Sinica-IASL Chinese Spelling Check System at SIGHAN-7

Ting-Hao Yang*
Institute of Information Systems and Applications
National Tsing-Hua University
tinghaoyang@iis.sinica.edu.tw

Yu-Lun Hsieh*
Institute of Information Science
Academia Sinica
morphe@iis.sinica.edu.tw

Yu-Hsuan Chen
Institute of Information Science
Academia Sinica
smallright@iis.sinica.edu.tw

Michael Tsang
Electrical Engineering and Computer Sciences
University of California, Berkeley
themichaeltsang@gmail.com

Cheng-Wei Shih
Institute of Information Science
Academia Sinica
dapi@iis.sinica.edu.tw

Wen-Lian Hsu
Institute of Information Science
Academia Sinica
hsu@iis.sinica.edu.tw

Abstract

We developed a Chinese spelling check system for error detection and error correction subtasks in the 2013 SIGHAN-7 Chinese Spelling Check Bake-off. By using the resources of Chinese phonology and orthographic components, our system contains four parts: high confidence pattern matcher, the detection module, the correction module, and the merger. We submitted 2 official runs for both subtasks. The evaluation result show that our system achieved 0.6016 in error detection F-score of subtask 1, and 0.448 in correction accuracy of subtask 2.

1 Introduction

Chinese spelling check is a task which detects and corrects errors in text. These errors may result from writing, optical character recognition (OCR), typing, and so on. Chinese spelling check has been considered useful in many area such as language learning or error-tolerated language processing, and there are many researches around this topic (Y.-Z. Chen, Wu, Yang, Ku, & Chen, 2011; Liu et al., 2011; Wu, Chen, Yang, Ku, & Liu, 2010).

The SIGHAN Bake-off 2013 Chinese Spelling Check contains two subtasks. The first subtask requires each team to detect whether a sentence contains errors. If the answer is yes, the error location(s) should be provided. For each sentence in subtask2, there is at least one error. Participants have to locate and correct those errors in the sentence.

The organization of this paper is as follows. Section 2 describes the architecture and different modules in our spelling check system. Section 3 shows our evaluation results and some discussion. Lastly, Section 4 concludes this work and shares some insights we gained participating this Bake-off.

2 Method

Our system can be divided into four parts. They are high confidence pattern matcher, detection module, correction module and merger. High confidence pattern matcher finds patterns that are very unlikely to contain any error, and exclude them from the rest of the process. Detection module is used to detect the error locations in a sentence. Correction module generates suggestions for erroneous words. Merger receives these suggestions and chooses the most possible result. Figure 1 shows the structure of our system.

* Authors with equal contributions.

Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), pages 93–96,
Nagoya, Japan, 14 October 2013.

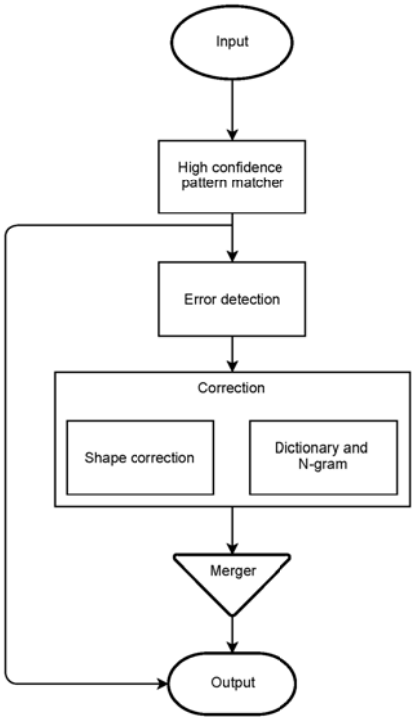

Figure 1. Structure of the Sinica-IASL Spelling Check System

2.1 High Confidence Pattern Matchers

Reliable Phonological Sequence Matcher

There are many homophones in Chinese. In order to detect errors caused by such a phonological similarity, we prepared an in-house dictionary with words which are longer than 2 characters, converted these words to phonetic symbols (注音) to form a syllable-word mapping table. Then we apply every syllable sequences in a syllable-annotated corpus based on CIRB (K.-H. Chen) to compute the matching percentage of each syllable-word pair in the mapping table. Syllable-word pairs with high percentage are considered as high confidence syllable-to-word patterns, means that these syllable sequences most likely map to the corresponding words in the corpus.

In the reliable phonological sequence matching of our system, we first convert the input sentence to phonetic symbols. If the phonetic symbols match one of the high confidence syllable-word pairs, the module checks the difference between the mapping word and the original input sequence. The overlapped characters are marked as correct. The remaining characters in the original input sentence are marked as error candidates and the correcting suggestions will be deliver to the correction module. For example, we found the syllable sequence of an input sentence "銼折" match a high confidence phonetic-word pair "ㄘㄨㄛˋ ㄓㄜˊ" — "挫折" (setback). Then the overlapped character "折" is marked as correct. On the other hand, the character "銼" will be regarded as an error candidate and the correcting suggestion of "挫" will be preserved.

Reliable Long Words

This module handles errors that happened in long Chinese in-vocabulary words based on the idea of maximum matching. We collected Chinese words, idioms, and sayings which are longer than four words, such as "一毛不拔" (stingy). Any part of the input text that exactly matches the patterns in this list are considered reliable, thus are marked as error-free.

Frequent Errors

We collected frequent errata and misused words from a dataset of junior high school students' composition. For example, "一但" is a frequent error from the correct one "一旦" (once). Whenever this module finds a part of the text contained in this frequent error list, a suggestion will be generated based on this list.

2.2 Detection Module

Detection module locates possible errors by integrating information from high confidence pattern matchers in Section 2.1 and word segmentation result described below, and passes these error locations to the correction module.

Word Segmentation

We used CKIP Chinese word segmentation tool (Ma & Chen, 2003) to get segmented sentences. Our presumption is that words containing erroneous characters are more likely to be split into different segments. For example, "佈告欄" (bulletin board) would be tagged as one segment by the tool, while the erroneous case "怖告欄" would be split into singlets such as "怖", "告" and "欄". This module would then check for consecutive singles and try to merge them into one segment. Then those segments were verified by a two-step checking. The first one is using a dictionary (Ministry of Education, 1994) to ensure there is no out-of-vocabulary being generated. The second one is using the frequency of n-grams from Google web 1T. The frequency of the generated segment has to surpass the pre-set threshold. Only those suggestions that pass at least one of the checks are kept.

2.3 Correction Module

Possible error positions from detection modules are received by the following correction modules to generate candidates for corrections. Both similar pronunciation and shape correcting process will be activated, and the results will be sent to the merger for the final decision.

Homophone Dictionary and N-gram Correction

We check the received error locations and generate possible corrections by using homophones and Google web 1T n-gram frequency. For example, there is an error "書貴" and the detection modules say that "貴" is an error. This module will generate possible candidates by finding all homophones of "貴". The frequency of each candidate in Google web 1T n-gram is used as the confidence. In this case, the frequency of "書櫃" (bookcase) is higher than the frequency of the original text, and all other homophones. Thus, a correction for "書貴" is given by this module as "書櫃".

Errors with Similar Shape

Shape correction module utilized data from Xiaoxuetang Chinese character database (National-Taiwan-University & Academia-Sinica, 2013), which consists of decomposed components of almost every Chinese character, to find corrections with similar shapes. We retrieved the components of each character that were marked as a possible error by the detection module, and calculate the Damerau-Levenshtein edit distance (Damerau, 1964; Levenshtein, 1966) between this character and all other characters. We slightly altered this edit distance formula to favor those with identical parts regardless of the order. For example, a character with parts (A, B) are considered more similar to (B, A) than to (A, D). From our observation of the training data, this method can better rank the most similar characters. We then select those characters that have an edit distance score less than 1, and filter out the ones that do not form a word with its neighboring 1 to 3 characters using a dictionary (Ministry of Education, 1994).

Across-the-board Search and Correction

This process will only be activated when no answer was provided by any previous modules. It checks all locations which are not covered by high confidence pattern matcher, and generates suggestions that have similar shapes to the characters in these locations using the shape correction module. We do not consider phonetic errors in this step because we assume phonetic errors can be detected by previous modules.

	Run 1	Run 2	Best	Average
False-Alarm Rate	0.3	**0.1857**	0.0229	0.4471
Detection Accuracy	0.713	**0.754**	0.861	0.654
Detection Precision	0.5161	**0.5873**	0.9091	0.4603
Detection Recall	0.7467	0.6167	1	0.89
Detection F-Score	0.6103	0.6016	0.7642	0.6068
Error Location Accuracy	0.605	**0.686**	0.82	0.549
Error Location Precision	0.2673	**0.3714**	0.7102	0.2793
Error Location Recall	0.3867	0.39	0.6167	0.54
Error Location F-Score	0.3161	**0.3805**	0.5854	0.3682

Table 1. Evaluation Results of Subtask 1

	Run 1	Run 2	Best	Average
Location Accuracy	0.468	**0.49**	0.663	0.418
Correction Accuracy	0.429	**0.448**	0.625	0.409
Correction Precision	0.4286	0.4476	0.705	0.6956

Table 2. Evaluation Results of Subtask 2.

2.4 Merger

The merger receives all suggestions from the aforementioned correction modules, and decides whether a suggestion is accepted or not. In our system, we used a probabilistic language model trained by LDC news corpus as the kernel of this merger. This module generates possible combinations of suggestions and calculates scores. The combination of suggestions with the best score is selected as our answer.

3 Experimental Results

We submitted two runs to compare the effect of high confidence patterns. Run 1 used patterns which have a confidence level of 50% or higher, and run 2 used those having over 80%. Table 1

and 2 are our experimental results for subtask 1 and 2, respectively. Bold typed numbers indicate that our performance is above the average.

We can see that, generally speaking, our performance of both subtasks is above average among participants. The effect of the confidence level of our high confidence patterns can be observed when we compare the results of our 2 runs. Using a higher confidence threshold (run 2) would yield a higher accuracy, while a lower threshold (run 1) would sometimes yield a higher recall.

4 Conclusion

This paper introduced our Sinica-IASL Chinese spelling checking system, implemented for the 2013 SIGHAN-7 Bake-off. By using phonological and orthographical data of Chinese characters, dictionaries and frequent error data, we were able to achieve reasonable performances. During the process of our work, we noticed that about 80% of the texts are covered by all words in our dictionary. The minimum coverage of a sentence is 50%. It implies that we can handle at least 50% of the text by only using a dictionary. If we use frequent n-grams, the coverage is over 90%. A method for finding useful n-grams is a way to boost our performance. The experimental results showed that there is plenty of room for improvement in our system's ability to detect errors. Further works also include using a web corpus to find frequent errors, possible error locations and corrections. In conclusion, our system can benefit from more resources in order to become a more competitive Chinese spelling checker.

Acknowledgement

We would like to thank the reviewers of the SIGHAN 2013 Program Committee for their helpful suggestions and comments on this paper, and Dr. Chiang-Chi Liao for his assistance on this work. This research was supported by the National Science Council of Taiwan under Grants NSC101-3113-P-032-001.

References

Kuang-Hua Chen. Chinese Information Retrieval Benchmark. http://lips.lis.ntu.edu.tw/cirb/index.htm

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.* Paper presented at the Engineering Education and Life-Long Learning.

Fred J. Damerau. (1964). *A technique for computer detection and correction of spelling errors.* Paper presented at the Communications of the ACM 7.3.

Vladimir I. Levenshtein. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady., 10*.

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 Transactions on Asian Language Information Processing, 10*, 1-39.

Wei-Yun Ma and Keh-Jiann Chen. (2003). *Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff.* Paper presented at the The Second SIGHAN Workshop on Chinese Language Processing.

R.O.C. Ministry of Education. (1994). *教育部重編國語辭典修訂本 Revised Chinese Dictionary*.

National-Taiwan-University and Academia-Sinica (Producer). (2013). 小學堂文字學資料庫 Xiaoxuetang Philology Database. Retrieved from http://xiaoxue.iis.sinica.edu.tw/

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.* Paper presented at the CIPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing.