and Future Work

In this paper, we proposed a hybrid approach to automatic generating Chinese corpus for spelling check with labeled spelling errors. Specifically, OCR- and ASR-based methods were used to generate labeled spelling errors by replacing visually and phonologically resembled characters. Human evaluation confirmed that our proposed method can produce common errors that are likely to be made by human and such errors can serve as effective annotated spelling errors for CSC. In our experiment, a neural tagging model was trained on the generated corpus and the results tested on benchmark datasets confirmed the quality of the corpus, which further demonstrated the effectiveness of our corpus generation approach for CSC.

The large scale annotated dataset generated by the proposed approach can potentially serve as useful resources in helping improving the performance of data-driven models for CSC, because the availability of large scale annotated data is the first vital step before applying any algorithms or models. In this paper, we do not put too many efforts into model design for CSC, which we leave as potential future work. To facilitate related research in the community and benefit other researchers, we make our code and data in this work publicly available on: https://github.com/wdimmy/Automatic-Corpus-Generation.

## Acknowledgements

The authors want to express special thanks to Xixin Wu for his suggestions and help in the experiment of ASR. Besides, the authors would like to thank Li Zhong, Shuming Shi, Garbriel Fung, Kam-Fai Wong, and three anonymous reviewers for their help and insightful comments on various aspects of this work.

## References

Gary Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123.

Keith P Braho, Jeffrey P Pike, and Lori A Pike. 2014. Methods And Systems for Identifying Errors in A Speech Recognition System. US Patent 8,868,421.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. AIShell-1: An Open-source Mandarin Speech Corpus and A Speech Recognition Baseline. *arXiv preprint arXiv:1709.05522*.

Jill Burstein and Martin Chodorow. 1999. Automated Essay Scoring for Nonnative English Speakers. In *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing*, pages 68–75. Association for Computational Linguistics.

Chao-Huang Chang. 1995. A New Approach for Automatic Chinese Spelling Correction. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, volume 95, pages 278–283.

Chao-Huang Chang et al. 1993. HMM-based Part-of-speech Tagging for Chinese Corpora. *Very Large Corpora: Academic and Industrial Perspective*.

Tao-Hsing Chang, Hsueh-Chih Chen, Yuen-Hsien Tseng, and Jian-Liang Zheng. 2013. Automatic Detection and Correction for Chinese Misspelled Words Using Phonological and Orthographic Similarities. In *Proceedings of the Seventh SIGHAN*

*Workshop on Chinese Language Processing*, pages 97–101.

Kuan-Yu Chen, Hung-Shin Lee, Chung-Han Lee, Hsin-Min Wang, and Hsin-Hsi Chen. 2013. A study of Language Modeling for Chinese Spelling Check. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing (SIGHAN-13)*, pages 79–83.

Yong-Zhi Chen, Shih-Hung Wu, Ping-Che Yang, Tsun Ku, and Gwo-Dong Chen. 2011. Improve the Detection of Improperly Used Chinese Characters in Students' Essays with Error Model. *International Journal of Continuing Engineering Education and Life Long Learning*, 21(1):103–116.

Shichao Dong, Gabriel Pui Cheong Fung, Binyang Li, Baolin Peng, Ming Liao, Jia Zhu, and Kam-Fai Wong. 2016. ACE: Automatic Colloquialism, Typographical and Orthographic Errors Detection for Chinese Language. In *COLING (Demos)*, pages 194–197.

Gabriel Fung, Maxime Debosschere, Dingmin Wang, Bo Li, Jia Zhu, and Kam-Fai Wong. 2017. NLPTEA 2017 Shared Task–Chinese Spelling Check. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 29–34.

Jianfeng Gao, Xiaolong Li, Daniel Micol, Chris Quirk, and Xu Sun. 2010. A Large Scale Ranker-based System for Search Query Spelling Correction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 358–366. Association for Computational Linguistics.

Alex Graves. 2013. Generating Sequences With Recurrent Neural Networks. *arXiv pre-print*, abs/1308.0850.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Networks*, 18(5):602–610.

Matthew W Hartley and David E Reich. 2005. Method and System for Speech Recognition Using Phonetically Similar Word Alternatives. US Patent 6,910,012.

Chuen-Min Huang, Mei-Chen Wu, and Ching-Che Chang. 2007. Error Detection and Correction Based on Chinese Phonemic Alphabet in Chinese Text. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 463–476. Springer.

Chuen-Min Huang, Mei-Chen Wu, and Ching-Che Chang. 2008. Error Detection and Correction Based on Chinese Phonemic Alphabet in Chinese Text. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 16(supp01):89–105.

Frederick Jelinek. 1997. *Statistical Methods for Speech Recognition*. MIT press.

Wenbin Jiang, Liang Huang, Qun Liu, and Yajuan Lü. 2008. A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-speech Tagging. *Proceedings of ACL-08: HLT*, pages 897–904.

Satoshi Kaki, Eiichiro Sumita, and Hitoshi Iida. 1998. A Method for Correcting Errors in Speech Recognition Using the Statistical Features of Character Co-occurrence. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 653–657. Association for Computational Linguistics.

Lung-Hao Lee, RAO Gaoqi, Liang-Chih Yu, XUN Endong, Baolin Zhang, and Li-Ping Chang. 2016. Overview of nlp-tea 2016 Shared task for Chinese Grammatical Error Diagnosis. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 40–48.

C-L Liu, M-H Lai, K-W Tien, Y-H Chuang, S-H Wu, and C-Y Lee. 2011. Visually and Phonologically Similar Characters in Incorrect Chinese Words: Analyses, Identification, and Applications. *ACM Transactions on Asian Language Information Processing (TALIP)*, 10(2):10.

Chao-Lin Liu, Kan-Wen Tien, Min-Hua Lai, Yi-Hsuan Chuang, and Shih-Hung Wu. 2009. Phonological and Logographic Influences on Errors in Written Chinese Words. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 84–91. Association for Computational Linguistics.

Deryle Lonsdale and Diane Strong-Krause. 2003. Automated Rating of ESL Essays. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2*, pages 61–67. Association for Computational Linguistics.

Bruno Martins and Mário J Silva. 2004. Spelling correction for search engine queries. In *Advances in Natural Language Processing*, pages 372–383. Springer.

Catherine McBride-Chang, Hua Shu, Aibao Zhou, Chun Pong Wat, and Richard K Wagner. 2003. Morphological Awareness Uniquely Predicts Young Children's Chinese Character Recognition. *Journal of educational psychology*, 95(4):743.

G Nagy. 1988. Chinese Character Recognition: A Twenty-five-year Retrospective. In *Pattern Recognition, 1988., 9th International Conference on*, pages 163–167. IEEE.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, EPFL-CONF-192584. IEEE Signal Processing Society.

Ray Smith. 2007. An Overview of the Tesseract OCR Engine. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 629–633. IEEE.

Yan Song and Fei Xia. 2012. Using a Goodness Measurement for Domain Adaptation: A Case Study on Chinese Word Segmentation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3853–3860, Istanbul, Turkey.

Weiwei Sun. 2011. A Stacked Sub-Word Model for Joint Chinese Word Segmentation and Part-of-speech Tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1385–1394. Association for Computational Linguistics.

Xiang Tong and David A Evans. 1996. A Statistical Approach to Automatic OCR Error Correction in Context. In *Fourth Workshop on Very Large Corpora*.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to Sighan 2015 Bake-off for Chinese Spelling Check. *ACL-IJCNLP 2015*, page 32.

Kimberly Voll, Stella Atkins, and Bruce Forster. 2008. Improving the Utility of Speech Recognition Through Error Detection. *Journal of digital imaging*, 21(4):371.

Yih-Ru Wang and Yuan-Fu Liao. 2015. Word vector/conditional random field-based chinese spelling error detection for sighan-2015 evaluation. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 46–49.

Yih-Ru Wang, Yuan-Fu Liao, Yeh-Kuang Wu, and Liang-Chun Chang. 2013. Conditional Random Field-based Parser and Language Model for Traditional Chinese Spelling Checker. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 69–73.

Shih-Hung Wu, Yong-Zhi Chen, Ping-Che Yang, Tsun Ku, and Chao-Lin Liu. 2010. Reducing the False Alarm Rate of Chinese Character Error Detection and Correction. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*, 8(1):29–48.

Shaohua Yang, Hai Zhao, Xiaolin Wang, and Bao-liang Lu. 2012. Spell Checking for Chinese. In *LREC*, pages 730–736.

Junjie Yu and Zhenghua Li. 2014. Chinese Spelling Error Detection and Correction Based on Language Model, Pronunciation, and shape. *CLP 2014*, page 220.

Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, Hsin-Hsi Chen, et al. 2014. Overview of SIGHAN 2014 Bake-off for Chinese Spelling Check. In *Proceedings of the 3rd CIPSSIGHAN Joint Conference on Chinese Language Processing (CLP'14)*, pages 126–132.

Bo Zheng, Wanxiang Che, Jiang Guo, and Ting Liu. 2016. Chinese Grammatical Error Diagnosis with Long Short-Term Memory Networks. *NLPTEA 2016*, page 49.