Description of HLJU Chinese Spelling Checker for SIGHAN Bakeoff 2013

Yu He
School of Computer Science and Technology
Heilongjiang University
Harbin 150080, China
heyucs@yahoo.com

Guohong Fu
School of Computer Science and Technology
Heilongjiang University
Harbin 150080, China
ghfu@hotmail.com

Abstract

In this paper, we describe in brief our system for Chinese Spelling Check Backoff sponsored by ACL-SIGHAN. It consists of three main components, namely potential incorrect character detection with a multiple-level analysis, correction candidate generation with similar character sets and correction scoring with n-grams. We participated in all the two sub-tasks at the Bakeoff. We also make a summary of this work and give some analysis on the results.

1 Introduction

As one typical task in written language processing, spelling check is aiming at detecting incorrect characters within a sentence and correcting them. While a number of successful spelling checker have been available for English and many other other alphabetical languages, it is still a challenge to develop a practical spelling checker for Chinese due to its language-specific issues, in particular the writing system of Chinese without explicit delimiters for word boundaries. Furthermore, no data set are commonly available for spelling check in Chinese. As such, ACL-SIGHAN sponsor a Backoff on Chinese spelling check, which consists of two subtasks, namely spelling error detection and spelling error correction.

Based on the task specification the data sets for SIGHAN Backoff 2013, we develop a spelling checker for Chinese. It consists of three main components, namely potential incorrect character detection with a multiple-level analysis, correction candidate generation with similar character sets and correction scoring with n-grams. We have participated in all the two sub-tasks at the Bakeoff. We also make a summary of this work and give some analysis on the results.

The rest of this paper is organize as follows. First, we describe in brief our system for Chinese spelling check in Section 2. Then in Section 3, we present the settings or configuration of our system for different subtasks, and report the relevant results at this Bakeoff. Finally, we give our conclusions on this work in Section 4.

2 Proposed Method

2.1 System Architecture

Figure 1 shows the architecture of our system. It works in three main steps. Given a plain Chinese sentence with/without spelling errors, our system first segments it to words. Then, a multi-level analysis module is used to detect potential incorrect characters within the input and thus a 5401×5401 similarity matrix generated from the similar character set (viz. the Bakeoff 2013 CSC Datasets) (Liu et al., 2011) is further employed to generate set of corrections for the input. Finally, n-grams are used to score and decode a sentence as the best correction for the input. For convenience, we refer to this sentence as output sentence. If the output sentence is same as the original input sentence, then the input sentence does not contain any spelling errors; Or else, it has incorrect characters, and the output sentence would be its correction.

In the figure above, CLA is the abbreviation for character level analysis, WLA means word level analysis and CLA_2 represents context level analysis.

2.2 Potential Incorrect Character Detection

2.2.1 Types of incorrect words in Chinese

In general, Chinese words with incorrect characters (refered to as incorrect Chinese words thereafter) have three main ways of segmentations.

(1) The segmentation of an incorrect Chinese word would be a sequence of single-character

Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), pages 84–87,
Nagoya, Japan, 14 October 2013.

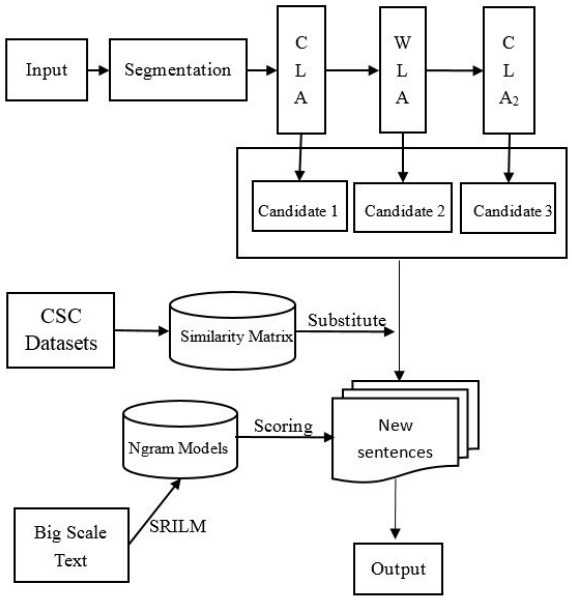

Figure 1: The architecture of our system

words. For example, 煩腦 is a common incorrect form of the word 煩惱 fan-nao 'trouble', and it will be segmented into two separate single-character words, namely, 煩 and 腦, during word segmentation;

(2) The segmentation of an incorrect word is still a word. For example, 措折 is a typical incorrect form of the word 挫折 cuo-zhe 'setback', and is usually segmented into one word 措折(Vt) after word segmentation. Here, we refer such

(3) An incorrect word and its adjacent characters in the context will form another word after segmentation. For example, the fragment 一番新的进攻 yi-fan-xin-de-jin-gong 'some new offensive' may be wrongly written as 一翻新的进攻. Here, the incorrect word 翻 fan 'turn over' and its left character 新 xin 'new' in the fragment will form a verb 翻新 fan-xin 'retread' during word segmentation.

In terms of the above different ways of segmentation, we can classify incorrect words in Chinese into three types, namely as character-level errors (CLEs), word-level errors (WLEs) and context-level errors (CLEs), respectively.

2.2.2 Detecting incorrect characters with a multi-level analysis

In order to reduce the space and noise in decoding for spelling error correction, we employ a three-level analysis strategy to identify the above mentioned three types of incorrect words and thus detect all potential incorrect characters within the input sentence.

Character－level analysis is for detection CLES within a given input sentence. After the statistical analyzing of the large formal tokenized text, we use the following formula to calculate each character w's probability to be a single-character word:

$$P_{sw}(w) = \frac{Count(w \text{ is a single word})}{Count(w)}$$

For each single-character word within a segmented sentence, if the value of P_{sw} is less than a given threshold, then we will regards it as a candidate incorrect character.

Word-level analysis is for identifying WLEs. In this case, we just need to take all the out-of-vocabulary words (OOVs) as the candidates for the word layer processing. The candidate contains two cases, one is the wrong word, and the other is OOVs.

Context-level analysis is for CLEs. Here, we use n-grams to detect such types of errors. Considering the previous example 一翻新的, we can observe that $P(\text{的}|\text{一,翻新}) = 0$, $P(\text{的}|\text{翻新}) = 0.385$, $P(\text{翻新}|\text{一}) = 0$ through n-gram models, indicating that the word does not exist. This may be due to the word 一 yi 'one'. However, the incorrect character is 翻 rather than 一. So we can conclude that its neighbor is not reliable when CLEs occur. Thus we take "一" and all words around it within a window range of 1 as the incorrect character candidates.

2.3 Correction Candidate Generation with Similarity Matrix

Organizers have provided us with a group of similar character sets (CSC)(Liu et al., 2011), which includes similar shape and similar pronunciation, and latter are divided into "same sound and same tone (SS)", "same sound and different tone (SD)", "similar sound and same tone (MS)" and "similar sound and different tone (MD)" and so on. As follows:

Similar Shape: 可, 何呵坷奇河柯苛阿倚寄崎...
Similar Pronunciation: 右, 幼鼬誘宥柚祐有侑...

Through the statistical analyzing of the sample data, we found that the similar pronunciation er-

rors accounting for more than 80%, nevertheless, only 10% of similar shape errors, and the other errors accounted for about 10%. Therefore, we believe that similar pronunciation words should have greater weight. We take the words (5401 words) of the data set as the matrix's rows and columns, the elements of the matrix are the similar weights between two words. The degree of similarity is divided into five levels, namely, the same sound and same tune (SS) is 1, similar morphology is 2, the same sound and different tune (SD) is 3, similar sound and same tone (MS) is 4 and similar sound and same tone (MS) is 5. So we can get a diagonal matrix (the value of diagonal elements all is 0), called the similarity matrix.

2.4 Correction Scoring with n-grams

We take all the candidate words and the words around them within the window range of 1 in a sentence S to be replaced by the similar words successively. Using the following formula to calculate the new sentence $S^{'}$ probability score.

$$score(S^{'}) = \prod \alpha \times P_{trigram}(w|w_{i-1}w_{i-2}) + (1-\alpha) \times \beta \times P_{bigram}(w|w_{i-1} + (1-\alpha) \times (1-\beta) \times P_{unigram}(w)$$

The value of α in the models determine the weight of $P_{trigram}$, the greater of α, the more greater proportion of $P_{trigram}$; The value of β determine the weight of P_{bigram} .

3 Experimental Results

Our system participated in both subTask at the Chinese Spelling Check Bakeoff. This section reports the results and discussions on its evaluation.

3.1 Experimental Settings

As mentioned above, SIGHAN Bakeoff 2013 consists of two sub-tasks: namely error detection (viz. Subtask 1) and error correction (viz. Subtask2). For the error detection task, the system should return the locations of the incorrect characters for a given Chinese sentence that may have or do not have spelling errors, while in Subtask2, the system should return the locations of the incorrect characters within the input and correct them. Obviously, Subtask2 is a follow-up problem of error detection for sentences with errors.

In SIGHAN Bakeoff 2013, ninth measures for subTask1 and three measures for subTask2 are employed to score the performance of a spelling correction system. They are False-Alarm Rate(FAR), Detection Accuracy(DA), Detection Precision(DP), Detection Recall(DR), Detection F-score(DF), Error Location Accuracy(ELA), Error Location Recall(ELR), Error Location F-score(ELF), Location Accuracy(LA), Correction Accuracy(CA) and Correction Precision(CP).

In our system, we employ the SRILM Toolkit(Stolcke and others, 2002) to build n-gram models for spelling correction selection from the Academia Sinica Segmentation Corpus(3.0) (Chen et al., 1996). Furthermore, we use the similar character sets (CSC datasets)(Liu et al., 2011) to build the similarity matrix for correct sentence candidate generation. In addition, we also uses Academia Sinica Segmentation System (CKIP)(Ma and Chen, 2003) to perform word segmentation.

4 Experimental results and discussion

We use three different sets of parameters presented three sets of results, namely HLJU_Run1, HLJU_Run2 and HLJU_Run3. See the table 1 below for details:

	Model α	Model β	Similarity
Run1	0.8	0.8	$5 \leq$
Run2	0.8	0.8	$2 \leq$
Run3	0.6	0.8	$5 \leq$

Table 1: Parameter Selection.

α and β have been introduced in section 2.4. The Similarity less than or equal a value x , it represents only consider the similarity less than x characters in similarity matrix. For example, the Similarity of Run2 is less than 2, so we consider only two cases, the "same sound and same tone (SS)" and "similar shape".

Table 2 shows the result of sub-Task1 and Table 3 shows the result of sub-Task2. The "Best" indicates the high score achieved in Chinese Spelling Check task. The "Average" represents the average level. The numbers in boldindicate the highest values of each metric.

From the above table, we can see that results are not satisfactory, and many metrics from the best score is still a certain gap. The value of FAR is too high, and the precision is low. It means our method causes a lot of false positive errors and shows our system is not strictly for candidate list. And the parameter setting remains to be determined. In ad-

	FAR	DA	DP	DR	DF	ELA	ELP	ELR	ELF
Run1	0.6857	0.514	0.3798	0.98	0.5474	0.301	0.1047	0.27	0.1509
Run2	**0.6529**	**0.529**	**0.3849**	0.9533	**0.5484**	**0.339**	**0.1292**	**0.32**	**0.1841**
Run3	0.6929	0.51	0.3782	**0.9833**	0.5463	0.296	0.1038	0.27	0.15
Average	0.3222	0.698	0.5847	0.7454	0.6064	0.591	0.3472	0.3887	0.3418
Best	0.0229	0.861	0.9091	1	0.7642	0.82	0.7102	0.6167	0.5854

Table 2: Evaluation Results of Sub-Task1.

	LA	CA	CP
Run1	0.265	0.225	0.2432
Run2	**0.323**	**0.277**	**0.3081**
Run3	0.264	0.222	0.2403
Average	0.415	0.3788	0.5026
Best	0.663	0.625	0.705

Table 3: Evaluation Results of Sub-Task2.

dition, I think there are some other reasons for this results:

1) There are some errors in the training and CSC data set, and we do not deal with it;

2) Our methods are still based on ngram models for correcting spelling errors, and we failed to the breakthrough.

However, the performance of Run2 is much better than other schemes. We can conclude that low character similarity has no any help for the correction task.

5 Conclusion

In this paper, we have presented a spelling checker for Chinese. It consists of three main modules, namely potential incorrect character detection with a multiple-level analysis, correct sentence candidate generation with similar character sets and correction scoring with n-grams. We have participated in all the two sub-tasks at the ACL-SIGHAN Chinese Spelling Check Bakeoff. Since our system is still under development, the results are not satisfactory. For future work, we hope to explore more complicated techniques to achieve precise error detection and correction decoding.

Acknowledgments

This study was supported by National Natural Science Foundation of China under Grant No.60973081 and No.61170148, the Returned Scholar Foundation of Educational Department of Heilongjiang Province under Grant No.1154hz26, and Harbin Innovative Foundation for Returnees under Grant No.2009RFLXG007, respectively.

References

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica corpus: Design methodology for balanced corpora. *Language*, 167:176. `http://rocling.iis.sinica.edu.tw/CKIP/publication.htm`.

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 (TALIP)*, 10(2):10.

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 the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 168–171. Association for Computational Linguistics.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*.