Automatically Build Corpora for Chinese Spelling Check Based on the Input Method

Jianyong Duan[1,2], Lijian Pan[1,2], Hao Wang[1,2], Mei Zhang[1], and Mingli Wu[1]

[1] North China University of Technology
[2] CNONIX National Standard Application and Promotion Lab
duanjy@ncut.edu.cn, {panlijian1215,wanghaomails}@gmail.com

Abstract. Chinese Spelling Check (CSC) is very important for Chinese language processing. To utilize supervised learning for CSC, one of the main challenges is that high-quality annotated corpora are not enough in building models. This paper proposes new approaches to automatically build the corpora of CSC based on the input method. We build two corpora: one is used to check the errors in the texts generated by the Pinyin input method, called p-corpus, and the other is used to check the errors in the texts generated by the voice input method, called v-corpus. The p-corpus is constructed using two methods, one is based on the conversion between Chinese characters and the sounds of the characters, and the other is based on Automatic Speech Recognition (ASR). The v-corpus is constructed based on ASR. We use the misspelled sentences in real language situation as the test set. Experimental results demonstrate that our corpora can get a better checking effect than the benchmark corpus.

Keywords: corpora · Chinese spelling check · input method.

1 Introduction

All of the reasons for the spelling errors, a major one comes from the misuse of Chinese input methods on daily texts [12]. At present, the most popular Chinese input method is the Pinyin[3] input method [5], at the same time, the voice input method is getting more increasingly widely, such as machine translation, intelligent question and answer, voice navigation, data entry, etc. They are two mainstream of Chinese input methods. Table 1 shows two misspelled sentences generated by the two input methods.

Chinese misspelled sentences	Correction	Methods
火势逐渐向四周漫(man4)延 The fire gradually spreads around	蔓(man4)	P-method
任务是商(shang1)场(chang3)与(yu3)辽(liao2)库 The task is to generate corpus	生(sheng1)成(cheng2)语(yu3)料(liao4)	V-method

Table 1. Two misspelled sentences. Characters with red marks are misspelled characters. Correction denotes the correct character. Source Error denotes the input method which generates the spelling errors. P-method and V-method denote the Pinyin input method and the voice input method, respectively.

[3] Pinyin is the annotation of Chinese pronunciation. https://en.wikipedia.org/ wiki/Pinyin

To use supervised learning for CSC, we need a large number of annotated sentences like the sentences in Table 1. However, there is one major limitation that annotated corpora are not enough. Thus, this paper proposes approaches for automatically building the p-corpus and the v-corpus. The two corpora contain the misspelled sentences whose forms are consistent with that generated by the Pinyin input method and the voice input method, respectively.

The pronunciation of Chinese characters consists of two parts: sound and tone [7]. Such as "漫" (man4 "vine"), the sound is "song" and the tone is "4[4]". Since it is unaffected by the tone when using the Pinyin input method, there are two main types of spelling errors: the *m*isuse of *s*ame *s*ound characters (M-SS) and the *m*isuse of si*m*ilar *s*ound characters (M-MS)[5][7]. Hence, the p-corpus contains two types of sentences: M-SS type sentences with M-SS type errors and M-MS type sentences with M-MS type errors. The former are generated based on the conversion between Chinese characters and the sounds of the characters, and the later are generated based on the ASR.

At present, people mainly focus on improving the accuracy of speech recognition [10, 1]. As far as we know, few people have done spelling check from the results of the recognition. Hence the existing spelling check systems often cannot check the misspelled texts by the voice input method. Take Google spelling check system as an example, as shown in Figure 1, the first misspelled sentence is generated by the voice input method, and the system can't check it out. The second misspelled sentence is generated by the Pinyin input method, and the system checks it out. So if we want to use supervised learning to check the texts generated by the voice input method, we need to build the v-corpus. We collect the misspelled sentences generated by ASR tools to construct the v-corpus.

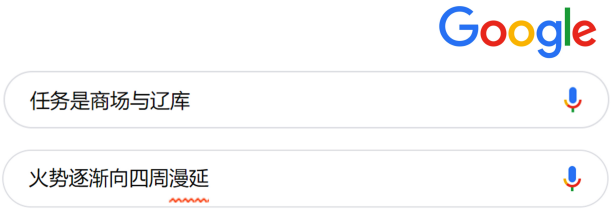


Fig. 1. The check results of the Google spelling check system. Word with red wavy lines denotes the misspelled word detected by the system.

Qualitative assessment of the corpora by measuring the similarity between the misspelled sentences in the corpora and those in real language situation. The evaluation results demonstrate that the two types of misspelled sentences are very similar, that is to say, people will make such spelling errors. In the quantitative evaluation, we treat CSC as a sequence tagging problem on characters, in which the correct or misspelled characters are tagged as C or M, respectively. A supervised model (BiLSTM-CRF) is

[4] Chinese tones range from 1 to 4

[5] According to [6], sound edit distance 1 covers about 90% of spelling errors, and sound edit distance 2 accounts for almost all of the remaining spelling errors. Thus we consider two characters with sound edit distances 1 or 2 as similar characters. Such as "震" (zhen4 "shock") and "正" (zheng4 "positive"), their sound edit distance is 1; hence, they are similar characters.

trained for spelling check [8]. The evaluation results demonstrate that our corpora are better than the benchmark corpus.

The rest of this paper is organized as follows. In Section 2, we briefly introduce how previous researchers obtained annotated corpora. Section 3 details the approaches of automatically building the corpora. A series of experiments are presented in Section 4. Finally, conclusions and future work are given in Section 5.

2 Related Work

Annotating spelling errors is an expensive and challenging task [12]. Most of the previous researchers used methods of collecting the misspelled sentences in real language situation to construct corpora [13, 17, 6]. The data in [13] is collected from the handwritten composition of primary school students. The data in [17] is collected from online papers is not handwritten. The data in [12] is collected from the Chinese misspelled sentences generated by ASR tool and OCR tool. In addition, most of them want to use the corpora generated through one or several input methods to check the texts generated by all input methods [12]. Nevertheless, different input methods produce different forms of spelling errors [14, 16], it is difficult to generate all types of errors by using one or several input methods. The following illustrates the difference of the errors generated by different input methods.

(1) When using the Pinyin input method, there is the misuse of confusing characters, such as the "漫" (man4 "overflow") and "蔓" (man4 "vine") in Table 1. People often can't distinguish them correctly, which leads to spelling errors. Nevertheless, when using the voice input method, such errors will hardly occur.

(2) There is no tone information when using the Pinyin input method [18, 14]. However, when using the voice input method, there is tone information. For example, when using the voice input method to input "抱负" (bao4fu4 "ambition"), the word with the same sound and same tone as "抱负" (bao4fu4) may be output, such as "暴富" (bao4fu4 "rich"). Nevertheless, when using the Pinyin input method, the word with the same sound but the different tone from "抱负" (bao4fu4) may be output, such as "包袱" (bao1fu2 "burden").

(3) In each sentence, the number of errors generated by different methods is various. According to [4], there may be two errors per student essay on average, which reflects the fact that when using the Pinyin input method, each sentence will not contain more than two spelling errors on average. However, according to statistics, nearly one-quarter of the Chinese misspelled sentences produced by the voice input method contain over two errors.

Therefore, this paper proposes new methods for automatically constructing the p-corpus and the v-corpus for the two major input methods.

3 Building the Corpora

This section will introduce three parts. In Section 3.1, we introduce the reasons for the spelling errors. Section 3.2 and Section 3.2 detail the approaches of automatically constructing the p-corpus and v-coupus, respectively.

3.1 Reasons for the Spelling Errors

How the errors occur when using the Pinyin input method. Using the Pinyin input method will bring two main types of spelling errors: M-SS and M-MS type sentences. The total number of Chinese characters exceeds 85,000, yet these characters are only pronounced in 420 different ways [7], which leads to the fact that many Chinese characters share a single pronunciation [14]; thus, M-SS type sentences often appear [7]. There are two major reasons for the generation of M-MS type sentences. Firstly, when using the Pinyin input method, insertion, deletion, replacement, and transposition may occur, which will lead to the generation of the M-MS type sentences [18, 5]. Secondly, people living in different regions may have different pronunciation systems [7], and some people cannot distinguish the fuzzy sounds, such as "eng" and "en", "s" and "sh", etc. At the same time, most Pinyin input methods support fuzzy sound input[6]. After enabling fuzzy sounds, such as "sh–s", input "si" can also come out "十" (shi2 "ten"), and input "shi" can also come out "四" (si4 "four"), which brings great help to people with different pronunciation systems. It is obvious that the fuzzy sound input is one of the reasons for the generation of M-MS type sentences [18, 7]. **How the errors occur when using the voice input method.** When using the voice input method, there are two main factors leading to spelling errors. One is the input pronunciation is not standard, and the other is speech recognition accuracy is not high enough [1].

3.2 Building the p-corpus

This section will introduce four parts. First, we will introduce the raw data of building the p-corpus. The second is the setting of the number of the errors in each sentence. The third and the last will introduce the methods of generating M-SS and M-MS type sentences, respectively.

The raw data used for generating M-SS type sentences is some authoritative news corpora, including Agence France Presse, People's Daily, etc[7]. The raw data used for generating M-MS type sentences is from the publicly spoken Mandarin speech library AlShell[8][2], which contains correct texts information and corresponding audio information. We discard sentences whose proportion of Chinese characters is less than 50% [15] and divide these texts into complete sentences using clause-ending punctuations such as periods "。", "？", "！",etc.

Before generating the p-corpus, we must determine how many errors are produced in each sentence. Many people have done research on this issue. [12] proposed the number of errors in one sentence should not exceed 2, while [11] proposed an average of 2.7 errors in one misspelled sentence. When using the Pinyin input method, the basic unit of input is a word, not a single character [18]. For example, when using

[6] According to statistics, there are 11 groups of fuzzy sounds in Chinese characters: z-zh, c-ch, s-sh, l-n, f-h, r-l, an-ang, en-eng, in-ing, ian-iang, uan-uang.

[7] https://catalog.ldc.upenn.edu/LDC2011T13, these articles reported have undergone a rigorous editing process and are considered to be all correct.

[8] http://www.openslr.org/resources/33/data_aishell, this speech library is transcoded by professional voice proofreaders and pass strict quality inspection. The correct rate of AlShell is above 95%.

the Pinyin input method to input the sentences: “任务是生成语料库”, the basic input unit is the word (任务/是/生成/语料库), not the character (任/务/是/生/成/语/料/库). Therefore, this paper lets every sentence contain a misspelled word. The word could consist of one character or more [3], and the length of the word is determined by the word segmentation[9]. According to statistics, each misspelled sentence in p-corpus has an average of 1.54 errors.

Generate M-SS type sentences. Figure 2 shows the generation process of an M-SS type sentence. Firstly, the sentences are processed by word segmentation. Secondly, a Chinese word in each sentence is randomly selected. Thirdly, we use the pypinyin[10] toolkit to extract sounds of the words. Fourthly, we use the Pinyin2Hanzi[11] toolkit to convert the sounds into corresponding Chinese words. Lastly, M-SS type sentences are generated by replacing the original words with the generated words.

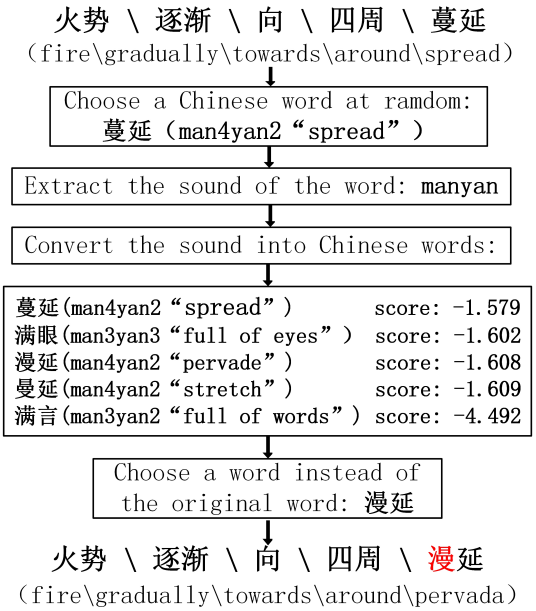


Fig. 2. The generation process of an M-SS type sentence.

Note that, when the sounds are converted to Chinese words, all the Chinese words with the same sounds will be generated, and each word has a corresponding score[12]. When using the generated words to replace the original words, we set the corresponding replacement probability for each generated word. When the words are the same as the original words, the replacement probability is 0. Then, the words different from the original words are sorted in descending order. The score of the i-th word is set to

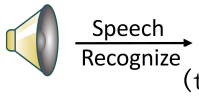
[9] The word segmentation tool used in this paper is jieba. https://github.com/fxsjy/jieba
[10] It can extract the sounds of the Chinese characters. https://github.com/mozillazg/python-pinyin
[11] It can convert the sounds into Chinese characters. https://github.com/letiantian/Pinyin2Hanzi
[12] The score is calculated based on the HMM principle. In general, the more commonly used words, the higher the score. https://github.com/letiantian/Pinyin2Hanzi

$Socre(i)$, and the corresponding replacement probability is set to $RP(i)$. Equation 1 gives the calculation process of $RP(i)$. In general, the higher the score, the greater the replacement probability.

$$\begin{aligned} RP(i) &= \frac{1/\text{Socre}(i)}{Sum} \\ Sum &= \sum_{i=1}^{n} \left(\frac{1}{Socre(i)} \right) \end{aligned} \tag{1}$$

$RP(i)$ represents the replacement probability of the i-th word, $Socre(i)$ denotes the score of the i-th word, and n denotes the number of the words different from the original words.

Generate M-MS type sentences. A major challenge in generating MS-type sentences is that there are no rules to follow [18]. Our paper proposes a method for generating M-MS type sentences using Baidu ASR interface[13]. The basic generation method is shown in Figure 3. It is worth noting that Baidu ASR interface will generate multiple types of errors, and we just collect the sentences with wrong words having similar sounds (Similar sounds means that the pinyin editing distance is 1 or 2).



刺青 / 标注 / 技术 / 用途 / 广泛
(tattoo/marking/technology/use/widely)

Correct sentence:词性 / 标注 / 技术 / 用途 / 广泛
(part of speech/marking/technology/use/widely)

Fig. 3. An M-MS type sentence generated by ASR. The Chinese word marked in red is the misrecognized word. “词性” (ci2xing4 “part of speech”) is incorrectly recognized as “刺青”(ci4qing1 “tattoo”), and they are similar sounds.

When converting the Mandarin speech library AlShell into texts, there are many types of spelling errors. It is easy to identify the errors types by comparing with the corresponding correct sentences. We collect the M-MS type sentences with only one misspelled word. As a result, we generated 12,031 M-MS type sentences using the above method and the statistics are shown in Table 2. D(M-SS) represents the data of M-SS type sentences, D(M-MS) represents the data of M-MS type sentences, D represents the combination of D(M-SS) and D(M-MS), ASL represents the average sentences length, and ANE represents the average number of errors per sentence.

	Sentences	Characters	Errors	ASL	ANE
D(M-SS)	100000	2548514	153312	25.5	1.53
D(M-MS)	12031	233401	19250	19.4	1.6
D	112031	2781915	172562	24.8	1.54

Table 2. Statistics of the M-SS type sentences and M-MS type sentences.

3.3 Building the v-corpus

This section will introduce two parts, one is the types of the errors generated by the voice input method, and the other is the methods of constructing the v-corpus.

[13] https://github.com/baidubce/pie/tree/master

The misspelled sentences generated by using the voice input method can be divided into two categories according to whether the lengths of those are the same as the original sentences, as shown in Table 3.

Correct Sentences	Misspelled Sentences	Type
任务/是/生成/语料库(length=8) task/is/generate/corpus	任务/是/商场/与/辽库(length=8) task/is/mall/and/distant corpus	S
五氧化二磷/可以/溶于/水(length=10) phosphorus pentoxide/can/soluble/water	养花/二零/可以/溶于/水(length=9) raising flowers/20/can/soluble/water	D

Table 3. Two categories of sentences are generated by the voice input method. S denotes the misspelled sentences the same length as the correct sentences. D denotes the misspelled sentences different from the correct sentences.

We use the Kaldi[14][9] and Baidu ASR interface to build the v-corpus. The basic principle is shown in Figure 3. We only collect S type sentences generated by the two ASR tools. Because when they are different in length, many labels will be marked incorrectly, which will bring lots of noise. Take the second sentence in Table 3 as an example, as shown in Figure 4, only the first 4 characters are incorrect. However, this situation causes all subsequent characters to be marked as misspelled characters.

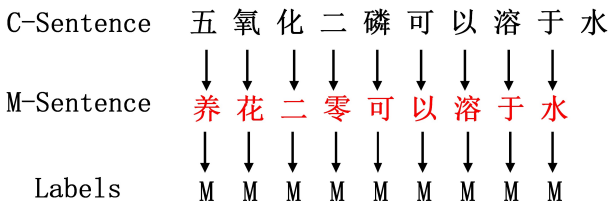


Fig. 4. The labels are marked incorrectly when the correct sentence is different from the misspelled sentence in length. C-Sentence denotes correct sentence, and M-Sentence denotes misspelled sentence.

The raw data is also the Mandarin speech library AIShell, and the v-corpus statistics are shown in Table 4.

	Sentences	Characters	Errors	ASL	ANE
v-corpus(Kaldi)	88717	2135646	187912	24.1	2.11
v-corpus(Baidu)	68376	1624578	135481	23.8	1.98
v-corpus	157093	3760224	323393	24	2.06

Table 4. Statistics of the v-corpus. v-corpus(Kaldi) represents the corpus generated based on Kaldi, and v-corpus(Baidu) represents the corpus generated based on Baidu ASR interface.

4 Evaluation

We qualitatively and quantitatively evaluate the corpora. The qualitative evaluation aims to evaluate whether the misspelled sentences in our corpora can simulate those in real

[14] A speech recognition kit. https://github.com/kaldi-asr/kaldi

language situation. The quantitative evaluation aims to evaluate whether a better check effect can be achieved using our corpora than the benchmark corpus.

This paper uses the BiLSTM-CRF model to quantitatively evaluate, and the model diagram shows in Figure 5 [8]. BiLSTM layer is used to extract sentence features, and CRF layer is used to automatically complete sequence tagging.

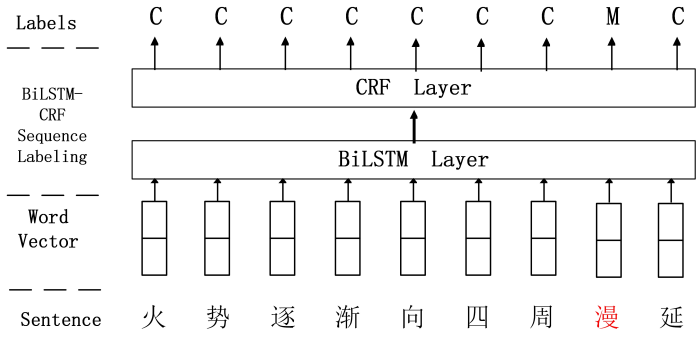


Fig. 5. BiLSTM-CRF model structure diagram.

4.1 Evaluate the p-corpus

We use the corpus provided by [12] as the benchmark corpus. It is worth noting that they [12] did not build different corpora from the perspective of input methods, but wanted to build a corpus to check all forms of text. The statistics of the benchmark corpus are shown in Table 5.

Qualitative evaluation. We find some texts generated by the Pinyin input method, including student papers, published books and articles published on the Internet. A total of 2000 sentences with Chinese spelling errors were selected. The number of the M-SS type sentences and the M-MS type sentences are 1698 and 302, respectively, and the ratio of them is close to 17:3. Hence, we construct the p-corpus according to this ratio. The statistics of the p-corpus is shown in Table 5.

	Sentences	Characters	Errors	ASL	ANE
b-corpus	80000	1632458	132524	20.41	1.66
p-corpus(M-SS)	68000	1734316	104051	25.5	1.53
p-corpus(M-MS)	12000	216951	19231	18.1	1.6
p-corpus	80000	1951267	123282	24.4	1.54

Table 5. Statistics of the benchmark corpus and the p-corpus. b-corpus represents the benchmark corpus, p-corpus(M-SS) represents the M-SS type sentences in p-corpus, and p-corpus(M-MS) represents the M-MS type sentences in p-corpus.

We randomly select 250 sentences in the p-corpus and in real language situation respectively. The two types of sentences construct the test set. In addition, we invite 5 college students and giving each person 50 misspelled sentences in the p-corpus and 50 misspelled sentences in real language situation. Let them pick out the sentences in the p-corpus. The quality of the corpus is measured by S-Recall (the recall from the

students' tests) and S-Precision (the precision from the students' tests), and Equation 2 shows the calculation process of the S-Recall and the S-Precision. The test results demonstrate in Table 6.

$$\text{S-Recall} = \frac{NP}{100} \quad \text{S-Precision} = \frac{NP}{\text{Total}} \tag{2}$$

Where Total denotes the total number of misspelled sentences selected by the students. NP denotes the *n*umber of misspelled sentences selected by the students belonging to the *p*-corpus. 100 denotes the number of sentences assigned to each college student.

	S1	S2	S3	S4	S5
Total	13	19	28	32	7
NP	8	11	17	18	4
S-Recall	0.08	0.11	0.17	0.18	0.04
S-Precision	0.62	0.58	0.61	0.56	0.58

Table 6. Qualitative assessment results. S1 to S5 represent 5 college students respectively.

It can be seen from Table 6 that S-Recall is very low, which means that the two types of sentences are very similar and it is difficult to distinguish between the two. In the test set, the number of the two types sentences is the same, so if S-Precision is equal to 0.5, it can be considered that the college students can't distinguish the two. The experimental S-Precision is about 0.6, which is very close to 0.5. Thus, we can believe the sentences in the p-corpus can simulate misspelled sentences in real language situation.

Quantitative evaluation. As far as we know, no one has built a corpus specifically for checking the texts generated by the Pinyin input method. So this paper uses 2000 misspelled sentences collected in real language situation as the test set. We set up five training sets of different sizes: Trn-10k, Trn-20k, Trn-30k, Trn-40k, Trn-50k. The quality of the p-corpus is measured by calculating precision, recall, and F1 [11]. The test results are shown in Table 7. By observing the test results, we draw the following conclusions.

	Trn-10K		Trn-20K		Trn-30K		Trn-40K		Trn-50K	
	bc	pc	bc	pc	bc	pc	bc	pc	bc	pc
Precision	41.31	44.57	50.36	59.91	56.42	69.11	61.39	75.71	61.02	77.12
Recall	47.29	51.35	61.22	66.25	71.52	77.57	75.14	82.22	81.01	87.43
F1	43.87	47.82	54.94	62.39	62.11	72.92	68.58	78.19	70.29	80.94

Table 7. The test results of the benchmark corpus and the p-corpus. bc denotes the b-corpus, and pc denotes the p-corpus.

(1) Compared to the benchmark corpus, the sentences in the p-corpus are closer to those in real language situation. As we all know, the closer the sentences in the corpus are to those in real language situation, the better the test results will be. We can see from Table 7 that compared with the benchmark corpus, the p-corpus has achieved better test results, so we can believe that our corpus is better.

(2) The size of the training data set is very important. From Table 7, as the training sets become larger, the three indicators have a steady upward trend, which indicates the

model has learned more information. Thus we can draw such a conclusion that the size of the training sets is very important for data-driven approaches.

(3) As the sizes of the two training sets grow, benchmark corpus brings more noise. From Table 7, as the sizes of the two training sets grow, precision improvement is very obvious. However, the increase in recall rate is not very significant, which indicates that benchmark corpus causes more wrong tags. Therefore, we can believe that the benchmark corpus brings more noise.

4.2 Evaluate the v-corpus

The sentences in the v-corpus are generated by the ASR tools, and they come from the real language situation; hence we just only quantitatively evaluate the v-corpus. When evaluating the quality of the v-corpus, the training sets are 50k in size, and the test sets are 5k. In addition to using the benchmark corpus for testing (called Benchmark Test), we also do three sets of comparison tests: Corresponding Test, Cross Test, and Mixed Test. Figure 6 shows the four sets tests.

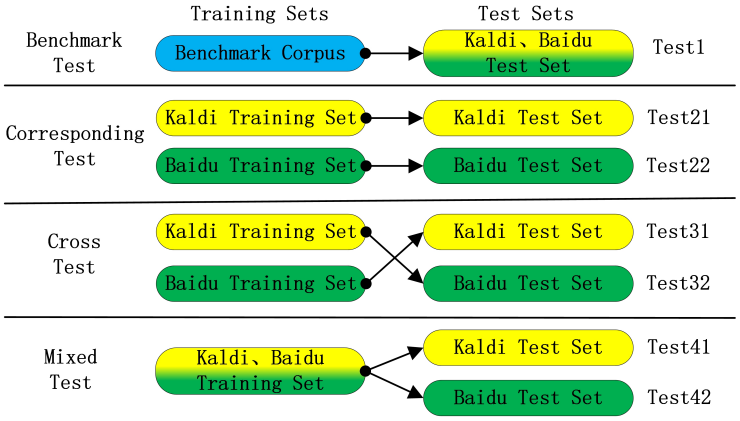


Fig. 6. The display of the four sets tests.

Benchmark Test (called Test 1): the training set is benchmark corpus, and the test set is generated by Kaldi and Baidu ASR interface together. **Corresponding Test:** the training sets and the test sets are generated by the same ASR tool. The evaluation of the training set and the test set both from Kaldi is called Test21. The evaluation of the training set and the test set both from Baidu ASR interface is called Test22. **Cross Test:** the training sets and the test sets are generated by different ASR tools. The evaluation of the training set from Kaldi and test set from the Baidu ASR interface is called Test31. The evaluation of the training set from the Baidu ASR interface and the test set from Kaldi is called Test32. **Mixed Test:** the training set is generated by the two tools together, and the test sets generated by different ASR tools. In detail, the evaluation of the test set from the Kaldi is called Test41 and the evaluation of the test set from the Baidu ASR interface is called Test42.

Table 8 shows the results of the four sets tests. We have the following conclusions.

Compared to the benchmark corpus, the v-corpus could get a better checking effect. From Table 8, these results of the Corresponding Test, the Cross Test, and the

	Benchmark Test	Corresponding Test		Cross Test		Mixed Test	
	Test1	Test21	Test22	Test31	Test32	Test41	Test42
Precision	58.01	77.96	78.12	71.21	70.98	74.42	73.91
Recall	69.33	87.67	85.81	82.22	82.41	85.16	86.12
F1	63.19	83.38	81.56	76.69	76.16	79.81	80.81

Table 8. The results of the four sets tests.

Mixed Test are higher than the Benchmark Test, which means that our corpus is more suitable for checking the texts generated by the voice input method.

Different ASR tools generate different forms of the errors. The results of the Corresponding Test is higher than the Cross Test and the Mixed Test, at the same time, the results of the Cross Test is lower than the Corresponding Test and the Mixed Test. Therefore, we can believe that the forms of the errors are different when they are generated by different ASR tools.

The generalization ability will be improved when the training sets are generated by different ASR tools. There are many different ASR tools, and it's hard to train the corresponding spelling check model for every ASR tool. The results of the Mixed Test gives us good inspiration. Although the results of the Mixed Test is not as good as the Corresponding Test, it better than the Cross Test and the Benchmark Test. Therefore, when we want to check the texts generated by different ASR tools, the training set should be generated by using multiple ASR tools as much as possible.

5 Conclusions and Future Work

At present, due to the lack of a large number of high quality annotated corpora, many advanced data-driven models cannot be applied to the task of CSC. This paper proposes new approaches to automatically build spelling corpora based on the input method. The corpora are used to check the texts generated by the Pinyin input method and the voice input method, respectively. The evaluation results demonstrate that the misspelled sentences in our corpora can simulate those in real language situation, and using them for the task of CSC can get a better effect than the benchmark corpus. A complete spelling checker is a writing assistance tool which provides users with better word suggestions by automatically detecting spelling errors in documents. Therefore, in the future work, we plan to develop error correction based on spelling check.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61672040) , Beijing Urban Governance Research Center and the North China University of Technology Startup Fund. The corresponding author is Hao Wang.

References

1. Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., et al.: End to end speech recognition in english and mandarin (2016)

2. Bu, H., Du, J., Na, X., Wu, B., Zheng, H.: Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In: 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). pp. 1–5. IEEE (2017)
3. Chang, T.H., Chen, H.C., Tseng, Y.H., Zheng, J.L.: Automatic detection and correction for chinese misspelled words using phonological and orthographic similarities. In: Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing. pp. 97–101 (2013)
4. Chen, Y.Z., Wu, S.H., Yang, P.C., Ku, T., Chen, G.D.: 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 (2011)
5. Chen, Z., Lee, K.F.: A new statistical approach to chinese pinyin input. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (2000)
6. Hsieh, Y.M., Bai, M.H., Huang, S.L., Chen, K.J.: Correcting chinese spelling errors with word lattice decoding. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) **14**(4), 18 (2015)
7. Liu, C.L., Lai, M.H., Tien, K.W., Chuang, Y.H., Wu, S.H., Lee, C.Y.: Visually and phonologically similar characters in incorrect chinese words: Analyses, identification, and applications. ACM Transactions on Asian Language Information Processing (TALIP) **10**(2), 10 (2011)
8. Liu, Y., Zan, H., Zhong, M., Ma, H.: Detecting simultaneously chinese grammar errors based on a bilstm-crf model. In: Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. pp. 188–193 (2018)
9. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The kaldi speech recognition toolkit. Tech. rep., IEEE Signal Processing Society (2011)
10. Sak, H., Senior, A., Rao, K., Beaufays, F.: Fast and accurate recurrent neural network acoustic models for speech recognition. arXiv preprint arXiv:1507.06947 (2015)
11. Wang, D., Fung, G.P.C., Debosschere, M., Dong, S., Zhu, J., Wong, K.F.: A new benchmark and evaluation schema for chinese typo detection and correction. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
12. Wang, D., Song, Y., Li, J., Han, J., Zhang, H.: A hybrid approach to automatic corpus generation for chinese spelling check. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2517–2527 (2018)
13. Wu, S.H., Liu, C.L., Lee, L.H.: Chinese spelling check evaluation at sighan bake-off 2013. In: Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing. pp. 35–42 (2013)
14. Yang, S., Zhao, H., Wang, X., Lu, B.l.: Spell checking for chinese. In: LREC. pp. 730–736 (2012)
15. Yongwei, Z., Qinan, H., Fang, L., Yueguo, G.: Cmmc-bdrc solution to the nlp-tea-2018 chinese grammatical error diagnosis task. In: Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. pp. 180–187 (2018)
16. Yu, J., Li, Z.: Chinese spelling error detection and correction based on language model, pronunciation, and shape. In: Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing. pp. 220–223 (2014)
17. Yu, L.C., Lee, L.H., Tseng, Y.H., Chen, H.H.: Overview of sighan 2014 bake-off for chinese spelling check. In: Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing. pp. 126–132 (2014)
18. Zheng, Y., Li, C., Sun, M.: Chime: An efficient error-tolerant chinese pinyin input method. In: Twenty-Second International Joint Conference on Artificial Intelligence (2011)