A Maximum Entropy Approach to Chinese Spelling Check

Dongxu Han, Baobao Chang
Institute of Computational Linguistics, Peking University
Key Laboratory of Computational Linguistics(Peking University), Ministry Education, China
Beijing, 100871, P.R.China
{handx,chbb}@pku.edu.cn

Abstract

Spelling check identifies incorrect writing words in documents. For the reason of input methods, Chinese spelling check is much different from English and it is still a challenging work. For the past decade years, most of the methods in detecting errors in documents are lexicon-based or probability-based, and much progress are made. In this paper, we propose a new method in Chinese spelling check by using maximum entropy (ME). Experiment shows that by importing a large raw corpus, maximum entropy can build a well-trained model to detect spelling errors in Chinese documents.

1 Introduction

Because of the popularity of computers, more and more documents are produced. For the carelessness of human or errors of OCR image recognition, many spelling errors occur in documents, which seriously interferes documents quality. Proofreading by human to correct the errors is laborious and expensive, so an automatic approach is badly in need. Automatic spelling check can identify incorrect writing words in documents, which plays an important role in documents writing and OCR post-processing.

Research on automatic spelling check of English documents began in the 1960s (Damerau F.J., 1964), many studies have been proposed and quite good results have been obtained. While spelling check of Chinese is still a challenging work due to some special processing difficulties arising from Chinese writing, which hardly occur in spelling check of English.

In English writing, each word is directly input by Latin letters, so the spelling errors are only the situation that one letter is mistaken written to another, such as writing "bcg" instead of "bag", or "son glasses" instead of "sun glasses". The former is a non-word spelling error, meaning the form of input word is definitely incorrect and latter is a real-word spelling error, meaning the form of input word can be found in the dictionary but incorrectly used.

In Chinese writing, unlike English, all legal characters (we call them hanzi) have been stored in a font lib and Chinese input system builds an effective map between Latin letters and hanzi fonts. For the reason of input methods, Chinese characters would not take the non-word errors such as missing or adding a part of character to form an illegal character in the dictionary. That is, all Chinese spelling errors are real-word errors. The treatment of real-word errors needs analyzing the context, which is much harder than the treatment of non-word errors. Chinese spelling check is still a challenging work.

In this paper, we propose a new but simple method in Chinese spelling check by using maximum entropy (ME) models. We train a maximum entropy model for each Chinese character based on a large raw corpus and use the model to detect the spelling errors in documents. Tentative experiment in the bakeoff shows the simple strategy works. However, further refinement and methodology combinations seem still needed to produce state-of-arts results.

The rest of the paper is organized as follows: In section 2, we give a brief introduction to the Chinese spelling check. In section 3 we introduce our approach to Chinese spelling check using maximum entropy model. Section 4 is description and discussion of our experiments. Section 5 is the conclusion.

Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), pages 74–78,
Nagoya, Japan, 14 October 2013.

2 Previous work

Research on Chinese automatic spelling check approaches appeared in 1990s (shih et al. 1992). Most of them are generally based on lexicon methods and statistic methods.

Lexicon-based methods use dictionaries, which contain as much as possible language information, such as word information, characters and words frequency information, encoding information, part-of speech tagging information and similar character information. Chinese characters are usually mistakenly written as some other characters, because their shapes or pronunciations are very similar or even the same in pronunciation. Such characters are called Chinese similar characters, and most of Chinese spelling errors are caused by them. In order to improve the performance of spelling check, these similar characters are summarized to similar character dictionaries, for example, the shape similar characters set and the pronunciation similar characters set provided by the bakeoff organization, are both similar character dictionaries (Liu, 2011).

Chang (1995) replaced each character in a sentence with another similar character by a large-enough similar character dictionary and calculated the replaced sentence score, to judge whether a character should be replaced with another. Zhang et al. (2000a) made use of characters and words frequency, similar character dictionary, and part of speech (POS) tagging information to detect dubious areas and generate candidate words. Zhang et al. (2000b, 2000c) used WuBi encoding information and Lin (2002) used Chong-Je encoding information to estimate dubious characters. These kinds of methods achieve success in some aspects, like Liu (2011) using Chong-Je encoding information could detect 93.37% error characters.

Statistic-based methods usually use a huge language corpus and the product of conditional probabilities to compute the appearance probability of a sentence (shih et al. 1992). Moreover, most of the statistic-based methods of Chinese spelling check jointly use lexicon-based methods together so as to achieve better performance. Like Ren (1994) used language model with word frequency dictionary, and Huang (2007) used language model with word dictionaries and similar character dictionary.

In the following, we will introduce our approach to Chinese spelling check in statistic-based methods totally without lexicon-based methods.

3 Chinese spelling check based on maximum entropy model

In this section, we first formalize spelling check as classifying each character into right or wrong categories based on the characters before and after it. We then briefly describe our feature setting in modeling the spelling check task using maximum entropy model.

3.1 Reformulating error characters detecting as a classification problem

Deciding whether a character is correctly or incorrectly written can be treated as a classification problem. To do this, we train each character a model that can classify the character into two categories named right or wrong, which means the character is correctly or incorrectly used.

In a Chinese sentence, no character can exist independently. They are all associated with the characters previous or next. In order to gain the whole data meaning, a complete context must be extracted, not just the target character. For example, when we train the character "國" (country), we select the n-gram "中華民國十三年" as the training data. In this way, we import a large raw corpus, segment the corpus into sentences by the pronunciations and remove these pronunciations, from the sentences we extract the n-grams whose middle character is the character to be trained (for example "國"). Then, the training data of character "國" could be like this:

中華民國十三年
$H_{-2}H_{-1}$美國總統布
到市區國會山莊
需要跨國 $H_1H_2H_3$
… … … … … …

In the training data, if there is not enough characters after the target character in an n-gram, we use padding characters "H_1", "H_2" and "H_3" as the characters after it (here is "國"). So are the characters "H_{-1}", "H_{-2}" and "H_{-3}".

To judge whether a target character is correctly or incorrectly written, in the training data of the target character, there should be enough positive instances and negative instances for classification training. Intuitively, the positive instances are all the n-grams in the corpus whose middle character is the target character, and the negative instances are all the n-grams in the corpus whose middle character should not be the target character but mistaken written as the target character. But usually there are no incorrectly used characters in corpus, so we don't have the negative in-

stances like that way. Our method is that we replace all the n-grams in the corpus whose middle character is not the target character with the target character and choose these n-grams as the negative instances. In this way, the amounts of positive instances and negative instances are seriously imbalanced, the former too few and the latter too many. In order to reduce the amount of the negative instances, we import the similar character dictionaries provided by the bakeoff organization. We select the n-grams whose middle character is the similar character of the target character as the negative n-grams.

Then the positive instances are labeled right and the negative instances are labeled wrong. Also for example "國":

Right	中華民國十三年
Right	$H_{-2}H_{-1}$美國總統布
Right	需要跨國 $H_1H_2H_3$
Wrong	小吃店國商店 H_1
Wrong	一個月國更長時
Wrong	巡守巾國不讓鬚
… … … … … … … … … … …	

We use maximum entropy to train the training data and achieve corresponding model of each character. We extract each character in the test data to be the n-gram in the same way and classify the n-grams into right or wrong categories by the character corresponding model, judging whether the character is correctly or incorrectly written, achieving the result of Chinese spelling check.

3.2 Feature templates

In our raw corpus, after segmented by pronunciations, the average length of characters in sentences is 7.443, so the n-gram we set here is seven-gram, namely we extract both 3 characters before and after the target character as a training seven-gram.

The target character is set C_0, the characters previous are set C_{-1}, C_{-2}, C_{-3} and the characters next are set C_1, C_2 and C_3. We have following maximum entropy feature templates:

(a) C_n (n=-3, -2, -1, 1, 2, 3)
(b) C_nC_{n+1} (n=-3, -2, 1, 2)
(c) $C_{-1}C_1$, $C_{-1}C_2$, $C_{-2}C_1$, $C_{-2}C_2$
(d) $C_{-2}C_{-1}C_1$, $C_{-1}C_1C_2$
(e) $C_{-2}C_{-1}C_1C_2$

From feature templates above, we can see that we train the character through the information of characters before and after it, so the C_0 actually cannot be used.

4 Experiments and discussions

We choose to use maximum entropy toolkit[1] as our model learner and we use traditional Chinese part of Chinese Gigaword corpus as our training data.

4.1 Training data

The traditional Chinese part of Gigaword corpus has about 800 million characters, covering over 9000 different characters. We select 5311 different characters mainly appear in the corpus, covering over 95% of the corpus.

Corresponded to the 5311 different characters, 5311 training data are made, each of which contains around 7.48 million seven-grams, and 5311 maximum entropy models are trained.

4.2 Error characters selection

Each character is associated with the characters previous or next, so if a target character with the character before or after it together appear in the test corpus, they are highly likely to appear in the training data. Then the target character would be highly classified into the right character category. Conversely, if a target character with the character before or after it together could not be found in the training data, the target character would be highly classified into the wrong category.

Affected by the incorrectly written character, even though the characters before and after it are correctly written, they all may be classified into wrong character category, for they are missed with the incorrectly written character in the training corpus. In the same way, if a certain character is classified into wrong character category while the characters before and after it are all classified into the right character category, it is highly likely mistakenly classified. We need to set thresholds to judge whether the characters are really incorrectly written or mistaken classified in the above two situations:

(a) To the situation that continuous two or more characters are classified into wrong character category, if all the calculated probabilities of the wrong character category of these characters are over the threshold X1, they will be treated as incorrectly written characters.
(b) To the situation that a single character is classified into wrong character category while the characters before and after it are all classified into the right character cate-

[1] Download from https://github.com/lzhang10/maxent/

gory, if the calculated probabilities of the wrong character category of the single character is over the threshold X2, it will be treated as incorrectly written characters.

In our experiment, we find that if the threshold X1 is set to 0.95 and the threshold X2 is set to 0.99, most of the characters incorrectly written can detected.

Though we set thresholds above, there are still too many mistaken classified characters. We need to set each character an accurate threshold, forming a cutoff table to filter out the mistaken classified characters.

We use the maximum entropy toolkit to classify the characters in the Dry-Run test set data, and achieve all the calculated probabilities of the wrong character category of incorrectly written characters. We calculate the mean probabilities of the wrong character category X, and set the smallest probability higher than X of each character as the threshold of the character. In our experiment, the X we calculated is 0.977.

As the number of the incorrectly written characters in the Dry-Run test set data is limited, we couldn't get all the probabilities of the characters. In order to avoid these characters mistaken classified as much as possible, a relatively high threshold is set. In our experiment, the threshold of it is set to 0.9999.

Corresponding to the 5311 characters in the experiment, we have 5311 characters thresholds. Using the cutoff table, we could achieve a better result on Chinese spelling check.

4.3 Experimental results

Spelling check performance is evaluated by F-score F=2RP/(R + P). The recall R is the ratio of the correctly identified spelling error sentences of the checker's output to all spelling error sentences in the gold-standard and the precision P refers to the ratio of the correctly identified spelling error sentences of the checker's output to all identified error sentences of the checker's output. Moreover, False-Alarm Rate and Detection Accuracy are also introduced to evaluate spelling check. The former is the ratio of the checker's output to all spelling error sentences with false positive error detection results to testing sentences without errors in the gold-standard, and the latter is the ratio of the checker's output to all spelling sentences with correctly detected results to all testing sentences.

Table 1: Performance of the final test

False-Alarm Rate	0.3986
Detection Accuracy	0.678
Detection Precision	0.4795
Detection Recall	0.8567
Detection F-score	0.6149
Error Location Accuracy	0.5
Error Location Precision	0.1474
Error Location Recall	0.2633

From the result, we achieve a relative better Detection Recall. As the maximum entropy can storage the knowledge of characters appearing together, most of the illegal continuous characters can be detected, and they are highly likely incorrectly written characters.

However, the Detection Precision is relative not high, as the maximum entropy mistaken classifies many single characters with high probabilities of the wrong character category such as "我", "的", "是", "不", "在" and so on. These characters are high frequency characters, almost appearing in every sentence. Even though the maximum entropy can classify over 99% of these characters correctly, the rest 1% mistaken classified would pull down the Detection Precision.

5 Conclusion

In this paper, we propose a maximum entropy method in Chinese spelling check. As the maximum entropy can storage the knowledge of characters appearing together, most of the illegal collocation can be detected. It also grows the problem that it could not handle the high frequency characters well, which affects the spelling check result a lot.

It is our first attempt on Chinese spelling check, and tentative experiment shows we achieve a not bad result. We don't use lexicon-based methods, easy to operate is the merit of our simple method.

However, we still have a long way from the state-of-arts results. Much work needs to be done, and further refinement and methodology combinations seem still needed. We need to find a better way to solve the problems of high frequency characters. In this work, we ignore the association of the n-grams formed by continuous characters. We need to explore a better way to train them. We also need to probe into other machine learning classifying tools, like Support Vector Machine (SVM).

6 Acknowledgment

This work is supported by National Natural Science Foundation of China under Grant No. 61273318.

Reference

Chao-Huang Chang. 1995. *A new approach for automatic Chinese spelling correction*, Proceedings of Natural Language Processing Pacific Rim Symposium: 278-283.

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.

Chuen-Min Huang, Mei-Chen Wu and Ching-Che Chang. 2007. *Error detection and correction based on Chinese phonemic alphabet in Chinese text*. Proceedings of the fourth conference on Modeling Decisions for Artificial Intelligence (MDAIIV). Springer Berlin Heidelberg: 463-476.

Damerau F.J., 1964. *A technique for computer detection and correction of spelling errors*. Communication of the ACM, 7(3):171-176.

Fuji Ren, Hongchi Shi and Qiang Zhou. 1994. *A hybrid approach to automatic chinese text checking and error correction*. Proceedings of the ARPA Workshop on Human Language Technology: 76-81.

Lei Zhang, Ming Zhou, Changning Huang, Lu Mingyu. 2000a. *Approach in automatic detection and correction of errors in Chinese text based on feature and learning* (In Chinese). Proceedings of the 3rd world congress on Intelligent Control and Automation, Hefei: 2744–2748.

Lei Zhang, Ming Zhou, Changning Huang, Haihua Pan. 2000b. *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.

Lei Zhang, Ming Zhou, Changning Huang, Maosong, Sun. 2000c. *Automatic Chinese text error correction approach based-on fast approximate Chinese word-matching algorithm*. In Intelligent Control and Automation, Proceedings of the 3rd World Congress on IEEE, 4: 2739-2743.

Shih, D.S. et al., 1992. *A statistical method for locating typo in Chinese sentences*. CCL Research Journal (8):19-26.

Yih-jeng Lin, Feng-long Huang, Ming-shing Yu. 2002. *A chinese spelling error correction system*. Proceedings of seventh conference on artificial intelligence and applications.