NTOU Chinese Spelling Check System in SIGHAN-8 Bake-off

Wei-Cheng Chu and Chuan-Jie Lin
Department of Computer Science and Engineering
National Taiwan Ocean University
No 2, Pei-Ning Road, Keelung 202, Taiwan R.O.C.
{wcchu.cse, cjlin}@ntou.edu.tw

Abstract

This paper describes details of NTOU Chinese spelling check system in SIGHAN-8 Bakeoff. Besides the basic architecture of the previous system participating in last two CSC tasks, three new preference rules were proposed to deal with Simplified Chinese characters, variants, sentence-final particles, and DE-particles. A new sentence likelihood function was proposed based on frequencies of space-removed version of Google *n*-gram datasets. Two formal runs were submitted where the best one was created by the system using Google *n*-gram frequency information.

1 Introduction

Automatic spell checking is a basic and important technique in building NLP systems. It has been studied since 1960s as Blair (1960) and Damerau (1964) made the first attempt to solve the spelling error problem in English. Spelling errors in English can be grouped into two classes: non-word spelling errors and real-word spelling errors.

A non-word spelling error occurs when the written string cannot be found in a dictionary, such as in *fly fron* Paris*. The typical approach is finding a list of candidates from a large dictionary by edit distance or phonetic similarity (Mitten, 1996; Deorowicz and Ciura, 2005; Carlson and Fette, 2007; Chen *et al.*, 2007; Mitten 2008; Whitelaw *et al.*, 2009).

A real-word spelling error occurs when one word is mistakenly used for another word, such as in *fly form* Paris*. Typical approaches include using confusion set (Golding and Roth, 1999; Carlson *et al.*, 2001), contextual information (Verberne, 2002; Islam and Inkpen, 2009), and others (Pirinen and Linden, 2010; Amorim and Zampieri, 2013).

Spelling error problem in Chinese is quite different. Because there is no word delimiter in a Chinese sentence and almost every Chinese character can be considered as a one-character word, most of the errors are real-word errors.

On the other hand, there is also an *illegal-character error* where a hand-written symbol is not a legal Chinese character (thus not collected in a dictionary). Such an error cannot happen in a digital document because all characters in Chinese character sets such as BIG5 or Unicode are legal.

There have been many attempts to solve the spelling error problem in Chinese (Chang, 1994; Zhang *et al*., 2000; Cucerzan and Brill, 2004; Li *et al*., 2006; Liu *et al*., 2008). Among them, lists of visually and phonologically similar characters play an important role in Chinese spelling check (Liu *et al*., 2011).

This bake-off is the third Chinese spelling check evaluation project. A CSC system will be evaluated in two levels: error detection and error correction. The task is organized based on some research works (Wu *et al*., 2010; Chen *et al*., 2011; Liu *et al*., 2011; Yu *et al*., 2014).

2 NTOU CSC System Description

This year, the architecture of NTOU CSC system mostly follows the previous version, only that three new preference rules are added. The architecture of previous NTOU CSC system is explained as follows.

Figure 1 shows the architecture of NTOU Chinese spelling checking system. A sentence under consideration is first word-segmented. New sentences are generated by replacing candidates of spelling errors with their similar characters one at a time. New sentences are also word-segmented. Their likelihoods of being

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 137–143,
Beijing, China, July 30-31, 2015.

acceptable Chinese sentences are measured by sorted by *n*-gram linguistic model. If the new sentence with the top-1 likelihood is better than the original sentence, a spelling error is reported.

There are 6 kinds of **confusion sets** used in this system. One of them was generated from the Four-Corner Code system, proposed by us in CSC 2014 (Chu and Lin, 2014). The other 5 were provided by the organizers in CSC 2013 (Wu *et al.*, 2013). They are characters with the same sound in the same tone, characters with the same sound in different tones, characters with similar sound in the same tone, characters with similar sound in different tones, and visually similar characters.

There are three cases of spelling error candidates in our system. Two of them have been described in our CSC 2014 system description paper. Multi-word replacement will be explained in Section 3.1.

One-character word replacement: every one-character word in the original sentence is considered as a spelling error candidate and should be replaced with its similar characters in its confusion set. For example, "座" in Topic A2-0101-2 is a one-character word and its similar characters are 柞坐雁挫..., the replacement is as follows.

A2-0101-2, Original:
所以我們沒位子可以座
Replaced:
所以我們沒位子可以柞
所以我們沒位子可以坐
所以我們沒位子可以雁
所以我們沒位子可以挫
...

Multi-character word replacement: the method to create multi-character word confusion sets has been proposed by Lin and Chu (2015). Given a multi-character word, if one of the characters is replaced with a similar character and becomes another legal word, these two words are considered as collected into each other's multi-character word confusion set. The resource to create our word confusion set is the Revised Mandarin Dictionary by the Ministry of Education[1].

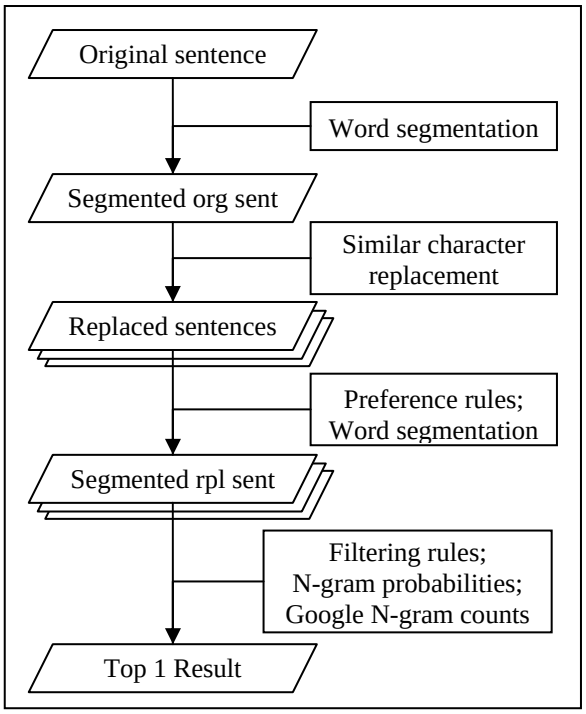

Figure 1. Architecture of NTOU Chinese Spelling Check System

Every multi-character word in the original sentence is considered as a spelling error candidate and should be replaced with its similar words. For example, "不過" and "漢子" in Topic A2-1308-1 are multi-character words. Their similar words are "補過", "不果"…, "蚶子", "漢字"... The replacement is as follows.

A2-1308-1, Original:
不過一個漢子也看不懂
Replaced:
補過一個漢子也看不懂
不果一個漢子也看不懂
不過一個蚶子也看不懂
不過一個漢字也看不懂
...

Two **filtering rules** are again adopted this year.

Rule-1 No error in personal names: discard a replacement if it becomes a personal name; it is unlikely to see errors in personal names. Take C1-1701-2 in the CLP Bakeoff 2014 CSC test set as an example as an example. When the one-character word "位" is replaced by its similar character "魏", "魏產齡" is identified as a personal name, so this replacement is discarded.

[1] http://dict.revised.moe.edu.tw/

C1-1701-2, Original segmented:
每 位 產 齡 婦女
Replaced *and discarded*:
每 魏產齡(PERSON) 婦女

Rule-2 Stopword filtering: discard a replacement if the original character is a personal anaphora (你'you', 我'I', 他她它祂牠'he/she/it') or numbers from 1 to 10 (一二三四五六七八九十).

N-gram linguistic models, word-unigram, word-bigram, and POS-bigram models, were trained by using a large Chinese corpus, Academic Sinica Balanced Corpus (Chen *et al*., 1996).

N-gram preference score is defined as $[P(S_{new}) / P(S_{org}) - 1]$, where $P(S)$ is the probability of a sentence S in a language model. When sorting, word-bigram preference score has the higher priority, word-unigram preference score has the second priority, and POS-bigram preference score has the lowest priority.

If the top-1 sentence is a newly generated sentence, and all of its preference scores are not lower than predefined thresholds, report it as an error with the location of the replacement. Otherwise, report "no error". The threshold of word-bigram preference score is 0.0571, and 0.0171 for word-unigram, 0 for POS-bigram preference scores.

3 New Features in 2015

3.1 Multi-word replacement

In our observation, a spelling error occurs in at least three different cases. The first case is that the error alone is identified as a one-character word. The second case is that one character in a multi-character word is misused but the wrong word is still a legal word. The third case is that the erroneous character, combining with the character to its left or to its right, is misidentified as a multi-character word. Take Topic 00043 in the SIGHAN7 Bakeoff 2013 CSC Datasets as an example. The error "帶" occurs in a multi-character word "膠帶", but the correct word "塑膠袋" is a longer word.

Topic 00043, Original:
外面也會包塑膠帶啦
Segmented:
外面 也 會 包 塑 膠帶 啦
Correct:
外面 也 會 包 塑膠袋 啦

To deal with such an error case, we proposed a new replacement procedure: if a multi-character word is preceded or followed by a one-character word, each character in this multi-character word is substituted with its similar characters one by one. Again, take Topic 00043 as an example. "外面" and "膠帶" are multi-character words and adjacent to one-character words, so they are candidates of spelling errors. By replacing similar characters of "外", "面", "膠", and "帶", newly generated sentences are as follows.

Topic 00043, Segmented:
外面 也 會 包 塑 膠帶 啦
Replaced:
舀面也會包塑膠帶啦
外麵也會包塑膠帶啦
外面也會包塑穋帶啦
外面也會包塑膠袋啦
...

3.2 Preference rules

Three kinds of preference rules were proposed this year to deal with special cases: Simplified Chinese characters or variants, sentence-final particles, and DE-particles. If any of the rules are matched, an error is reported immediately.

Rule 1: Simplified and variant Chinese character detection

Because the sentences in the datasets are written in Traditional Chinese, all Simplified Chinese characters or variants of Traditional Chinese characters appearing in the datasets are marked as errors.

A mapping table (Lin *et al*., 2012) from variants (including Simplified Chinese characters) to their corresponding Traditional Chinese characters is adopted to correct such a kind of errors.

Take B1-0840-2 in the CLP Bakeoff 2014 CSC Datasets as an example of Simplified Chinese character replacement, where "尔" is a Simplified Chinese character and should be replaced with its corresponding Traditional Chinese character "爾" directly.

B1-0840-2, Original:
首尔是韓國的首都
Correct:
首爾是韓國的首都

Take B1-3981-1 in the CLP Bakeoff 2014 CSC Datasets as an example of variant replacement, where "僔" is a variant of the more-common Traditional Chinese character "咱", so it should be replaced directly.

B1-3981-1, Original:
然後僔倆就一塊兒出去打球
Correct:
然後咱倆就一塊兒出去打球

Rule 2: Sentence-final particle detection

In our observation, some sentence-final particles were frequently misspelled in the datasets, including "嗎", "吧", and "啊". We collected the errors in the dataset whose corrections were these particles and created the following three replacement rules:

1. If a sentence ends with a one-character word "碼" or "馬", it should be replaced with "嗎".
2. If a sentence ends with a one-character word "把" or "巴", it should be replaced with "吧".
3. If a sentence ends with a one-character word "阿", it should be replaced with "啊".

The following examples show the application of these rules.

B1-0381-2, Original:
你喜歡西式的餐廳馬？
Correct:
你喜歡西式的餐廳嗎？

B1-1125-4, Original:
應該沒有問題把？
Correct:
應該沒有問題吧？

B1-1589-1, Original:
像討論活動啊，遊戲阿，
Correct:
像討論活動啊，遊戲啊，

Rule 3: DE-particle detection

In Chinese, "的", "得", and "地" serve as function words in various different cases. They are grouped together and receive a special POS "DE". However, despite their usages are different, they are easily messed up with one another, even for native speakers.

Patterns	Correction
得/地 Na	的
得/地 PERIODCATEGORY	的
VC 的/地 VC	得
VA 的/地 VH	得
VCL 的/地 VH	得
VH 的/得 VE	地

Table1. Replacement Rules for DE-particles

To deal with such kind of errors, we extracted most frequently-seen POS patterns in the training set. Table 1 lists the 6 patterns learned and used in our system. To demonstrate how to apply these rules, take B1-0184-3 in the CLP Bakeoff 2014 CSC Datasets as an example. The DE-particle "得" is followed by a common noun (whose POS is "Na") and matched the first DE-particle replacement rule in Table 1, so it is replaced with "的".

B1-0184-3, Original:
我 得 英文(Na) 那麼 好
Correct:
我的英文那麼好

3.3 Google N-gram Scoring Functions

As described in Section 2, our previous language models were trained by Academia Sinica Balanced Corpus. We found that the volume and vocabulary of ASBC was not large enough. So we turn to use Chinese Web 5-gram dataset [2] instead. Several *n*-gram scoring functions have been proposed by Lin and Chu (2015). Some examples from the Chinese Web 5-gram dataset are given here:

Unigram:	稀釋劑	17260
Bigram:	蒸發量 超過	69
Trigram:	能量 遠 低於	113
4-gram:	張貼 色情 圖片 或	73
5-gram:	幸好 我們 發現 得 早	155

Moreover, in order to avoid interference of word segmentation errors, we further design some likelihood scoring functions which utilize substring frequencies instead of word *n*-gram frequencies.

By removing space between *n*-grams in the Chinese Web 5-gram dataset, we constructed a new dataset containing identical substrings with

[2] https://catalog.ldc.upenn.edu/LDC2010T06

Run	FPAlarm	Accuracy	Precision	Recall	F1
Formalrun1_NTOU	9.09	**54.45**	**66.44**	18.00	28.33
Formalrun2_NTOU	**57.27**	42.27	42.20	**41.82**	**42.01**

Table 2: Formal Run Performance in Error-Detection Level

Run	FPAlarm	Accuracy	Precision	Recall	F1
Formalrun1_NTOU	9.09	**53.27**	**63.24**	15.64	25.07
Formalrun2_NTOU	**57.27**	39.00	38.11	**35.27**	**36.64**

Table 3: Formal Run Performance in Error-Correction Level

their web frequencies. For instances, n-grams in the previous example will become:

Zhar=3: 稀釋劑	17260
Zhar=5: 蒸發量超過	69
Zhar=5: 能量遠低於	113
Zhar=7: 張貼色情圖片或	73
Zhar=8: 幸好我們發現得早	155

where *Zhar*(*S*) is defined as the number of Chinese or other characters in a sentence *S*. Note that if two different n-gram sets become the same after removing the space, their will merge into one entry with the summation of their frequencies. Simplified Chinese words were translated into Traditional Chinese in advanced.

Given a sentence *S*, let *SubStr*(*S*, *n*) be the set of all substrings in *S* whose *Zhar* values are *n*. We define ***Google string frequency*** *gsf*(*u*) of a string *u* to be its frequency data provided in the modified Chinese Web 5-gram dataset. If a string does not appear in that dataset, its *gsf* value is defined to be 0.

Equation 1 give the definition of ***averaged weighted log frequency score*** $GS_{wgt}(S)$ which sums up the logarithm of frequencies of all substrings with length *n*, averages scores at the same *n* level, and multiplies log*n*.

$$GS_{wgt}(S) = \sum_{n=2}^{12}\left(\frac{\log n}{|SubStr(S,n)|}\times\sum_{u\in SubStr(S,n)}\log(gsf(u))\right) \quad \text{Eq. 1}$$

Now the ***Google n-gram preference score*** is defined as Eq 2.

$$GS_{prf}(S_{new},S_{org}) = \frac{GS_{wgt}(S_{new})}{GS_{wgt}(S_{old})} - 1 \quad \text{Eq. 2}$$

As the same algorithm of error detection as described in Section 2, a top-1 replacement should have a Google *n*-gram preference score no lower than the threshold 0.0002 so that it could be reported as an error correction.

4 Experimental Results

We submitted 2 formal runs this year by two different statistics-based systems. The first system checks the word-unigram, word-bigram, and POS-bigram preference scores of the top-1 sentence to decide the occurrence of a spelling error, as described in Section 2. The second system uses Google *n*-gram preference scores instead to check the occurrence of a spelling error, as described in Section 3.3.

Table 2 and 3 illustrate the evaluation results of formal runs. As we can see, the first system guesses errors more correctly but too cautiously. The second system, on the other hand, proposed more errors so it achieved a higher recall rate and a higher F-score.

5 Conclusion

It is our third time to participate in a Chinese spelling check evaluation project. Based on our previous CSC system, we further proposed three preference rules to handle three special cases: (1) Simplified Chinese characters or variants; (2) sentence-final particles, and (3) DE-particles. Moreover, a new sentence-likelihood scoring function, ***averaged weighted log frequency score***, was proposed which used Google *n*-gram frequency information.

Two formal runs were submitted this year. The first one was predicted by three *n*-gram language models trained by a large corpus ASBC. The second one was predicted by the system which used Google *n*-gram averaged weighted log frequency scores to decide the occurrence of errors. The evaluation results show the system using Google *n*-gram frequency information outperformed the traditional language models.

References

R.C. de Amorim and M. Zampieri. 2013. “Effective Spell Checking Methods Using Clustering

Algorithms," *Recent Advances in Natural Language Processing*, 7-13.

C. Blair. 1960. "A program for correcting spelling errors," *Information and Control*, 3:60-67.

A. Carlson, J. Rosen, and D. Roth. 2001. "Scaling up context-sensitive text correction," *Proceedings of the 13th Innovative Applications of Artificial Intelligence Conference*, 45-50.

A. Carlson and I. Fette. 2007. "Memory-Based Context-Sensitive Spelling Correction at Web Scale," *Proceedings of the 6th International Conference on Machine Learning and Applications*, 166-171.

C.C. Chang and C.J. Lin. 2011. "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, 2:27:1-27.

C.H. Chang. 1994. "A pilot study on automatic chinese spelling error correction," *Journal of Chinese Language and Computing*, 4:143-149.

K.J. Chen, C.R. Huang, L.P. Chang, and H.L. Hsu. 1996. "Sinica corpus: Design methodology for balanced corpora," *Language, Information and Computation* (*PACLIC 11*), 167-176.

Q. Chen, M. Li, and M. Zhou. 2007. "Improving Query Spelling Correction Using Web Search Results", *Proceedings of the 2007 Conference on Empirical Methods in Natural Language* (*EMNLP-2007*), 181-189.

Y.Z. Chen, S.H. Wu, P.C. Yang, T. Ku, and G.D. 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*, 21(1):103-116.

W.C. Chu and C.J. Lin. 2014. "NTOU Chinese Spelling Check System in CLP Bake-off 2014," *Proceedings of The 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 210-215.

S. Cucerzan and E. Brill. 2004. "Spelling correction as an iterative process that exploits the collective knowledge of web users," *Proceedings of EMNLP*, 293-300.

F. Damerau. 1964. "A technique for computer detection and correction of spelling errors." *Communications of the ACM*, 7:171-176.

S. Deorowicz and M.G. Ciura. 2005. "Correcting Spelling Errors by Modelling Their Causes," *International Journal of Applied Mathematics and Computer Science*, 15(2):275-285.

A. Golding and D. Roth. 1999. "A winnow-based approach to context-sensitive spelling correction," *Machine Learning*, 34(1-3):107-130.

A. Islam and D. Inkpen. 2009. "Real-word spelling correction using googleweb 1t 3-grams," *Proceedings of Empirical Methods in Natural Language Processing* (*EMNLP-2009*), 1241-1249.

M. Li, Y. Zhang, M.H. Zhu, and M. Zhou. 2006. "Exploring distributional similarity based models for query spelling correction," *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 1025-1032.

C.J. Lin, J.C. Zhan, Y.H. Chen, and C.W. Pao. 2012. "Strategies of Processing Japanese Names and Character Variants in Traditional Chinese Text," *International Journal of Computational Linguistics & Chinese Language Processing*, 17(3), 87-108.

Chuan-Jie Lin and Wei-Cheng Chu. 2015. "A Study on Chinese Spelling Check Using Confusion Sets and N-gram Statistics," *International Journal of Computational Linguistics and Chinese Language Processing*, to be appeared.

W. Liu, B. Allison, and L. Guthrie. 2008. "Professor or screaming beast? Detecting words misuse in Chinese," *The 6th edition of the Language Resources and Evaluation Conference*.

C.L. Liu, M.H. Lai, K.W. Tien, Y.H. Chuang, S.H. Wu, and C.Y. 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.

R. Mitton. 1996. *English Spelling and the Computer*, Harlow, Essex: Longman Group.

R. Mitton. 2008. "Ordering the Suggestions of a Spellchecker Without Using Context," *Natural Language Engineering*, 15(2):173-192.

T. Pirinen and K. Linden. 2010. "Creating and weighting hunspell dictionaries as finite-state automata," *Investigationes Linguisticae*, 21.

S. Verberne. 2002. Context-sensitive spell checking based on word trigram probabilities, Master thesis, University of Nijmegen.

C. Whitelaw, B. Hutchinson, G.Y. Chung, and G. Ellis. 2009. "Using the Web for Language Independent Spellchecking and Autocorrection," *Proceedings Of Conference On Empirical Methods In Natural Language Processing* (*EMNLP-2009*), 890-899.

S.H. Wu, Y.Z. Chen, P.C. Yang, T. Ku, and C.L. Liu. 2010. "Reducing the False Alarm Rate of Chinese Character Error Detection and Correction," *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing* (*CLP 2010*), 54-61.

S.H. Wu, C.L. Liu, and L.H. Lee. 2013. "Chinese Spelling Check Evaluation at SIGHAN Bake-off

2013," *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing* (*SIGHAN'13*), 35-42.

L.C. Yu, L.H. Lee, Y.H. Tseng, and H.H. Chen. 2014. "Overview of SIGHAN 2014 Bake-off for Chinese Spelling Check," *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing* (*CLP'14*), 126-132.

L. Zhang, M. Zhou, C.N. Huang, and H.H. Pan. 2000. "Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm," *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 248-254.