Candidate Scoring Using Web-Based Measure for Chinese Spelling Error Correction

Liang-Chih Yu
Yuan Ze University
135 Yuan-Tung Road, Chung-Li City
Taoyuan County, Taiwan, 32003
lcyu@saturn.yzu.edu.tw

Chao-Hong Liu **Chung-Hsien Wu**
National Cheng Kung University
No. 1, University Road
Tainan City, Taiwan, 70101
chl,chwu@csie.ncku.edu.tw

Abstract

Chinese character correction involves two major steps: 1) Providing candidate corrections for all or partially identified characters in a sentence, and 2) Scoring all altered sentences and identifying which is the best corrected sentence. In this paper a web-based measure is used to score candidate sentences, in which there exists one continuous error character in a sentence in almost all sentences in the Bakeoff corpora. The approach of using a web-based measure can be applied directly to sentences with multiple error characters, either consecutive or not, and is not optimized for one-character error correction of Chinese sentences. The results show that the approach achieved a fair precision score whereas the recall is low compared to results reported in this Bakeoff.

1 Introduction

Errors existing in Chinese sentences can be classified into five categories: 1) Deletion, 2) Insertion, 3) Substitution, 4) Word-Order and 5) Non-Word errors (C.-H. Liu, Wu, & Harris, 2008; C.-L. Liu, Lai, Chuang, & Lee, 2010; C.-L. Liu, Tien, Lai, Chuang, & Wu, 2009; C.-H. Wu, Liu, Harris, & Yu, 2010). Deletion errors occur when there are missing Chinese characters/words in a sentence; Insertion errors occur when there are grammatically redundant characters/words; Substitution errors occur when characters/words are mis-typed by similar, either visually or phonologically, ones; Word-Order errors occur when the word order of a sentence does not conform to the language, which is a common error type exists in writings of second-language learners; Non-Word errors occur when a Chinese character is written incorrectly by hand, e.g., miss of a stroke.

Of the five error types, the Substitution errors is addressed in this SIGHAN-7 Chinese Spelling Check bakeoff and might be referred to as "Chinese spelling error" to emphasize its resemblance to counterparts in spelling-based languages such as English. It should be noted that Non-Word errors is also a kind of Chinese spelling errors. It is also a common error type in hand-writings of second-language learners. However, since it only exists in hand-writings of humans and because all characters used in computers are legal ones, it is not necessary to address this kind of spelling errors when given erroneous texts are of electronic forms.

The task addressed in SIGHAN-7 is a restricted type of Substitution errors, where there exists at most one continuous error (mis-spelled) character in its context within a sentence, with only one exception in which there is a two-character error (Chen, Wu, Yang, & Ku, 2011; C.-L. Liu et al., 2010; S.-H. Wu, Chen, Yang, Ku, & Liu, 2010). This allows the system to assume that when a character is to be corrected, its adjacent characters are correct. The correction procedure is comprised of two consecutive steps: 1) Providing candidate corrections for each character in the sentence, and 2) Scoring the altered correction sentences and identifying which is the best corrected sentence (C.-H. Liu et al., 2008; C.-H. Wu et al., 2010). In this paper, a web-based measure is employed in the second step to score and identify the best correction sentence (Macias, Wong, Thangarajah, & Cavedon, 2012).

This paper is organized as follows. Section 2 describes the system architecture for spelling error correction. Section 3 provides the details

Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), pages 108–112,
Nagoya, Japan, 14 October 2013.

of the model using web-based measure to score candidate corrections. In Section 4 the experimental setup and results are detailed. The last section summarized the conclusions and future work of this paper.

2 System Overview

SIGHAN-7 bakeoff is comprised of two subtasks, 1) Error Detection and 2) Error correction. Each of the sub-tasks requires the system to report positions where the errors occur. The philosophy behind the separation of the two subtasks lies in the belief that it is easier to detect if there is an error than to locate that error and provide correction to it.

In this paper, we took a different philosophy to address spelling error correction problem, in which there is no separate error detection method to detect if there is an error character or where the error is in a sentence. In this paper there is the one error correction method for both subtasks. In our system, if a character is reported erroneous, there is always a correction to that character; the correction method itself serves as an error detection mechanism.

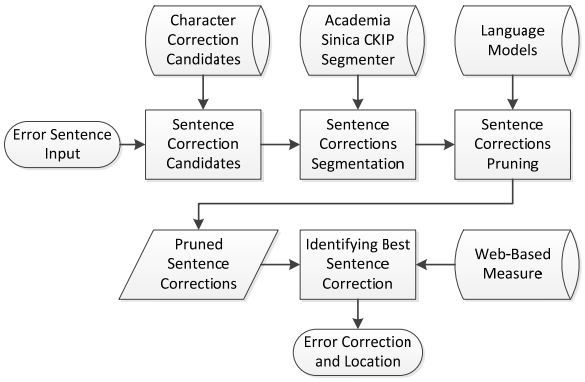

Fig. 1. System overview of the proposed spelling error correction using web-based measure.

The overview of our system is shown in Fig. 1. The input of the system is a sentence in which an erroneous character might occur. To recover the possible error, each character in the sentence is assumed to be an error one, and is given replacements (which are possible corrections to a character) using visually and phonologically similar sets provided by SIGHAN-7 bakeoff. For a character s_i in a sentence $S = (s_1, s_2, \dots, s_n)$ of n characters, there will be m possible corrections $S_i^1, S_i^2, \dots, S_i^m$ and the best correction sentence $\hat{S}_i$, concerning s_i, can be derived using equation 1.

$$\hat{S}_i = \underset{j=1,\dots,m}{\operatorname{argmax}} LM\left(Seg\left(S_i^j\right)\right) \quad (1)$$

where $Seg(S_e)$ is the function returning Chinese segmentation results of sentence S_e, and $LM(S_g)$ returns the language model score of a segmented sentence S_g.

Therefore, sentence correction candidates $\hat{S}_1, \hat{S}_2, \dots, \hat{S}_n$ are derived, corresponding to the best correction characters, $\hat{s}_1, \hat{s}_2, \dots, \hat{s}_n$, respectively. Finally, Equation 2 is used to determine which candidate is the best correction, $\hat{S}$.

$$\hat{S} = \underset{i=1,\dots,n}{\operatorname{argmax}} R\left(\hat{S}_i, \hat{s}_i\right) \quad (2)$$

where $R(S_c, s_c)$ returns the relatedness between a correction sentence S_c and its corresponding correction character s_c (Macias et al., 2012). The description of R is presented in Section 3.

In the proposed system, if the derived correction, $\hat{S}$, is identical to input sentence, S, it reports that there is no error in the sentence. On the contrary, if $\hat{S}$ is not identical to S, which indicates there is one character difference, the system then reports the sentence is detected erroneous along with the resulting correction character. Therefore there is no independent error detection module or procedure in our system; error detection itself depends on if the resulting corrections are identical to input sentences.

3 Web-Based Measure

There are two major directions to improve error correction system, 1) Finding correct and concise candidate sets for erroneous texts, and 2) Using measures such as language model scores to determine which correction sentence is the best result (C.-H. Liu et al., 2008; C.-H. Wu et al., 2010). In both directions, measures used to prune out unlikely candidates and determine the best correction are the fundamental technique. In SIGHAN-7 bakeoff, the visually and phonologically similar characters are provided as correction candidates. Therefore the focus of the proposed system lies in the second direction, i.e., to provide a measure that will rank correct candidates higher against other candidates.

To provide information about which of the candidates is a better correction, language models and pointwise mutual information (PMI) are commonly used (Chen et al., 2011; C.-L. Liu et al., 2010; C.-H. Wu et al., 2010). Although the information is usually trained with a large corpus

such as Chinese Gigaword, they are still insufficient in general-domain applications.

To overcome this data insufficiency problem, web-based measures for estimating distances/similarities and relatedness have been proposed as alternative metric for several NLP applications (Cilibrasi & Vitanyi, 2009; Cilibrasi & Vitanyi, 2007; Gracia & Mena, 2008; Lovelyn Rose & Chandran, 2012). In this paper, we modified a web-based definition of semantic relatedness metric proposed by (Macias et al., 2012). Equation 2 is re-written as Equation 3.

$$\hat{S} = \underset{i=1,\dots,n}{\operatorname{argmax}} R(\hat{S}_i, \hat{s}_i) = \underset{i=1,\dots,n}{\operatorname{argmax}} \frac{\sum_{\forall k \in \hat{S}_i} W(k, \hat{s}_i)}{|\hat{S}_i|} \quad (3)$$

where W is the "normalized web relatedness" and k is a comprising character in the sentence correction candidate $\hat{S}_i$. $|\hat{S}_i|$ indicates the number of characters in $\hat{S}_i$. The definition of W is provded in Equation 4.

$$W(k, s) = e^{-0.6 \times D(k,s)} \quad (4)$$

where D is the "normalized web distance" and is defined in Equation 5.

$$D(k, s) = \frac{\log(\max(|k|, |s|)) - \log(|k \cap s|)}{\log(|G|) - \log(\min(|k|, |s|))} \quad (5)$$

where $|G|$ is the number of Wikipedia Chinese pages, which is 3,063,936 as of the time the system is implemented.

It should be noted that Macias-Galindo et al.'s original work is used in English texts. Currently we have not administered any preliminary experiment to find better setups of these equations.

4 Experiments and Discussions

In the proposed system, Academia Sinica's CKIP Chinese Segmenter is used to derive segmentation results (Ma & Chen, 2003) and the language model (trigrams using Chen and Goodman's modified Kneser-Ney discounting) is trained using SRILM with Chinese Gigaword (LDC *Catalog No.*: LDC2003T09) (Stolcke, 2002).

In a brief summary of the results, our system did not perform well in the final test of SIGHAN-7 bakeoff. The authors would like to defend the proposed method with a major problem in the runtime of the final test. In theory, the $\hat{S}_1, \hat{S}_2, \dots, \hat{S}_n$ as derived in Equation 1 should all be estimated using the web-based measure using Equation 2. However, since the number of sentences in the final test is huge (Sub-Tasks 1 and 2 each has 1,000 paragraphs and each paragraph contains about five Chinese sentences), the enormous number of queries sent to the search engine (Yahoo!) has caused our experiments being banned for several times. To solve this problem, two strategies were used to complete the final test, 1) only three of the candidates $\hat{S}_1, \hat{S}_2, \dots, \hat{S}_n$ (ranked the highest three using n-gram) are considered in the final test using web-based measure, and 2) three computers with different physical IP addresses were setup for the experiment. Therefore, the potential of the proposed method is far from fully exploited. A post-workshop experiment will be administered for further analysis of the method.

Table 1. Comparisons on Error Location Accuracy in SIGHAN-7 Sub-Tasks 1 and 2.

Sub-Task 1 (Detection)	**Error Location Accuracy**
NCKU&YZU-1	0.705
NTHU-3	**0.820**
SinicaCKIP-3	0.771
SJTU-3	0.809
NCYU-2	0.652
NCYU-3	0.748
Sub-Task 2 (Correction)	**Location Accuracy**
NCKU&YZU-1	0.117
NTHU-3	0.454
SinicaCKIP-3	0.559
SJTU-3	0.370
NCYU-2	**0.663**
NCYU-3	**0.663**

The comparisons of the proposed system and highly ranked systems in SIGHAN-7 are excerpted in this section. The first result that attracts our attention is error location accuracy as shown in Table 1. This is a common measure in both Sub-Tasks and is defined as "number of sentences error locations are correctly detected" over "number of all test sentences". The report of our system (NCKU&YZU-1) on error location accuracy in Sub-Task 1 (Detection) is 0.705, whereas it is only 0.117 in Sub-Task 2 (Correction). This result puzzled the authors because in our system, there is no error detection module. Similar results on both Sub-Tasks are expected since the same error correction method is used. A possible explanation is that the final test corpora of the two Sub-Tasks exhibited substantial differences in the composition of correct and erroneous sentences or in sentential characteristics. The results of other systems reported in both

Sub-Tasks seem to support this point of view. However, further analysis on the test corpora is still needed to clarify this problem.

Table 2. Comparisons on Error Location measures in SIGHAN-7 Sub-Task 1.

Error Location (Detection)	Accuracy	Precision	Recall
NCKU&YZU-1	0.705	0.410	0.137
NTHU-3	**0.820**	0.670	0.520
SinicaCKIP-3	0.771	0.500	**0.617**
SJTU-3	0.809	**0.710**	0.417

Table 3. Comparisons on Error Detection measures in SIGHAN-7 Sub-Task 1.

Error Detection	Accuracy	Precision	Recall
NCKU&YZU-1	0.729	0.650	0.217
NTHU-3	**0.861**	0.846	0.657
SinicaCKIP-3	0.842	0.692	0.853
SJTU-3	0.844	0.909	0.533
NTOU-1	0.314	0.304	**1.000**

Tables 2 and 3 show the results on error location detection and error detection in Sub-Task 1 (Detection). The difference between these two is that "error location detection" requires the detected location is correct while "error detection" will report correctly detected even the locations in sentences is not correct. Therefore it is expected that scores of Error Location Detection are a little bit higher than those of Error Detection. Our system exhibits a relative smaller difference between these two scores, 2.4%, compared to other systems.

The major weakness of our system is its low recall rate, which might be the result of not applying an error detection module. Therefore an error detection method using web-based measure will be examined in our future work.

Table 4. Comparisons on False-Alarm Rate and Detection Accuracy in SIGHAN-7 Sub-Task 1.

Error Detection	False-Alarm Rate	Accuracy
NCKU&YZU-1	0.050	0.729
NTHU-3	0.051	**0.861**
SinicaCKIP-3	0.163	0.842
SJTU-3	**0.023**	0.844

Table 5. Comparisons on Correction Accuracy and Precision in SIGHAN-7 Sub-Task 2.

Error Correction	Accuracy	Precision
NCKU&YZU-1	0.109	0.466
NTHU-3	0.443	0.700
SinicaCKIP-3	0.516	0.616
SJTU-3	0.356	**0.705**
NCYU-2	**0.625**	0.703
NCYU-3	**0.625**	0.703

Table 4 shows the error detection accuracy of our system is significantly lower although False-Alarm Rate is relatively small. The correction accuracy and precision are also much lower than high-ranked systems in the Bakeoff, as shown in Table 5. Further investigation is required to examine if more thoroughly exploiting web-based measures will provide useful additional information for the purpose of Chinese spelling error detection and correction.

5 Conclusions and Future Work

In this paper, a web-based measure is employed in addition to language models as a metric to score sentence correction candidates. The goal of this approach is to exploit as much texts (i.e., the web) as possible to provide useful information for error correction purposes.

The approach's major obstacle to participate in the Bakeoff's final test is our limited resources to access the results of search engines within two days. This has forced our final participating system to only take advantage of web-based measure in correction candidates' very last decisions. Further experiments administered on more thorough uses of web-based measure are required in the applications of Chinese spelling errors detection and correction.

The results of our system have confirmed the value of using a separate error detection module, i.e., detecting if there is an error in a sentence regardless where the error situated, such that sentences with no (detected) errors won't go through the error correction module.

Our direct future work would consist of 1) the inclusion of a separate error detection module, and 2) the administering of experiments exploiting web-based measure conforming to the method described in Section 3. A decomposition approach of web-based measure is also desirable to minimize runtime reliance on search engines.

Acknowledgments

This work was supported by National Science Council (NSC), Taiwan, under Contract number: 102-2221-E-155-029-MY3.

References

Chen, Y.-Z., Wu, S.-H., Yang, P.-C., & Ku, T. (2011). Improve the Detection of Improperly Used Chinese Characters in Students' Essays with Error Model. *International Journal of Continuing Engineering Education and Life Long Learning, 21*(1), 103-116.

Cilibrasi, R. L., & Vitanyi, P. (2009). Normalized Web Distance and Word Similarity. *arXiv preprint arXiv:0905.4039*.

Cilibrasi, R. L., & Vitanyi, P. M. (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering, 19*(3), 370-383.

Gracia, J., & Mena, E. (2008). Web-based Measure of Semantic Relatedness *Web Information Systems Engineering-WISE 2008* (pp. 136-150): Springer.

Liu, C.-H., Wu, C.-H., & Harris, M. (2008). *Word Order Correction for Language Transfer Using Relative Position Language Modeling.* Paper presented at the 6th International Symposium on Chinese Spoken Language Processing (ISCSLP'08).

Liu, C.-L., Lai, M.-H., Chuang, Y.-H., & Lee, C.-Y. (2010). *Visually and Phonologically Similar Characters in Incorrect Simplified Chinese Words.* Paper presented at the The 23rd International Conference on Computational Linguistics: Posters.

Liu, C.-L., Tien, K.-W., Lai, M.-H., Chuang, Y.-H., & Wu, S.-H. (2009). *Phonological and Logographic Influences on Errors in Written Chinese Words.* Paper presented at the The 7th Workshop on Asian Language Resources.

Lovelyn Rose, S., & Chandran, K. (2012). Normalized Web Distance Based Web Query Classification. *Journal of Computer Science, 8*(5), 804-808.

Ma, W.-Y., & Chen, K.-J. (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.

Macias, D., Wong, W., Thangarajah, J., & Cavedon, L. (2012). *Coherent Topic Transition in a Conversational Agent.* Paper presented at the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH).

Stolcke, A. (2002). *SRILM-an extensible language modeling toolkit.* Paper presented at the 7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002.

Wu, C.-H., Liu, C.-H., Harris, M., & Yu, L.-C. (2010). Sentence Correction Incorporating Relative Position and Parse Template Language Models. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(6), 1170-1181.

Wu, S.-H., Chen, Y.-Z., Yang, P.-c., Ku, T., & Liu, C.-L. (2010). *Reducing the False Alarm Rate of Chinese Character Error Detection and Correction.* Paper presented at the The CIPS-SIGHAN Joint Conference on Chinese Language Processing.