Introduction to CKIP Chinese Spelling Check System for SIGHAN Bakeoff 2013 Evaluation

Yu-Ming Hsieh[1,2] Ming-Hong Bai[1,2] Keh-Jiann Chen[1]
[1] Institute of Information Science, Academia Sinica, Taiwan
[2] Department of Computer Science, National Tsing-Hua University, Taiwan
morris@iis.sinica.edu.tw, mhbai@sinica.edu.tw,
kchen@iis.sinica.edu.tw

Abstract

In order to accomplish the tasks of identifying incorrect characters and error correction, we developed two error detection systems with different dictionaries. First system, called CKIP-WS, adopted the CKIP word segmentation system which based on CKIP dictionary as its core detection procedure; another system, called G1-WS, used Google 1T uni-gram data to extract pairs of potential error word and correction candidates as dictionary. Both detection systems use the confusion character set provided by the bakeoff organizer to reduce the suggested correction candidates. A simple maximizing tri-gram frequency model based on Google 1T tri-gram was designed to validate and select the correct answers. The CKIP group of Academia Sinica participated in both Sub-Task1 (Error Detection) and Sub-Task2 (Error Correction) in 2013 SIGHAN bakeoff. The evaluation results show that the performances of our systems are pretty good on both tasks.

1 Introduction

Spelling check, an automatic mechanism to detect and correct document inputting errors, is a common task for every written languages. How to detect and correct error spellings in a document is an important and difficult task in particular for Chinese language. Since many Chinese characters have similar shape and similar pronunciation, improper use of characters in Chinese essays are hard to be detected (Liu et. al, 2011). Therefore, most Chinese character detection systems are built based on confusion sets and a language model. Some new systems also incorporate NLP technologies for Chinese character error detection in recent years (Huang et al., 2007; Wu et al., 2010). Huang et al. (2007) used a new word detection function in the CKIP word segmentation toolkit (Ma and Chen, 2003) to detect error candidates. With the help of a dictionary and confusion set, the system will be able to judge whether a monosyllabic word is probably error or not. The system we designed for this contest adopts CKIP word segmentation module for unknown word detection too, confusion sets for providing possible candidate characters, and a large-scale corpus for constructing language model for validation and correction of words.

In order to accomplish these two spelling check tasks, we designed two error detection systems with the capability of providing suggested correction candidates. Each system uses different dictionary for its knowledge source. The first system uses the CKIP dictionary, called CKIP-WS; another uses the correction pair dictionary extracted from Google 1T uni-gram data, called G1-WS. In CKIP-WS, we detect possible occurrences of errors through unknown word detection process (Chen and Bai, 1998). So that deeper morphological analysis is carried out only where morphemes of unknown word are detected (Chen and Ma, 2002). As a result, some false alarms caused by proper names and determinant-measure compounds can be avoided. For G1-WS, we build an error suggestion dictionary (or template) to match potential error spellings and suggest correction candidates. Finally we use an n-gram language model to select the corrected characters as our system output.

Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), pages 59–63,
Nagoya, Japan, 14 October 2013.

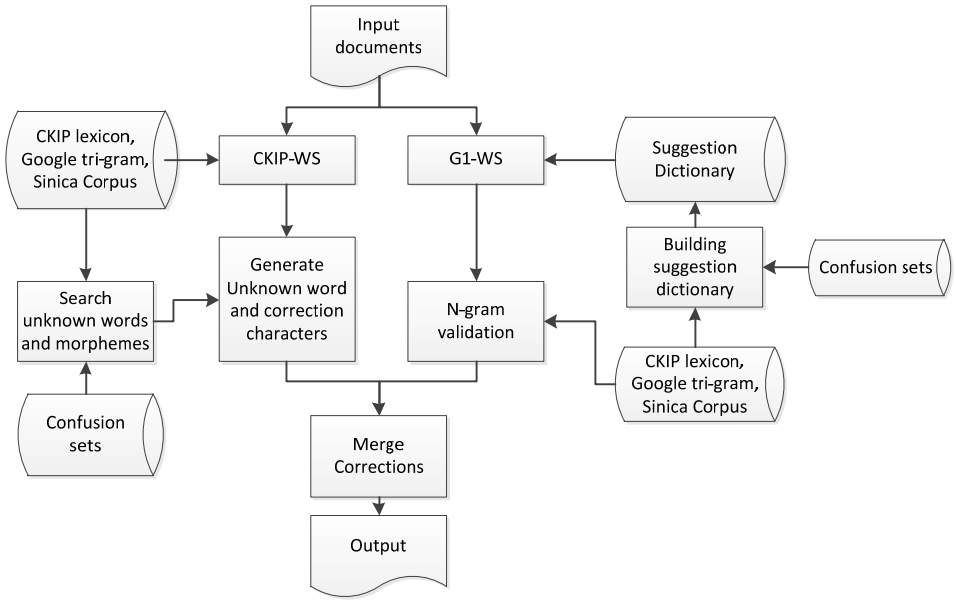

Figure 1. Flowchart of the system

The paper is organized as follows. Section 2 describes the architecture of our system. Section 3 states the bakeoff results evaluated by SIGHAN. In the section 4, we have some relevant discussions and provide analysis on the system performances. Section 5 is the conclusion.

2 System architecture

2.1 System flowchart

Figure 1 illustrates the block diagram of our Chinese Spelling Check system used in this contest. First, input documents are sent to two different error detection systems. The first one is CKIP-WS, which can detect error characters based on unknown word detection and n-gram verification. The second system is G1-WS, which treats error detection based on suggestion dictionary produced by using data of confusion sets, Sinica Corpus and Google Chinese 1T. Finally, the results of the two systems can be merged to get a final detection result. The details will be described in the following subsections.

2.2 Unknown word detection

The first step of our system is word segmentation to find possible error candidates. For example the input sentence “不怕措折地奮鬥” will be marked as “不() 怕() 措(?) 折(?) 地() 奮鬥()” by the unknown word detection process of the CKIP-WS, where (?) denotes the detected monosyllabic unknown word morpheme and () denotes common words. We focus on the morphemes marked with (?) only and provide possible replacement words by checking confusion sets and CKIP dictionary. After the process, the pattern “不 怕 {措,挫} 折 地 奮 鬥” is extracted. For another example, “也在一夕之門”. After the detection process, the system marks the sentence as “也() 在() 一() 夕(?) 之() 門()”. We use simple algorithm to produce “也 在 一 夕 之 {門, 間}” by left- or right- extension of the word by checking CKIP dictionary. To increase the recall rate, if there are still some monosyllabic words which are not stop words, those words will be also considered as possible error candidates. We will mark those problematic morphemes with (?) for further n-gram validation.

2.3 Building suggestion dictionary

In G1-WS, we first build a suggestion dictionary for potential error words. The data of the dictionary is extracted from Google 1T uni-gram. We use this uni-gram data, and the confusion set to search for similar word pairs and ranks the pair of words by their frequencies. The word of low frequency is considered as error candidate and the high frequency similar word is considered as correction suggestion. Note that the above process is based on the fact that Google 1T uni-gram contains many spelling-error words. Some extracted similar word pairs are shown follow:

Word	Suggestion
措折	挫折
讚同	贊同
讚助商	贊助商
…	

However, the extracted naive suggestion dictionary may have a lot of noises. So we use a simple method to confirm whether to adopt each similar word pair suggestions. First, we use word segmentation in Sinica Corpus by G1-WS. And then we count all words and suggestions. If the frequency ratio of freq(word)/freq(suggestion) > 0.1, we ignore this suggestion. The final G1-WS error detection and candidate suggestion process adopts the modified dictionary. After the first step CKIP-WS error detection, we use the new error detection system G1-WS with this suggestion dictionary to detect and provide additional correction suggestions.

2.4 Validation and correction by n-gram model

After two error detection steps an input document is marked with potential errors and suggest candidate characters. We were intended to develop a character n-gram language model to determine the best character sequence as the answers for detection and correction. However due to the limited developing time, we simply developed a maximizing tri-gram frequency approach instead. Based on the marked error spots, we set a window to count the frequency of these strings which contain potential errors. By simply maximizing tri-gram frequency based on Google 1T tri-gram data, we select the suggestion candidates with the highest string frequency as the answer.

For example, in "也 在 一 夕 之 {門,間}", in comparing with other string candidates as shown in Figure 2. We found the string of the highest frequency "在一夕之間" which is 37,709. So we detect the error spot and select '間' as the corrected character at the mean time.

```
L2L1C0: ("也在一夕之門", 0)
L1C0R1: ("在一夕之門，", 0)
C0R1R2: ("一夕之門，被", 0)
L1C0:   ("在一夕之門", 0)
C0R1:   ("一夕之門，", 0)
------------------------------
L2L1C0: ("也在一夕之間", 0)
L1C0R1: ("在一夕之間，", 0)
C0R1R2: ("一夕之間，被", 0)
L1C0:   ("在一夕之間", 37709)
C0R1:   ("一夕之間，", 0)
```

Figure 2. Calculating the frequency of the target string in Google tri-gram corpus.

3 Evaluation Results

3.1 Data

The resources adopted in our system are described below:

- CKIP lexicon[1]: The CKIP lexicon is an electronic dictionary containing 88,000 entries for Mandarin Chinese. We use this word information for checking whether the target lexicon is a word or not.
- Google 1T n-gram lexicon[2]: It consists of Chinese word n-grams and their frequency counts generated from over 800 million tokens of text. The length of the n-grams ranges from unigrams (single words) to 5-grams. We use tri-gram data for our n-gram validation process and use uni-gram data for building the suggestion dictionary.
- Confusion sets: Confusion sets are a collection of each individual Chinese character (Liu et al., 2011). There were 5401 confusion sets for each of the 5401 high frequency characters. We use this data to generate possible correction characters.
- Sinica Corpus[3]: We employ the ten-million-word Sinica Corpus, a balanced modern Chinese Corpus with word segmentation and PoS tag. We use this corpus to check and filter our correction data.

3.2 Evaluation metrics

There are several evaluation indexes provided by SIGHAN, i.e. false-alarm rate (FAR), detection accuracy (DA), detection precision (DP), detection recall (DR), detection F-score (DF), error location accuracy (ELA), error location precision (ELP), error location recall (ELR), error location F-score (ELF), location accuracy (LA), correction accuracy (CA), and correction precision (CP).

3.3 Results of our CKIP-WS system

Table 1 shows the evaluation results of our CKIP-WS system in error detection and error correction tasks. In SIGHAN evaluation report, the CKIP-WS system is 'SinicaCKIP-Run1'. In

[1] http://www.aclclp.org.tw/use_ced.php

[2] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2010T06

[3] http://db1x.sinica.edu.tw/kiwi/mkiwi/

both tasks, our system achieves good performance.

Task1	**FAR**			
	0.13			
	DA	**DP**	**DR**	**DF**
	0.84	0.7174	0.77	0.7428
	ELA	**ELR**	**ELP**	**ELF**
	0.773	0.5093	0.5467	0.5273
Task2	**LA**	**CA**	**CP**	**CF**
	0.482	0.442	0.5854	0.5037

Table 1. Results of our CKIP-WS system

3.4 Results of our final system

In our final system, we merge CKIP-WS and G1-WS output into final correction data. The evaluations of our final system are shown in table 2. For sub-task 1, FAR score rises 0.03, from 0.13 to 0.1619, and DF and ELF improve 0.0214 and 0.025 respectively. For sub-task 2, the CF has improved 0.0578. From these results, we know that the two systems of CKIP-WS and G1-WS have a complementary relationship. With a better suggestion dictionary, the system performance will be better.

Task1	**FAR**			
	0.1619			
	DA	**DP**	**DR**	**DF**
	0.842	0.6919	0.8533	0.7642
	ELA	**ELR**	**ELP**	**ELF**
	0.771	0.8533	0.6167	0.5523
Task2	**LA**	**CA**	**CP**	**CF**
	0.559	0.516	0.6158	0.5615

Table 2. Results of our final system

From the final summary of SIGHAN Bake-off, our final system ranks the top among 33 submitted systems for detection F-score (DF) and rank 3rd for error location F-score (ELF) in sub-task 1. For sub-task 2, our system ranks second among 30 submitted systems.

4 Discussions

The evaluation results show that our system arrives the top three in both Sub-Task 1 and Sub-Task 2. However, our system performance is still low in both recall and precision. Following are discussions on the recall and precision problems for our systems. We have observes some reasons accounted for recall problems:

- Some correct characters are not in the confusion sets, for examples, “不怕[固→苦]難”, “有特[絑→殊]的意義”, “深深地敬[佩→佩]這”, and etc.
- Dispute on the gold standard, for examples, “樹木 [經]不起 大雨的打擊”, “有時候同學的 [嘻]笑怒罵”, “不要 一時[胡]塗”.
- The word pairs as (再,在),(得,的) cannot be distinguished in our system, such as, “是 個 [在] 平凡 不過 的”, “覺 [的] 很 不 開心”, “都 過 的 很 快樂”, “從此變[的]不同”, and etc.
- No information on the related words, such as “圈差” (correct suggestion “圈叉”), and “二 連 罷” (correct suggestion “二連霸”).

As to the precision problem, we focus on the confusion set and n-gram language model:

- There are a lot of irrelevant characters in the confusion sets. There should be a way to filter out some of the irrelevant characters.
- A better n-gram language model needs to be developed.

The above discussions suggest that we should enrich our knowledge bases to increase the recall rate by including more suggestion candidates and on the other hand to design a more robust language model to increase the precision of the correction.

5 Conclusions

In this paper, we described the overview of our Chinese Spelling Check system for SIGHAN-7 bakeoff. We employ two word segmentation systems, and adopt some knowledge resources. With the help of these resources, we propose a method to select and filter these correction candidates. Finally, we merge these two systems’ outputs for SIGHAN evaluation. The evaluation results show that our approaches are promising. In the future, we will be trying to merge the two word segmentation to a uniform system and develop a more robust language model.

References

Keh-Jiann Chen and Ming-Hong Bai. 1998. Unknown Word Detection for Chinese by a Corpus-based Learning Method. International Journal of Computational Linguistics and Chinese Language Processing, 3(1):27-44.

Keh-Jiann Chen and Wei-Yun Ma. 2002. Unknown word extraction for Chinese documents. In proceedings of the 19th international conference on Computational linguistics (COLING 2002), pages 1-7.

Chuen-Ming Huang, Mei-Che 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 (MDAIIV), pages 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 Transactions on Asian Language Information Processing, 10(2), 10:1-39. Association for Computing Machinery, USA, June 2011.

Wei-Yun Ma and Keh-Jiann Chen. 2003. Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff. In Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pages 168-171.

Shih-Hung Wu, Yong-Zhi Chen, Ping-Che Yang, Tsun Ku, and Chao-lin Liu. 2010. Reducing the False Alarm Rate of Chinese Character Error Detection and Correction. In Proceeding of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2010), pages 54-61, Beijing, 28-29 Aug., 2010.