Introduction to NJUPT Chinese Spelling Check Systems in CLP-2014 Bakeoff

Lei Gu, Yong Wang, Xitao Liang
Nanjing University of Posts and Telecommunications, China
gulei@njupt.edu.cn, 13041127@njupt.edu.cn, centaolang@163.com

Abstract

Chinese spelling check (CSC) is an essential issue in the research field of Chinese language processing (CLP). This paper describes the details of two CSC systems we developed to solve this problem. The first system was built based on CRF model, and the modules of such system include word segmentation, error detection and error correction. Another system was based on 2-Chars&&3-Chars model, and its modules include bigram segmentation, error detection and error correction. Using the final test data set provided by CLP2014, the final experimental result of the system based on 2-Chars&&3-Chars model was better, which achieved 0.403 detection accuracy with 0.3344 detection precision and 0.3964 correction accuracy with 0.3191 correction precision.

1 Introduction

Language Spelling check is an important subject in the field of language processing both in Chinese and English. Compared with English, how to detect and correct spelling errors in Chinese sentences automatically is more difficult. In English, there are two classes of spelling errors: non-word spelling errors and real-word spelling errors. Non-word spelling errors generally refer to the wrong spelling words that not exist in a dictionary, such as a sentence 'buu some apples' where 'buu' is an error word which can't be found in a dictionary. Real-word spelling errors usually refer to the wrong words which are misused in sentences but exist in a dictionary, for example, in the sentence 'bye some apples', 'bye' is misused in such sentence but can be found in a dictionary.

Chinese spelling check is different. Firstly, for Chinese electronic documents, there are not non-word spelling errors, because each misspelled character is exist in reality, such as "產玲婦女 chan ling fu nv", "玲(齡) ling" is a character misused but exists in reality. Secondly, in English sentences, each word is separated by a space, so it's easier to detect misspelled words. But there are no word delimiters between words in Chinese sentences, and a Chinese word may consist of a single-character or more, so it's hard to decide whether a single-character is wrong or it's a part of a misspelled word.

Generally, phonologically similar or visually similar characters result in the misspelled words in Chinese sentences. For instance, "嬰兒個紓 ying er ge shu", "數 shu" is misspelled as "紓 shu" because both are pronounced as "shu". In the sentence "不斷曾加 bu duan ceng jia", "增 zeng" is misspelled as "曾 ceng" because "曾" is similar with "增" in visual. For most CSC systems, to correct the misspelled words, it's necessary to build a module to replace the wrong characters by similar characters extracted from the character confusion sets which are edited based on phonologically and visually similarity between characters. In our experiment, the confusion sets provided by SIGHAN Bake-off 2013 are used in both CSC systems (Wu et al., 2013).

Lots of colleges and research institutions have made efforts to solve such CSC problems in recent years. There have been two types of methods of spelling check: rule-based methods and statistical methods. Data driven, the statistical spelling check approaches appear to be more robust and performs better than simple rule-based methods (Chiu et al., 2013). Wang et al. (2013) built a system and its main idea is to exchange potential error character with its confusable ones and rescore the modified sentence using a conditional random field (CRF)-based word segmentation/part of speech (POS)

tagger and a tri-gram language model (LM) to detect and correct possible spelling errors. Lin and Chu (2013) also proposed a system and the modules in their system include word segmentation, N-gram model probability estimation, similar character replacement, and filtering rules. In this paper, we build two CSC systems based on CRF model and 2-Chars&&3-Chars model. The rest of this paper will introduce the two CSC systems in detail, and it's organized as follows. We will introduce the first system based on CRF model in section 2, in section 3 we'll describe the second system based on 2-Chars&&3-Chars model, at last we'll make conclusions in section 4.

2 System Based On CRF Model

As is shown in Figure 1, our system gets the input sentences firstly, then the sentences will be segmented by word which is based on CRF model, after the step of word segmentation, error words in the sentences segmented will be picked out by some rules and be dealed with the module of error correction. Details of the models will be discussed in the following subsections.

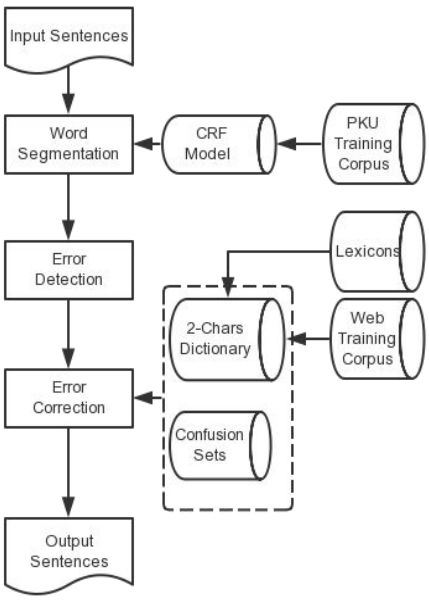

Figure 1. Framework of CSC system based on CRF model

2.1 Word Segmentation

Chinese word segmentation (CWS) is the first step for Chinese language processing. In recent years, Chinese spelling checkers have incorporated word segmentation (Chiu et al., 2013) and many word segmentation methods have been proposed. Such as support vector machine (SVM), conditional random field (CRF) and maximum entropy Markov models (MEMMs), among them, CRF-based approach has been shown to be effective with very low computational complexity (Wang et al., 2013).

The module word segmentation of our first CSC system uses condition random fields (CRF) approach. CRFs are a class of undirected graphical models with exponent distribution (Lafferty et al., 2001). A common used special case of CRFs is linear chain, which has a distribution of:

$$P_{\wedge}(\bar{y} \mid \bar{x}) = \frac{1}{Z_{\bar{x}}} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, \bar{x}, t)\right)$$

where $f_k(y_{t-1}, y_t, \bar{x}, t)$ is a function which is usually an indicator function; λ_k is the learned weight of feature f_k; and $Z_{\bar{x}}$ is the normalization factor. The feature function actually consists of two kinds of features, that is, the feature of single state and the feature of transferring between states.

In this system, we use a public tool CRF++ (Li et al., 2009) for CRF implementation and regard the PKU (Emerson, 2005) corpus as the training corpus.

The process of word segmentation using CRF++ is as follows:

a. Convert the simplified Chinese sentences in the PKU training corpus to traditional Chinese;

b. Train the CRF++ tool;

c. Segment the sentences inputted into this system.

2.2 Error Detection

If there're no misspelled words in a sentence, the sentence could be divided into serial correct words after ideal word segmentation. But if a sentence contains misspelled words, the segmentation could separate words containing misspelled character by serial single characters (Chang et al., 2013). For instance, the sentence "儘管婦女的數量不斷增加 jin guan fu nv de shu liang bu duan zeng jia" which has no misspelled words will be segmented into "儘管/婦女/的/数量/不斷/增加". However, the sentence "儘管婦女的數量不斷正加 jin guan fu nv de shu liang bu duan zheng jia" with an error word "正加" ("增 zeng" is misspelled as

“正 zheng”) will be segmented into “儘管/婦女/的/數量/不斷/正/加”. In this sentence “正加” is an error word, so it is segmented into serial two single characters “正” and “加”.

In the error detection module of our first CSC system , we make a rule that error occurs in the serial single characters generated by word segmentation. Like the serial characters “正” and “加”, one of these serial characters should be an error.

2.3 Error Correction

In this system, we build a 2-Chars dictionary extracted from a large number of lexicons and a web training corpus which is collected from lots of news reports, compositions and other data on the web.

The way to build a 2-Chars dictionary is as follows:

a. Segment the sentences in web training corpus by bigram. For example, “邁向充滿希望的新世紀 mai xiang chong man xi wang de xin shi ji” will be segmented as “邁向/向充/充滿/滿希/希望/望的/的新/新世/世紀”;

b. Count the frequency (indicates how many times a word presents in the web training corpus) of each word;

c. Add each word and its frequency into the 2-Chars dictionary.

The format of words in the 2-Chars dictionary is [Word:Frequency]. For example:

邁向:23　向充:3　充滿:75　滿希:7

希望:322　望的:16　的新:195　新世:25

世紀:230　...

In our system, we just deal with the error words consist of two characters. We take the serial single characters “正” and “加” for example. Firstly, “正” will be replaced by its similar character lists one by one, then the similar character will be combined with “加” to a new word “?加”. If the new word “?加” do exist in the 2-Chars dictionary, the similar character will be added into the candidates list. After the treatment of all similar characters of “正”, “加” will be replaced by its similar character lists like the processing of “正”. At last, if the length of candidates list is more than one, we will choose the new word with the highest frequency in the 2-Chars dictionary.

Table 1 and table 2 show the processing of “正” and “加”, from these two table, we can find that the frequency of new word “增加” consist of “增” and “加” is higher than “正下” and “正夾”, so “增” should be the correct character of “正”.

Confusion Sets	New Word	Exist In 2-Chars Dic?, Frequency
陣	陣加	False
禎	禎加	False
增	增加	True, 248
鴆	鴆加	False
...		

Table 1. The processing of “正”

Confusion Sets	New Word	Exist In 2-Chars Dic?, Frequency
家	正家	False
下	正下	True, 1
茄	正茄	False
夾	正夾	True, 1
...		

Table 2. The processing of “加”

2.4 Analysis Of The Result

We submitted two experimental results using two different number of lexicons. As shown in table 3 and table 4, the final results of the first CSC system are not so good.

The defect of word segmentation and the limit of 2-Chars dictionary may result in the bad result. Besides the future work of improving the performance of this CSC system, we propose another CSC system without word segmentation in section 3.

Run-2	Accuracy	Precision	Recall	F1
Detection Level	0.275	0.202	0.1525	0.1738
Correction Level	0.258	0.1645	0.1186	0.1379
False Positive Rate	0.6026			

Table 3. Run-2 result of system based on CRF model

Run-3	Accuracy	Precision	Recall	F1
Detection Level	0.2853	0.1885	0.1299	0.1538
Correction Level	0.2665	0.1416	0.0923	0.1117
False Positive Rate	0.5593			

Table 4. Run-3 result of system based on CRF model

3 System Based On 2-Chars&&3-Chars Model

Although the first module of most Chinese spelling checkers are word segmentation, there still exist many problems which may have bad influences on the next modules of the spelling checkers. Such as "但是 婴兒出生率不正加反而減少 dan shi ying er chu sheng lv bu zheng jia fan er jian shao" ("增 zeng" is misspelled as "正 zheng"), the result of word segmentation is "但是/婴兒/出生率/不正/加/反而/減少" where "不正" is regarded as a word which results in the neglect of wrong word "正加".

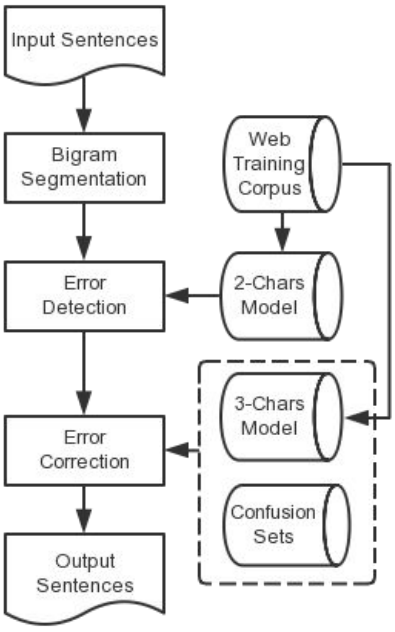

Figure 2. Framework of CSC system based on 2-Chars&&3-Chars model

According to the reasons above, we propose a system without word segmentation. Figure 2 shows the framework of our second system based on 2-Chars&&3-Chars model. After getting input sentences, system will segment them by bigram, then the next module based on 2-Chars model will detect errors in these segmented sentences. After error detection, a 3-Chars model is used to correct errors by some rules. Details of this system will be described in the following subsections.

3.1 Bigram Segmentation

A significant difference between the bigram segmentation and the word segmentation is: words in the sentences are non overlapping, but bigrams are overlapping.

With respect to the sentence "全球的婦女人口正加 quan qiu de fu nv ren kou zheng jia", the segmentation results of different methods are as follows:

By word: 全球/的/婦女/人口/正/加

By bigram: 全球/球的/的婦/婦女/女人/人口/口正/正加

Compared with word segmentation, it's easier to segment sentences by bigram, because it don't need any segmentation tools. In this CSC system, all sentences will be segmented by bigram. After segmentation, this system will detect errors in these bigrams.

3.2 Error Detection

In this system, we build a 2-Chars Model and a 2-Chars dictionary extracted from a web training corpus which is collected from lots of news reports, compositions and other data on the web. The format of words in this 2-Chars dictionary is the same as the dictionary in the first CSC system.

In the sentence "全球的婦女人口正加 quan qiu de fu nv ren kou zheng jia", "增 zeng" is misspelled as "正 zheng" so the result of bigram segmentation is: 全球/球的/的婦/婦女/女人/人口/口正/正加.

The module of error detection gets a string array consist of the results of segmentation. Take the first word "全球" as an example, we call it "Current-Word (C-Word)" and its next word "球的" is called "Next-Word (N-Word)". We make a rule that if C-Word ("全球") or N-Word ("球的") don't exist in the 2-Chars dictionary, the second character of C-Word "球" would be an error.

Using the rule above, the system will find "口正" isn't exist in the 2-Chars dictionary, then "正" is regraded as an error.

3.3 Error Correction

Like 2-Chars model, we also build a 3-Chars model. And we edited a 3-Chars dictionary just like 2-Chars dictionary but ignore the frequency of a word.

The method of building a 3-Chars dictionary is segmenting the sentences in web training corpus by trigram. For example, "邁向充滿希望的新世紀 mai xiang chong man xi wang de xin shi ji" will be segmented as "邁向充/向充滿/充滿希/滿希望/希望的/望的新/的新世/新世紀";

Compared with the format of words in 2-Chars dictionary, the format of 3-Chars words in the dictionary is as follows:

邁向充　向充滿　充滿希　滿希望
希望的　望的新　的新世　新世紀
. . .

As shown in the module of error detection, "正" is an error character in the word "口正" (C-Word). We combine "口正" with its next word "正加" (N-Word) into a new 3-Chars word "口正加", then the error "正" will be replaced by the characters extracted from its confusion sets. If a new 3-Chars word "口?加" can be found in the 3-Chars dictionary, the similar character will be regarded as the correct one.

Table 5 shows the method of determining whether a new word is correct or not. As shown in this table, "口增加" do exist in the 3-Chars dictionary, and "增" should be the correct one.

Confusion Sets	New Word	Exist In 3-Chars Dic?
陣	口陣加	False
禎	口禎加	False
增	口增加	True
鴆	口鴆加	False
...		

Table 5. The processing of "正"

3.4 Analysis Of The Result

Table 6 shows the result of system based on 2-Chars&&3-Chars model. We found that all the performances of this system is better than the system based on CRF model.

Run-1	Accuracy	Precision	Recall	F1
Detection Level	0.403	0.3344	0.1959	0.247
Correction Level	0.3964	0.3191	0.1827	0.2323
False Positive Rate	0.3898			

Table 6. The Result Of System Based On 2-Chars&&3-Chars Model

4 Conclusion

In this paper, we introduce two Chinese spelling check systems and the experimental results show that the CSC system without word segmentation do better than the system incorporated with word segmentation. The work to improve the performance of the system with word segmentation is still continued. And in the future, we'll do more research and work on the system based on 2-Chars&&3-Chars.

References

Tao-Hsing Chang, Hsueh-Chih Chen, Yuen-Hsien Tseng, and Jian-Liang Zheng. 2013. Automatic Detection and Correction for Chinese Misspelled Words Using Phonological and Orthographic Similarities. Proceedings of SIGHAN-7, 97-101.

Hsun-wen Chiu, Jian-cheng Wu, and Jason S. Chang. 2013. Chinese Spelling Checker Based on Statistical Machine Translation. Proceedings of SIGHAN-7, 49-53.

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. Proceeding of the Fourth SIGHAN Workshop on Chinese Language Processing, 123-133.

Yu He and Guohong Fu. 2013. Description of HLJU Chinese Spelling Checker for SIGHAN Bakeoff 2013. Proceedings of SIGHAN-7, 84-87.

Yu-Ming Hsieh, Ming-Hong Bai, and Keh-Jiann Chen. 2013. Introduction to CKIP Chinese Spelling Check System for SIGHAN Bakeoff 2013 Evaluation. Proceedings of SIGHAN-7, 59-63.

Zhongye Jia, Peilu Wang, and Hai Zhao. 2013. Graph Model for Chinese Spell Checking. Proceedings of SIGHAN-7, 88-92.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proceedings of International Conference of Machine Learning, 591-598.

Shoushan Li and Chu-Ren Huang. 2009. Word Boundary Decision with CRF for Chinese Word Segmentation. 23rd Pacific Asia Conference on Language, Information and Computation, 726-732.

Chuan-Jie Lin and Wei-Cheng Chu. 2013. NTOU Chinese Spelling Check System in SIGHAN Bake-off 2013. Proceedings of SIGHAN-7, 102-107.

Yih-Ru Wang, Yuan-Fu Liao, Yeh-Kuang Wu, and Liang-Chun Chang. 2013. Conditional Random Field-based Parser and Language Model for Traditional Chinese Spelling Checker. Proceedings of SIGHAN-7, 69-73.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013. Proceedings of SIGHAN-7, 35-42.

Zhiting Xu, Xian Qian, Yuejie Zhang, and Yaqian Zhou. 2007. CRF-based Hybrid Model for Word Segmentation, NER and even POS Tagging. Proceedings of Sixth SIGHAN Workshop on Chinese Language Processing, 167-170.