Introduction to SIGHAN 2015 Bake-off for Chinese Spelling Check

Yuen-Hsien Tseng[1], Lung-Hao Lee[1], Li-Ping Chang[2], Hsin-Hsi Chen[3]
[1]Information Technology Center, National Taiwan Normal University
[2]Mandarin Training Center, National Taiwan Normal University
[3]Dept. of Computer Science and Information Engineering, National Taiwan University
samtseng@ntnu.edu.tw, lhlee@ntnu.edu.tw,
lchang@ntnu.edu.tw, hhchen@ntu.edu.tw

Abstract

This paper introduces the SIGHAN 2015 Bake-off for Chinese Spelling Check, including task description, data preparation, performance metrics, and evaluation results. The competition reveals current state-of-the-art NLP techniques in dealing with Chinese spelling checking. All data sets with gold standards and evaluation tool used in this bake-off are publicly available for future research.

1 Introduction

Chinese spelling checkers are relatively difficult to develop, partly because no word delimiters exist among Chinese words and a Chinese word can contain only a single character or multiple characters. Furthermore, there are more than 13 thousand Chinese characters, instead of only 26 letters in English, and each with its own context to constitute a meaningful Chinese word. All these make Chinese spell checking a challengeable task.

An empirical analysis indicated that Chinese spelling errors frequently arise from confusion among multiple-character words, which are phonologically and visually similar, but semantically distinct (Liu et al., 2011). The automatic spelling checker should have both capabilities of identifying the spelling errors and suggesting the correct characters of erroneous usages. The SIGHAN 2013 Bake-off for Chinese Spelling Check was the first campaign to provide data sets as benchmarks for the performance evaluation of Chinese spelling checkers (Wu et al., 2013). The data in SIGHAN 2013 originated from the essays written by native Chinese speakers. Following the experience of the first evaluation, the second bake-off was held in CIPS-SIGHAN Joint CLP-2014 conference, which focuses on the essays written by learners of Chinese as a Foreign Language (CFL) (Yu et al., 2014).

Due to the greater challenge in detecting and correcting spelling errors in CFL leaners' written essays, SIGHAN 2015 Bake-off, again features a Chinese Spelling Check task, providing an evaluation platform for the development and implementation of automatic Chinese spelling checkers. Given a passage composed of several sentences, the checker is expected to identify all possible spelling errors, highlight their locations, and suggest possible corrections.

The rest of this article is organized as follows. Section 2 provides an overview of the SIGHAN 2015 Bake-off for Chinese Spelling Check. Section 3 introduces the developed data sets. Section 4 proposes the evaluation metrics. Section 5 compares results from the various contestants. Finally, we conclude this paper with findings and offer future research directions in Section 6.

2 Task Description

The goal of this task is to evaluate the capability of a Chinese spelling checker. A passage consisting of several sentences with/without spelling errors is given as the input. The checker should return the locations of incorrect characters and suggest the correct characters. Each character or punctuation mark occupies 1 spot for counting location. The input instance is given a unique passage number *pid*. If the sentence contains no spelling errors, the checker should return "pid, 0". If an input passage contains at least one spelling error, the output format is "pid [, location, correction]+", where the symbol "+" indicates there is one or more instance of the predicted element "[, location, correction]". "Location" and "correction", respectively, denote the location of incorrect character and its correct version. Examples are given as follows.

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 32–37,
Beijing, China, July 30-31, 2015.

- Example 1
  Input: (pid=A2-0047-1) 我真的洗碗我可以去看你
  Output: A2-0047-1, 4, 希, 5, 望
- Example 2
  Input: (pid=B2-1670-2) 在日本，大學生打工的情況是相當普偏的。
  Output: B2-1670-2, 17, 遍
- Example 3
  Input: (pid=B2-1903-7) 我也是你的朋友，我會永遠在你身邊。
  Output: B2-1903-7, 0

There are 2 wrong characters in Ex. 1, and correct characters "希," and "望" should be used in locations 4, and 5, respectively. In Ex. 2, the 17th character "偏" is wrong, and should be "遍". Location "0" denotes that there is no spelling error in Ex. 3

3 Data Preparation

The learner corpus used in our task was collected from the essay section of the computer-based Test of Chinese as a Foreign Language (TOCFL), administered in Taiwan. The spelling errors were manually annotated by trained native Chinese speakers, who also provided corrections corresponding to each error. The essays were then split into three sets as follows

(1) Training Set: this set included 970 selected essays with a total of 3,143 spelling errors. Each essay is represented in SGML format shown in Fig. 1. The title attribute is used to describe the essay topic. Each passage is composed of several sentences, and each passage contains at least one spelling error, and the data indicates both the error's location and corresponding correction. All essays in this set are used to train the developed spelling checker.

(2) Dryrun Set: a total of 39 passages were given to participants to familiarize themselves with the final testing process. Each participant can submit several runs generated using different models with different parameter settings of their checkers. In addition to make sure that the submitted results can be correctly evaluated, participants can fine-tune their developed models in the dryrun phase. The purpose of dryrun is to validate the submitted output format only, and no dryrun outcomes were considered in the official evaluation

(3) Test Set: this set consists of 1,100 testing passages. Half of these passages contained no spelling errors, while the other half included at least one spelling error. The evaluation was conducted as an open test. In addition to the data sets provided, registered participant teams were allowed to employ any linguistic and computational resources to detect and correct spelling errors. Besides, passages written by CFL learners may yield grammatical errors, missing or redundant words, poor word selection, or word ordering problems. The task in question focuses exclusively on spelling error correction.

```
<ESSAY title="學中文的第一天">
<TEXT>
<PASSAGE id="A2-0521-1"> 這位小姐說：你應該一直走到十只路口，再右磚一直走經過一家銀行就到了。</PASSAGE>
<PASSAGE id="A2-0521-2">應為今天是第一天，老師先請學生自己給介紹。</PASSAGE>
</TEXT>
<MISTAKE id="A2-0521-1" location="15">
<WRONG>十只路口</WRONG>
<CORRECTION>十字路口</CORRECTION>
</MISTAKE>
<MISTAKE id="A2-0521-1" location="21">
<WRONG>右磚</WRONG>
<CORRECTION>右轉</CORRECTION>
</MISTAKE>
<MISTAKE id="A2-0521-2" location="1">
<WRONG>應為</WRONG>
<CORRECTION>因為</CORRECTION>
</MISTAKE>
</ESSAY>
```

Figure 1. An essay represented in SGML format

4 Performance Metrics

Table 1 shows the confusion matrix used for performance evaluation. In the matrix, TP (True Positive) is the number of passages with spelling errors that are correctly identified by the spelling checker; FP (False Positive) is the number of passages in which non-existent errors are identified; TN (True Negative) is the number of passages without spelling errors which are correctly identified as such; FN (False Negative) is the number of passages with spelling errors for which no errors are detected.

The criteria for judging correctness are determined at two levels as follows.

(1) Detection level: all locations of incorrect characters in a given passage should be completely identical with the gold standard.

(2) Correction level: all locations and corresponding corrections of incorrect characters should be completely identical with the gold standard.

In addition to achieve satisfactory detection/correction performance, reducing the false positive rate, that is the mistaken identification of errors where none exist, is also important (Wu et al., 2010). The following metrics are measured at both levels with the help of the confusion matrix.

- False Positive Rate (FPR) = FP / (FP+TN)
- Accuracy = (TP+TN) / (TP+FP+TN+FN)
- Precision = TP / (TP+FP)
- Recall = TP / (TP+FN)
- F1= 2 *Precision*Recall/(Precision+Recall)

Confusion Matrix		System Result	
		Positive (Erroneous)	Negative (Correct)
Gold Standard	Positive	**TP**	**FN**
	Negative	**FP**	**TN**

Table 1. Confusion matrix for evaluation.

For example, if 5 testing inputs with gold standards are "A2-0092-2, 0", "A2-0243-1, 3, 健, 4, 康", "B2-1923-2, 8, 誤, 41, 情", "B2-2731-1, 0", and "B2-3754-3, 10, 觀", and the system outputs the result as "A2-0092-2, 5, 玩", "A2-0243-1, 3, 件, 4, 康", "B2-1923-2, 8, 誤, 41, 情", "B2-2731-1, 0", and "B2-3754-3, 11, 觀", the evaluation tool will yield the following performance:

- False Positive Rate (FPR) = 0.5 (=1/2) Notes: {"A2-0092-2, 5"}/{"A2-0092-2, 0", "B2-2731-1, 0"}
- Detection-level
  - Accuracy =0.6 (=3/5)

    Notes: {"A2-0243-1, 3, 4", "B2-1923-2, 8, 41", "B2-2731-1, 0"} / {"A2-0092-2, 5", A2-0243-1, 3, 4", "B2-1923-2, 8, 41", "B2-2731-1, 0", "B2-3754-3, 11"}
  - Precision = 0.5 (=2/4)

    Notes: {"A2-0243-1, 3, 4", "B2-1923-2, 8, 41"} / {"A2-0092-2, 5", A2-0243-1, 3, 4", "B2-1923-2, 8, 41", "B2-3754-3, 11"}
  - Recall = 0.67 (=2/3).

    Notes: {"A2-0243-1, 3, 4", "B2-1923-2, 8, 41"} / {A2-0243-1, 3, 4", "B2-1923-2, 8, 41", "B2-3754-3, 10"}
  - F1=0.57 (=2*0.5*0.67/(0.5+0.67))
- Correction-level
  - Accuracy =0.4 (=2/5)

    Notes: {"B2-1923-2, 8, 誤, 41, 情", "B2-2731-1, 0"} / {"A2-0092-2, 5, 玩", "A2-0243-1, 3, 件, 4, 康", "B2-1923-2, 8, 誤, 41, 情", "B2-2731-1, 0", "B2-3754-3, 11, 觀"}
  - Precision = 0.25 (=1/4)

    Notes: {"B2-1923-2, 8, 誤, 41, 情"} / {"A2-0092-2, 5, 玩", "A2-0243-1, 3, 件, 4, 康", "B2-1923-2, 8, 誤, 41, 情", "B2-3754-3, 11, 觀"}
  - Recall = 0.33 (=1/3)

    Notes: {"B2-1923-2, 8, 誤, 41, 情"} / {, "A2-0243-1, 3, 健, 4, 康", "B2-1923-2, 8, 誤, 41, 情", "B2-3754-3, 10, 觀"}
  - F1=0.28 (=2*0.25*0.33/(0.25+0.33))

5 Evaluation Results

Table 2 summarizes the submission statistics for 9 participant teams including 4 from universities and research institutions in China (CAS, ECNU, SCAU, and WHU), 4 from Taiwan (KUAS, NCTU & NTUT, NCYU, and NTOU), and one private firm (Lingage). Among 9 registered teams, 6 teams submitted their testing results. In formal testing phase, each participant can submit at most three runs that adopt different models or parameter settings. In total, we received 15 runs.

Table 3 shows the task testing results. The research team NCTU&NTUT achieved the lowest false positive rate at 0.0509. For the detection-level evaluations, according to the test data distribution, a baseline system can achieve an accuracy level of 0.5 by always reporting all testing cases as correct without errors. The system result submitted by CAS achieved promising performance exceeding 0.7. We used the F1 score to reflect the tradeoff between precision and recall. As shown in the testing results, CAS provided the best error detection results, achieving a high F1 score of 0.6404. For correction-level evaluations, the correction accuracy provided by the CAS system (0.6918) significantly outperformed the other teams. Besides, in terms of correction precision and recall, the spelling checker developed by CAS also outperforms the others, which in turn has the highest F1 score of 0.6254. Note

that it is difficult to correct all spelling errors found in the input passages, since some sentences contain multiple errors and only correcting some of them are regarded as a wrong case in our evaluation.

Table 4 summarizes the participants' developed approaches and the usages of linguistic resources. Among 6 teams that submitted the official testing results, NCYU did not submit the report of its developed method. None of the submitted systems provided superior performance in all metrics, though those submitted by CAS and NCTU&NTUT provided relatively best overall performance when different metric is considered. The CAS team proposes a unified framework for Chinese spelling correction. They used HMM-based approach to segment sentences and generate correction candidates. Then, a two-stage filter process is applied to re-ranking the candidates for choosing the most promising candidates. The NCTU&NTUT team proposes a word vector/conditional random field based spelling error detector. They utilize the error detection results to guide and speed up the time-consuming language model rescoring procedure. By this way, potential Chinese spelling errors could be detected and corrected in a modified sentence with the maximum language model score.

Participant (Ordered by abbreviations of names)	**#Runs**
Chinese Academy of Sciences (**CAS**)	3
East China Normal University (**ECNU**)	0
National Kaohsiung University of Applied Sciences (**KUAS**)	3
Lingage Inc. (**Lingage**)	0
National Chiao Tung University & National Taipei University of Technology (**NCTU & NTUT**)	3
National Chiayi University (**NCYU**)	1
National Taiwan Ocean University (**NTOU**)	2
South China Agriculture University (**SCAU**)	3
Wuhan University (**WHU**)	0
Total	**15**

Table 2. Submission statistics for all participants

Submission	**FPR**	**Detection-Level**				**Correction-Level**			
		Acc.	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**
CAS-Run1	0.1164	0.6891	0.8095	0.4945	0.614	0.68	**0.8037**	0.4764	0.5982
CAS-Run2	0.1309	**0.7009**	0.8027	0.5327	**0.6404**	**0.6918**	0.7972	**0.5145**	**0.6254**
CAS-Run3	0.2036	0.6655	0.7241	**0.5345**	0.6151	0.6491	0.7113	0.5018	0.5885
KUAS-Run1	0.2327	0.5009	0.5019	0.2345	0.3197	0.4836	0.4622	0.2	0.2792
KUAS-Run2	0.2091	0.5164	0.5363	0.2418	0.3333	0.4982	0.4956	0.2055	0.2905
KUAS-Run3	0.1818	0.5318	0.5745	0.2455	0.3439	0.5145	0.537	0.2109	0.3029
NCTU&NTUT-Run1	**0.0509**	0.6055	**0.8372**	0.2618	0.3989	0.5782	0.8028	0.2073	0.3295
NCTU&NTUT-Run2	0.0655	0.6091	0.8125	0.2836	0.4205	0.5809	0.7764	0.2273	0.3516
NCTU&NTUT-Run3	0.1327	0.6018	0.7171	0.3364	0.4579	0.5645	0.6636	0.2618	0.3755
NCYU-Run1	0.1182	0.5245	0.586	0.1673	0.2603	0.5091	0.5357	0.1364	0.2174
NTOU-Run1	0.0909	0.5445	0.6644	0.18	0.2833	0.5327	0.6324	0.1564	0.2507
NTOU-Run2	0.5727	0.4227	0.422	0.4182	0.4201	0.39	0.3811	0.3527	0.3664
SCAU-Run1	0.5327	0.3409	0.2871	0.2145	0.2456	0.3218	0.2487	0.1764	0.2064
SCAU-Run2	0.1218	0.5464	0.6378	0.2145	0.3211	0.5227	0.5786	0.1673	0.2595
SCAU-Run3	0.6218	0.3282	0.3091	0.2782	0.2928	0.3018	0.2661	0.2255	0.2441

Table 3. Testing results of our Chinese spelling check task.

Participant	Approaches	Linguistic Resources
CAS	• Candidate Generation • Candidate Re-ranking • Global Decision Making	• SIGHAN-2013 CSC Datasets • CLP-2014 CSC Datasets • SIGHAN-2015 CSC Training Data • Taiwan Web Pages as Corpus • Chinese Words and Idioms Dictionary • Pinyin and Cangjie Code Table • Web-based Resources
KUAS	• Rules-based Method • Linear Regression Model	• Chinese Orthographic Database
NCTU & NTUT	• Misspelling Correction Rules • CRF-based Parser • Word Vector/CRF-based Spelling Error Detector • Trigram Language Model	• CLP-2014 CSC Datasets • SIGHAN-2015 CSC Training Data • Sinica Corpus
NTOU	• N-gram Model • Rule-based Classifier	• SIGHAN 2013 CSC Datasets • CLP-2014 CSC Datasets • Showen Jiezi and the Four-Corner Encoding • Sinica Corpus • Google N-gram Corpus
SCAU	• Bi-gram Language Model • Tri-gram Language Model	• SIGHAN-2013 CSC Datasets • CLP-2014 CSC Datasets • CCL • SOGOU

Table 4. A summary of participants' developed systems

6 Conclusions and Future Work

This paper provides an overview of SIGHAN 2015 Bake-off for Chinese spelling check, including task design, data preparation, evaluation metrics, performance evaluation results and the approaches used by the participant teams. Regardless of actual performance, all submissions contribute to the knowledge in search for an effective Chinese spell checker, and the individual reports in the Bake-off proceedings provide useful insight into Chinese language processing.

We hope the data sets collected for this Bake-off can facilitate and expedite future development of effective Chinese spelling checkers. Therefore, all data sets with gold standards and evaluation tool are made publicly available at http://ir.itc.ntnu.edu.tw/lre/sighan8csc.html.

The future direction focuses on the development of Chinese grammatical error correction. We plan to build new language resources to help improve existing techniques for computer-aided Chinese language learning. In addition, new data sets obtained from CFL learners will be investigated for the future enrichment of this research topic.

Acknowledgments

We thank all the participants for taking part in our task. We would like to thank Bo-Shun Liao for developing the evaluation tool.

This research is partially supported by the "Aim for the Top University Project" and "Center of Learning Technology for Chinese" of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, R.O.C. and is also sponsored in part by the "International Research-Intensive Center of Excellence Program" of NTNU and Ministry of Science and Technology, Taiwan, R.O.C. under the Grant no. MOST 104-2911-I-003-301, MOST 102-2221-E-002-103-MY3, and MOST 103-2221-E-003-013-MY3.

References

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*, 10(2), Article 10, 39 pages.

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. *Processing of the 1st CIPS-SIGHAN Joint Conference on Chinese Language Processing* (*CLP-10*), pages 54-61.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese spelling check evaluation at SIGHAN Bake-off 2013. *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing* (*SIGHAN-13*), pages 35-42.

Lian-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2014. Overview of SIGHAN 2014 Bake-off for Chinese spelling check. *Processing of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing* (*CLP-14*), pages 126-132.