Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013

Shih-Hung Wu
Chaoyang Univ. of Technology
Taichung, Taiwan
shwu@cyut.edu.tw

Chao-Lin Liu
National Chengchi Univ.
Taipei, Taiwan
chaolin@nccu.edu.tw

Lung-Hao Lee
National Taiwan Univ.
Taipei, Taiwan
lhlee@ntnu.edu.tw

Abstract

This paper introduces an overview of Chinese Spelling Check task at SIGHAN Bake-off 2013. We describe all aspects of the task for Chinese spelling check, consisting of task description, data preparation, performance metrics, and evaluation results. This bake-off contains two subtasks, *i.e.*, error detection and error correction. We evaluate the systems that can automatically point out the spelling errors and provide the corresponding corrections in students' essays, summarize the performance of all participants' submitted results, and discuss some advanced issues. The hope is that through such evaluation campaigns, more advanced Chinese spelling check techniques will be emerged.

1 Introduction

Spelling check is a common task in every written language, which is an automatic mechanism to detect and correct human errors. A spelling checker should have both capabilities consisting of error detection and error correction. Spelling error detection is to indicate the various types of spelling errors in the text. Spelling error correction is further to suggest the correct characters of detected errors. Spelling check must be done within a context, say a sentence or a long phrase with a certain meaning, and cannot be done within one word (Mays et al., 1991).

However, spelling check in Chinese is very different from that in English or other alphabetic languages. There are no word delimiters between words and the length of each word is very short. There are several previous studies addressing the Chinese spelling check problem. Chang (1995) has proposed a bi-gram language model to substitute the confusing character for error detection and correction. Zhang et al. (2000) have presented an approximate word-matching algorithm to detect and correct Chinese spelling errors using operations of character substitution, insertion, and deletion. Ren et al. (2001) have proposed a hybrid approach that combines a rule-based method and a probability-based method to automatic Chinese spelling checking. Huang et al. (2007) have proposed a learning model based on Chinese phonemic alphabet for spelling check. Most of the Chinese spelling errors were originated from phonologically similar, visually similar, and semantically confusing characters (Liu et al., 2011). Empirically, there were only 2 errors per student essay on average in a learners' corpus (Chen et al., 2011). How to evaluate the false-alarm rate of a spelling check system with normal corpus was also a hard task (Wu et al., 2010). Up to date, there are no commonly available data sets for spelling check for Chinese. This motivates us to develop such data sets as benchmarks for fairly evaluating the performance of state-of-the-art Chinese spelling checkers.

At SIGHAN Bake-off 2013, we organize the Chinese Spelling Check task that provides an evaluation platform for developing and implementing automatic Chinese spelling checkers. Two subtasks, *i.e.*, error detection and error correction, are designed to evaluate complete function of a spelling checker. The first subtask focuses on the ability of error detection. Given a complete sentence, the checker should detect if there are errors in the input, and point out the error locations of incorrect characters. The second subtask aims at the quality of error correction. In addition to indicating the error locations, the checker should suggest the correct characters. The hope is that, through such evaluation campaigns, more advanced Chinese spelling check techniques will be emerged.

We give an overview of Chinese Spelling task at SIGHAN Bake-off 2013. The rest of this article is organized as the follows. Section 2 details the designed task, consisting of two subtasks, *i.e.*, error detection and error correction. Section 3 introduces the data sets provided in this eval-

Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), pages 35–42,
Nagoya, Japan, 14 October 2013.

uation. Section 4 proposes the evaluation metrics for both subtasks. Section 5 presents the results of participants' approaches for performance comparison. Section 6 elaborates on the semantic and pragmatic aspects of automatic correction of Chinese text. Finally, we conclude this paper with the findings and future research direction in the Section 7.

2 Task Description

The goal of this task is to evaluate the ability of a system on Chinese spelling check. The task can be further divided into two subtasks: error detection and error correction. We detail as the follows.

2.1 Subtask 1: Error Detection

For the error detection subtask, complete Chinese sentences with/without spelling errors will be given as the input, the system should return the locations of the incorrect characters. Each character or punctuation occupies 1 spot for counting location. The error detection problem is a yes/no question plus the locations of errors. If the input sentence (each given a serial number *NID*) contains no spelling errors, the system should return: *NID*, 0. If the input contains at least one spelling errors, the output format is: *NID, location [, location]**, where the symbol "*" indicates there is zero or more of the predicting element "[, location]". We give the following example for more information. In this example, the 27th character is wrong, the correct one should be "挫".

- Input: (*NID*=99999) 在我的人生中沒有風災大浪，但我看過許多勇敢的人，不怕措折的奮鬥，這種精神值得我們學習。
- Output: 99999, 27

2.2 Subtask 2: Error Correction

For the error correction subtask, the input texts are complete Chinese sentences with spelling errors. The system should return the locations of the incorrect characters, and must point out the correct characters. The error correction problem is a follow-up problem of error detection for checking spelling errors. Since the input sentence contains at least one spelling error, the output format is: *NID [, location, correction]*+, where "+" sign indicates there is one or more of the predicting element "[, location, correction]". Take the following example as instance, the 16th and 29th characters are wrong, the correct ones are "徵" and "間", respectively.

```
<DOC Nid="00018">
<p>有些人會拿這次的教訓來勉勵自己，好讓自己在打混摸魚時警悌，使自己比以前更好、更進步。
</p>
<TEXT>
<MISTAKE wrong_position=28>
<wrong>警悌</wrong>
<correct>警惕</correct>
</MISTAKE>
</TEXT>
</DOC>
```

Figure 1. A sample set in terms of XML format

- Input: (NID=88888) 擁有六百一十年歷史的崇禮門，象微著南韓人的精神，在一夕之門，被火燒得精光。
- Output: 88888, 16, 徵 29, 間

3 Data Preparation

3.1 Sample Set and Similar Character Set

We provided the Sample Set and Similar Character Set as the linguistic resources for this evaluation. The policy of our evaluation is an open test. Participants can employ any linguistic and computational resources to do identification and corrections.

In Sample Set, there are 700 samples selected from students' essays, which are represented in XML format shown in Figure 1. A half of these samples contain at least one error and the remaining samples do not contain any errors.

The set of Chinese characters with similar shapes, same pronunciations, and similar pronunciations is especially useful for this task. Details about these sets are described in the previous work (Liu et al., 2011). For example, the set of similar shape of the character "可" and the set of similar pronunciation of the character "隔" are listed as follows:

- Similar Shape: 可, 何呵坷奇河柯苛阿倚寄崎荷蚵軻.
- Similar Pronunciation: 隔, 郃革格咯骼閣膈閡葛鬲鎘蛤.

Test Set	Subtask1	Subtask2
# of sentences	1,000	1,000
# of sentences with errors	300	1,000
# of error characters	376	1265
Average # of errors in sentences with errors	1.253	1.265
Average length of sentences	68.711	74.328
Sentence-level error percentage (%)	30%	100%
Character-level error percentage (%) (with punctuation)	0.547%	1.702%
Character-level error percentage (%) (without punctuation)	0.611%	1.902%

Table 1. Descriptive statistics of the test sets

3.2 Test Set

Table 1 shows the statistics of our prepared test sets. The sentences were collected from 13 to 14-year-old students' essays in formal written tests. The average length of sentences is about 70 characters, which is a compromise to the writing style of the students. Most of the students cannot break their sentences into short and clear ones. To preserve the context, we kept the whole long sentences as they were written on the examination paper. The character-level error percentage is about 0.5% and 2% for subtask 1 and subtask 2, respectively. The error rate is higher than it was in the original corpus, since we deleted most sentences without any error to reduce the test set size.

There were 1,000 Chinese texts selected from students' essays that covered various common errors for each subtask, respectively. The teachers manually identified the errors embedded in Chinese sentences. There is some inconsistency between teachers on the standard of whether it is an error or not. There is no authority on the standard, which is an implicit consensus of the teachers. In our prepared test data set, 300 out of 1,000 test sentences contain errors in subtask 1. In subtask 2, each of the 1,000 test sentences contains one or more errors.

We found that there were some controversial cases, especially about the usage of Chinese idioms. There are many ways to express an idiom and some of them might be considered as errors. We did our best to reduce the inconsistency manually during the preparation of the test set by deleting the controversial cases. On the other hand, we preserved as many errors as possible in the test set, such that system developers could find the kinds of errors that students actually produced. There are some common errors that occur with high frequencies, but we did not delete them so that the distribution of errors can be kept and might be used for educational purposes.

We met some difficult issues during test set preparation. The first difficulty is to ensure that there is no more error other than the pointed out ones. There is almost no question that errors pointed out by the teachers are errors. However, there are errors we detected but not pointed out by teachers. Maybe they are minor errors that some teachers omitted or did not think they are errors. We manually deleted several sentences with such cases. The second difficulty is not to modify the sentences too much while preserving the original context. Since the test set is selected from students' essays, there are some ungrammatical sentences. We modified them such that the only errors are spelling errors not other syntactical errors or improper co-occurrences.

4 Performance Metrics

4.1 Metrics of Error Detection

For error detection subtask, we adopt sentence-level metrics for performance evaluation. Since the number of error characters is very small comparing to all the characters. It is not suitable to use the number of character to calculate accuracy. Therefore, in this bake-off, we adopt the numbers of sentences as the unit of performance metrics. The computation formulas are listed as follows:

- False-Alarm Rate (**FAR**)= # of sentences with false positive errors / # of testing sentences without errors
- Detection Accuracy (**DA**)= # of sentences with correctly detected results / # of all testing sentences
- Detection Precision (**DP**)= # of sentences with correctly detected errors / # of sentences the evaluated system reported to have errors
- Detection Recall (**DR**)= # of sentences with correctly detected errors / # of testing sentences with errors
- Detection F1 (**DF1**)= 2*DP*DR / (DP+DR)
- Error Location Accuracy (**ELA**)= # of sentences with correct location detection / # of all testing sentences
- Error Location Precision (**ELP**)= # of sentences with correct error locations / # of sentences the evaluated system reported to have errors

- Error Location Recall (**ELR**)= # of sentences with correct error locations / # of testing sentences with errors
- Error Location F1 (**ELF1**)= 2*ELP*ELR / (ELP+ELR)

The criterion for judging corrections is that the output should be completely identical with the gold standard. For example, give 5 testing inputs with gold standard shown as "0022, 43, 76", "0023, 0", "0024, 0", "0025, 72, 79", and "0026, 103". The system may output the results shown as "0022, 43, 55, 80", "0023, 10", "0024, 0", "0025, 72, 79", and "0026, 103". The evaluated tool will yield the following performance metrics:

- FAR=0.5 (=1/2)
  Notes: #{"0023"}/ #{"0023", "0024"}
- DA=0.75 (=4/5)
  Notes: #{"0022", "0024", "0025", "0026"}/ #{"0022", "0023", "0024", "0025", "0026"}
- DP=0.75 (=3/4)
  Notes: #{"0022", "0025", "0026"}/ #{"0022", "0023", "0025", "0026"}
- DR=1 (=3/3)
  Notes: #{"0022", "0025", "0026"}/ #{"0022", "0025", "0026"}
- DF1= 0.8571 (=2*0.75*1/(0.75+1))
- ELA=0.6 (=3/5)
  Notes: #{"0024, 0", "0025, 72, 79", "0026, 103"}/ #{"0022, 43, 76", "0023, 0", "0024, 0", "0025, 72, 79", "0026, 103"}
- ELP=0.5 (=2/4)
  Notes: #{"0025, 72, 79", "0026, 103"}/ #{"0022, 43, 55, 80", "0023, 10", "0025, 72, 79", "0026, 103"}
- ELR= 0.6667 (2/3)
  Notes: #{"0025, 72, 79", "0026, 103"}/ #{"0022, 43, 76", "0025, 72, 79", "0026, 103"}
- ELF1=0.5714 (=2*0.5*0.6667/(0.5+0.6667))

4.2 Metrics of Error Correction

For error correction subtask, we adopt the similar metrics. The computations are formulated as follows:

- Location Accuracy (**LA**)= # of sentences correctly detected the error location / # of all testing sentences
- Correction Accuracy (**CA**)= # of sentences correctly corrected the error / # of all testing sentences
- Correction Precision (**CP**)= # of sentences correctly corrected the error / # of sentences the system returns corrections.

The criterion for judging corrections is the same with subtask 1. Take a set of gold standard shown as {"00366, 1, 倘", "00367, 10, 的", "00368, 39, 嘩, 63, 葉, 89, 嫩", "00369, 16, 炭, 48, 作", "00370, 49, 已"} for example, if the system output the results: {"00366, 1, 趟", "00367, 10, 的", "00368, 39, 嘩, 63, 葉", "00369, 16, 炭, 48, 作"} , the evaluated tool will yield the follows:

- LA=0.6 (=3/5)
  Notes: #{"00366, 1", "00367, 10", "00369,

Participant (Ordered by abbreviations of names)	Subtask 1	Subtask2
Agency for Science, Technology and Research (**A*STAR**)	0	0
Heilongjiang University (**HLJU**)	3	3
National Kaohsiung University of Applied Sciences & National Taiwan Normal University (**KUAS & NTNU**)	1	1
Nara Institute of Science and Technology (**NAIST**)	3	3
National Chiao Tung University & National Taipei University of Technology (**NCTU & NTUT**)	2	2
National Chiayi University (**NCYU**)	3	3
Nanjing University of Posts and Telecommunications (**NJUPT**)	0	0
National Tsing Hua University (**NTHU**)	3	3
National Taiwan Ocean University (**NTOU**)	3	3
University of Oxford (**OX**)	0	0
Peking University (**PKU**)	3	0
Chinese Knowledge and Information Processing Group, IIS, Academia Sinica (**SinicaCKIP**)	3	3
Intelligent Agent Systems Lab, IIS, Academia Sinica (**SinicaIASL**)	2	2
Speech, Language and Music Processing Lab, IIS, Academia Sinica & National Taiwan University (**SinicaSLMP & NTU**)	3	3
Shanghai Jiao Tong University (**SJTU**)	3	3
University of Macau (**UMAC**)	0	0
Yuan Ze University & National Cheng Kung University (**YZU & NCKU**)	1	1
Total	33	30

Table 2. Result submission statistics of all participants

Participant	Approach	Usage of Provided Corpus	Additional Resources
HLJU	N-gram Model	Both	Sinica Corpus
KUAS & NTNU	Phonological similarity, Orthographic similarity, Bi-gram Linear Regression, Rule base Model	None	Sinica Corpus, Sinica Treebank, Chinese Electronic Dictionary, and Chinese Orthography Database
NAIST	Language Model + SVM, Language Model + Statistical Machine Translation Model + SVM	Both	Chinese Gigaword, Sinica Corpus of SIGHAN Bake-off 2005, and CC-CEDICT
NCTU & NTUT	CRF-based Chinese Parser, Trigram Language Model	Both	Sinica Corpus, CIRB030, the Taiwan Panorama Magazine 4 and the Wikipedia
NCYU	N-gram + Inverted Index	Both	E-HowNet, and Gathered corpus for training n-gram
NTHU	Machine Translation Language Model, Rule based model	Both	TWWaC, Sinica Corpus, Chinese dictionary, and Chinese Idioms
NTOU	Language Model + Heuristic Rules	Both	Sinica Corpus
PKU	Maximum Entropy Model	Both	Chinese Gigaword
SinicaCKIP	Unknown Word Detection, Word Segmentation, Language Model	Similar Character Set	CKIP lexicon, Sinica Corpus, and Google 1T n-gram
SinicaIASL	Reliable Phonological Sequence Matcher, Word Segmentation, Homophone Dictionary + N-gram Model, Shape Correction Module, Language Model	Both	Revised Chinese Dictionary, Xiaoxuetang Philology Database, LDC news corpus, Chinese Information Retrieval Benchmark (CIRB), Frequent Errors List from the Web, and Google 1T n-gram
SinicaSLMP & NTU	N-gram model, Topic model	Both	Chinese Gigaword, Sinica Corpus, and Search Engine (Baidu)
SJTU	Shortest Path Word Segmentation Algorithm, Language Model, Mutual Information	Both	SogouW Dictionary, Sinica corpus of SIGHAN Bake-off 2005, IRSTLM, and OpenCC
YZU & NCKU	Web-based Score	Similar Character Set	Chinese Gigaword, and Search Engine (Google)

Table 3. A summary of participants' developed systems

16, 48"}/ #{"00366, 1", "00367, 10", "00368, 39, 63, 89", "00369, 16, 48", "00370, 49"}

- CA=0.4 (=2/5)
  Notes: #{"00367, 10, 的", "00369, 16, 炭, 48, 作"}/ #{"00366, 1, 倘", "00367, 10, 的", "00368, 39, 嘩, 63, 葉, 89, 嫩", "00369, 16, 炭, 48, 作", "00370, 49, 已"}
- CP=0.5 (=2/4)
  Notes: #{ "00367, 10, 的", "00369, 16, 炭, 48, 作"}/ #{"00366, 1, 趟", "00367, 10, 的", "00368, 39, 嘩, 63, 葉", "00369, 16, 炭, 48, 作"}

5 Evaluation Results

Table 2 shows the participant teams and their testing submission statistics. This task of bake-off 2013 attracted 17 research teams. There are 9 teams that come from Taiwan, *i.e.*, KUAS & NTNU, NCTU & NTUT, NCYU, NTHU, NTOU, SinicaCKIP, SinicaIASL, SinicaSLMP & NTU, and YZU & NCKU. The other 5 teams originate from China, *i.e.*, HLJU, NJUPT, PKU, SJTU, and UMAC. The remaining 3 ones are A*STAR from Singapore, NAIST from Japan, and OX from United Kingdom.

Among 17 registered teams, 13 teams submitted their testing results. For formal testing, each participant can submit at most three runs that use different models or parameter settings. Table 3 summarizes the participants' developed approaches and the usage of linguistic resources for this bake-off evaluation. We can observe that most of participants adopt statistical approaches such as n-gram model, language model, machine translation model, and topic model. In addition to the Sample Set and the Similar Character Set,

Submission	FAR	DA	DP	DR	DF1	ELA	ELP	ELR	ELF1
HLJU-Run1	0.6857	0.5140	0.3798	0.98	0.5474	0.3010	0.1047	0.2700	0.1509
HLJU-Run2	0.6529	0.5290	0.3849	0.9533	0.5484	0.3390	0.1292	0.3200	0.1841
HLJU-Run3	0.6929	0.5100	0.3782	0.9833	0.5463	0.2960	0.1038	0.2700	0.1500
KUAS & NTNU-Run1	0.2257	0.7890	0.6099	0.8233	0.7007	0.6940	0.3753	0.5067	0.4312
NAIST-Run1	0.2929	0.7460	0.5504	0.8367	0.664	0.6450	0.3289	0.5000	0.3968
NAIST-Run2	0.0543	0.8120	0.7979	0.5000	0.6148	0.7640	0.5426	0.3400	0.4180
NAIST-Run3	0.2243	0.7770	0.5985	0.7800	0.6773	0.6980	0.3964	0.5167	0.4486
NCTU & NTUT-Run1	0.0243	0.7220	0.6964	0.1300	0.2191	0.7110	0.5000	0.0933	0.1573
NCTU & NTUT-Run2	0.8329	0.4110	0.3352	0.9800	0.4995	0.2570	0.1596	0.4667	0.2379
NCYU-Run1	0.2371	0.7380	0.5514	0.6800	0.609	0.6230	0.2405	0.2967	0.2657
NCYU-Run2	0.2129	0.7610	0.5850	0.7000	0.6374	0.6520	0.2813	0.3367	0.3065
NCYU-Run3	0.0929	0.8250	0.7451	0.6333	0.6847	0.7480	0.4431	0.3767	0.4072
NTHU-Run1	0.0386	0.8480	0.8663	0.5833	0.6972	0.8090	0.6733	0.4533	0.5418
NTHU-Run2	0.0471	0.8570	0.8520	0.6333	0.7265	0.8150	0.6637	0.4933	0.5660
NTHU-Run3	0.0514	**0.8610**	0.8455	0.6567	0.7392	**0.8200**	0.6695	0.5200	**0.5854**
NTOU-Run1	0.9800	0.3140	0.3043	**1.0000**	0.4666	0.1090	0.0963	0.3167	0.1477
NTOU-Run2	0.9429	0.3380	0.3111	0.9933	0.4738	0.1490	0.1138	0.3633	0.1733
NTOU-Run3	0.9257	0.3500	0.3150	0.9933	0.4783	0.1350	0.0877	0.2767	0.1332
PKU-Run1	0.1486	0.7020	0.5048	0.3533	0.4157	0.6380	0.2000	0.1400	0.1647
PKU-Run2	0.5286	0.5830	0.4061	0.8433	0.5482	0.3760	0.0738	0.1533	0.0996
PKU-Run3	0.3986	0.6780	0.4795	0.8567	0.6149	0.5000	0.1474	0.2633	0.1890
SinicaCKIP-Run1	0.1300	0.8400	0.7174	0.7700	0.7428	0.7730	0.5093	0.5467	0.5273
SinicaCKIP-Run2	0.2257	0.8040	0.6238	0.8733	0.7278	0.7030	0.3833	0.5367	0.4472
SinicaCKIP-Run3	0.1629	0.8420	0.6919	0.8533	**0.7642**	0.7710	0.5000	**0.6167**	0.5523
SinicaIASL-Run1	0.3000	0.7130	0.5161	0.7467	0.6103	0.6050	0.2673	0.3867	0.3161
SinicaIASL-Run2	0.1857	0.7540	0.5873	0.6167	0.6016	0.6860	0.3714	0.3900	0.3805
SinicaSLMP & NTU-Run1	0.4471	0.6540	0.4603	0.8900	0.6068	0.5490	0.2793	0.5400	0.3682
SinicaSLMP & NTU-Run2	0.1414	0.8350	0.7027	0.7800	0.7393	0.7460	0.4354	0.4833	0.4581
SinicaSLMP & NTU-Run3	0.1414	0.8360	0.7036	0.7833	0.7413	0.7490	0.4431	0.4933	0.4669
SJTU-Run1	0.4400	0.6620	0.4671	0.9000	0.6150	0.522	0.2249	0.4333	0.2961
SJTU-Run2	0.0957	0.8560	0.7690	0.7433	0.7559	0.8050	0.5931	0.5733	0.5830
SJTU-Run3	**0.0229**	0.8440	**0.9091**	0.5333	0.6722	0.8090	**0.7102**	0.4167	0.5252
YZU & NCKU-Run1	0.0500	0.7290	0.6500	0.2167	0.3250	0.7050	0.4100	0.1367	0.2050

Table 4. Testing results of error detection subtask

some linguistic resources are used popularly for this bake-off evaluation such as Chinese Gigaword and Sinica Corpus.

5.1 Results of Error Detection

The goals of this subtask are to detect whether a sentence contains errors or not and to identify the locations of the errors in the input sentences. Table 4 shows the testing results of subtask 1. In addition to achieving promising detection effects of error character, reducing the false-alarm rate, which is percentage of the correct sentences that are incorrectly reported containing error characters, is also important. The research teams, NTHU and SJTU, achieved very low false alarm rates, *i.e.*, less than 0.05, while maintaining relatively high detection recall rates, *i.e.*, more than 0.5. These results are what most of the previous studies did not accomplish.

Accuracy is usually adopted to evaluate the performance, but it is affected by the distribution of testing instance. The baseline can be achieved easily by always guessing without errors. That is accuracy of 0.7 in this evaluation. Some systems achieved promising effects of more than 0.8, regardless of detection accuracy or error location accuracy.

Since each participated teams can submit up to three runs, several teams sent different runs that aimed at optimizing the recall or precision rates. These phenomena guide us to adopt F1 score to reflect the tradeoff between precision and recall. In the testing results, SinicaCKIP achieved the best error detection results, if Detection F1 was concerned. NTHU accomplished the best detection effects of indicating error locations, which resulted the best Error Location F1.

In summary, different evaluation metrics were proposed to measure the performance of Chinese spelling checkers. It is difficult to find a perfect system that usually performs better than others, when different metrics are considered. In general, the systems implemented by NTHU, SJTU, and SinicaCKIP relatively outperform the others' developed systems in subtask1 evaluation.

Submission	LA	CA	CR
HLJU-Run1	0.2650	0.2250	0.2432
HLJU-Run2	0.3230	0.2770	0.3081
HLJU-Run3	0.2640	0.2220	0.2403
KUAS & NTNU-Run1	0.4440	0.3940	0.5058
NAIST-Run1	0.5080	0.4670	0.5765
NAIST-Run2	0.2610	0.2540	0.6530
NAIST-Run3	0.4870	0.4530	0.6155
NCTU & NTUT-Run1	0.0700	0.0650	0.5118
NCTU & NTUT-Run2	0.4850	0.4040	0.4040
NCYU-Run1	0.3690	0.3070	0.4850
NCYU-Run2	**0.6630**	**0.6250**	0.7030
NCYU-Run3	**0.6630**	**0.6250**	0.7030
NTHU-Run1	0.4180	0.4090	0.6956
NTHU-Run2	0.4420	0.4310	0.7020
NTHU-Run3	0.4540	0.4430	0.6998
SinicaCKIP-Run1	0.4820	0.4420	0.5854
SinicaCKIP-Run2	0.4990	0.4620	0.5416
SinicaCKIP-Run3	0.5590	0.5160	0.6158
SinicaIASL-Run1	0.4680	0.4290	0.4286
SinicaIASL-Run2	0.4900	0.4480	0.4476
SinicaSLMP & NTU-Run1	0.5070	0.4670	0.4670
SinicaSLMP & NTU-Run2	0.4890	0.4450	0.4450
SinicaSLMP & NTU-Run3	0.4940	0.4500	0.4500
SJTU-Run1	0.3720	0.3380	0.3828
SJTU-Run2	0.4750	0.4420	0.6360
SJTU-Run3	0.3700	0.3560	**0.7050**
YZU & NCKU-Run1	0.1170	0.1090	0.4658

Table 5. Results of error correction subtask

5.2 Results of Error Correction

For subtask 2, the systems need to identify the locations of the errors in the sentences and indicate the corresponding correct characters. Table 5 shows the testing results. For indicating the locations of errors, the research team came from NCYU accomplished the best Location Accuracy. Its achievement of 0.6630 significantly outperformed than the other teams. To further consider correction effects, NCYU also achieved the best Correction Accuracy of 0.6250. However, if the Correction Precision is concerned, the spelling checker developed by SJTU is the best one, which accomplished the effect of 0.7050.

In summary, it is difficult to make the correction on all errors embedded in the input sentences, since there are many sentences that contain more than one error. The achievements of systems implemented by NCYU and SJTU are relatively satisfactory for this subtask.

6 Discussion

The errors observed in everyday writings can be categorized into three different sources. The incorrect words are similar to the correct words either in sound, shape, and/or meaning. Characters of similar pronunciations are the most common source of errors. Characters of similar shapes are not as frequent, but still exist with a significant proportion (Liu et al., 2011).

The most challenging errors to detect and correct are those caused by semantically possible and contextually permissible words. This is a main cause for inter-annotator disagreement in preparing data sets. When a writer wrote "我用槌子處理這一份中藥" (I used a wood hammer to handle this set of Chinese medicine.), a spelling checker cannot tell whether the write might want to use "鎚子" (a metal hammer) or "錘子" (a pendulum) in the place of "槌子" (a wood hammer). As a consequence, it may be difficult for the spelling checker to detect all errors in a text without false alarms. It might be a good strategy to just issue a reminder to the writers these possible alternatives and to ask for confirmations from the writers.

There are confusing word pairs existing in everyday writings, *e.g.*, "紀錄" (record) and "記錄" (record). The basic principle is very clear: the former is a noun and the latter is a verb. However, not all contexts are clear as to which one should be used, e.g., the person who writes down the minutes of a meeting is a "記錄". Other equally confusing word pairs are ["需要" (need, verb), "須要"(need, noun)] and ["計畫" (plan, noun), "計劃"(plan, verb)].

Sometimes the incorrect characters are very competitive for replacing the correct characters due to their similarity at the lexical level, e.g., ["蔓延" (spread), "漫延" (an incorrect spelling of "蔓延")] and ["璀璨" (bright), "璀燦" (an incorrect spelling of "璀璨")]. Some of these incorrect spellings are becoming so popular among the younger generations such that it might be controversial to define "correctness" in the first place, e.g., ["伎倆" (trick), "技倆" (an incorrect spelling of "伎倆")].

7 Conclusions and Future Work

This paper describes the overview of Chinese spelling check evaluation at SIGHAN Bake-off 2013. We introduce the task designing ideas,

data preparation details, evaluation metrics, and the results of performance evaluation.

This bake-off motivates us to build more Chinese language resources for reuse in the future to possibly improve the state-of-the-art techniques for Chinese spelling checking. It also encourages researchers to bravely propose various ideas and implementations for possible breakthrough. No matter how well their implementations would perform, they contribute to the community by enriching the experience that some ideas or approaches are promising or impractical, as verified in this bake-off. Their reports in this proceeding will reveal the details of these various approaches and contribute to our knowledge and experience about Chinese language processing.

We hope our prepared data sets in this bake-off can serve as a benchmark to help developing better Chinese spelling checkers. More data sets that come from different Chinese learners will be investigated in the future to enrich this research topic for natural language processing and computer-aided Chinese language learning.

Acknowledgments

We thank Liang-Pu Chen and Ping-Che Yang, the research engineers of the institute for information industry, Taiwan, for their contribution of students' essays used in this Chinese Spelling Check task. We would like to thank Hsien-You Hsieh, Pei-Kai Liao, Li-Jen Hsu, and Hua-Wei Lin for their hard work to prepare the data sets for this evaluation.

This study was partially supported by the International Research-Intensive Center of Excellence Program of National Taiwan Normal University and National Science Council, Taiwan under grant 101WFA0300229.

References

Chao-huang Chang. 1995. *A New Approach for Automatic Chinese Spelling Correction*. In: Proceedings of Natural Language Processing Pacific Rim Symposium, pp. 278–283.

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*, Int. J. Cont. Engineering Education and Life-Long Learning, vol. 21, no. 1, pp.103-116.

Chuen-Min Huang, Mei-Chen Wu, and Ching-Che Chang. 2007. *Error Detection and Correction Based on Chinese Phonemic Alphabet in Chinese Text*. In: Proceedings of the Fourth Conference on Modeling Decisions for Artificial Intelligence, pp. 463-476.

Chao-Lin Liu, Min-Hua Lai, Kan-Wen Tien, Yi-Hsuan Chuang, Shih-Hung Wu, and Chia-Ying Lee. 2011. *Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications*, ACM Transaction on Asian Language Information Processing, vol. 10, no. 2, Article 10, 39 pages.

Eric Mays, Fred J. Damerau and Robert. L. Mercer. 1991. *Context based spelling correction*. Information Processing and Management, vol. 27, no. 5, pp. 517–522.

Fuji Ren, Hongchi Shi, and Qiang Zhou. 2001. *A hybrid approach to automatic Chinese text checking and error correction*. In: Proceedings of the 2001 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1693-1698.

Lei Zhang, Changning Huang, Ming Zhou, and Haihua Pan. 2000. *Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm*. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp.248–254.