Conditional Random Field-based Parser and Language Model for Traditional Chinese Spelling Checker

Yih-Ru Wang
National Chiao Tung University
HsinChu, Taiwan
yrwang@mail.nctu.edu.tw

Yuan-Fu Liao
National Taipei University of Technology, Taipei, Taiwan
yfliao@ntut.edu.tw

Yeh-Kuang Wu
Institute for Information Industry
Taipei, Taiwan
121590@gmail.com

Liang-Chun Chang
Institute for Information Industry
Taipei, Taiwan
lcchang@iii.org.tw

Abstract

This paper describes our Chinese spelling check system submitted to SIGHAN Bake-off 2013 evaluation. The 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. Experimental results on the Bakeoff 2013 tasks showed the proposed method achieved 0.50 location detection and 0.24 error location F-scores in subtask1 and 0.49 location and 0.40 correction accuracies and 0.40 correction precision in subtask2.

1 Introduction

Chinese spelling check is a difficult task for two reasons: (1) there are no word delimiters between words, (2) the length of each word is usually only one to three characters long. So it cannot be done within the word and must be solved within a context. Therefore, Chinese spell checking is usually divided into two steps: (1) segmentation of text into word sequence and (2) error checking of each word in sentence level.

Basically, word segmentation can be formulated as a sequential learning problem. In the past decade, many statistical methods, such as support vector machine (SVM) (Zhang, 2010), conditional random field (CRF) (Zhao, 2006), maximum entropy Markov models (MEMMs) (Berger, 1996), were proposed by NLP researchers to handle this sequential learning task. Among them, CRF-based approach has been shown to be effective with very low computational complexity.

On the other hand, error checking could be treated as an abnormal word sequence detection problem and is often based on language knowledge, and mainly includes rule-based methods and statistic-based methods. Rule-based methods use rule sets, which describe some exact dictionary knowledge such as word or character frequency, POS information and some other syntax or morphological features of a language, to detect dubious areas and generate candidate words list. This kind of methods achieves significant success in some special domains, but it is difficult to deal with open natural language. On the other hand, statistic-based methods often use a language model that is achieved by using some language knowledge and analyzing a huge of language phenomena on large corpus so more context information is utilized, and this kind of methods is suitable for general domains.

There are many advanced Chinese spelling check methods (Liu, 2011 and Chen, 2011). However, from the viewpoint of automatic speech recognition (ASR) research, the word segmentation and LM are the most important modules for ASR studies. Especially, it is known that a good LM can significantly improve ASR's recognition performance. And a sophisticated parser is required for building highly effective LM. So, in past few years, lots of works were conducted in our laboratory to build a CRF-

Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), pages 69–73,
Nagoya, Japan, 14 October 2013.

based word segmentation/POS tagger and a tri-gram LM to improve the performance of ASR.

Although, we have already applied our parser and LM to ASR and achieved many successes (Chen, 2012), we would like to take the chance of Bakeoff 2013 evaluation to examine again how generalization and sophistication our parser and LM are. Therefore, the focus of this paper is on how to integrate our parser and LM originally built for ASR to deal with the Chinese spelling check task.

2 The Proposed Framework

The block diagram of the proposed method is shown in Fig. 1. Our main idea is to exchange potential error character with its confusable ones and rescore the modified sentence using our CRF-based parser and tri-gram LM to detect and correct possible spelling errors.

In this scheme, the input text is first checked if there are some high frequency error words in the rule-based frontend. The sentence is then segmented into a word sequence using our CRF-based parser and scored with tri-gram LM. Each character in short words (less than 3 characters) is considered as potential error character and is replaced with characters that have similar shape or pronunciation. The modified sentence is then re-segmented and rescored to see if the score of the changed sentence is higher. This process is repeated until the best sentence with maximum LM score is found.

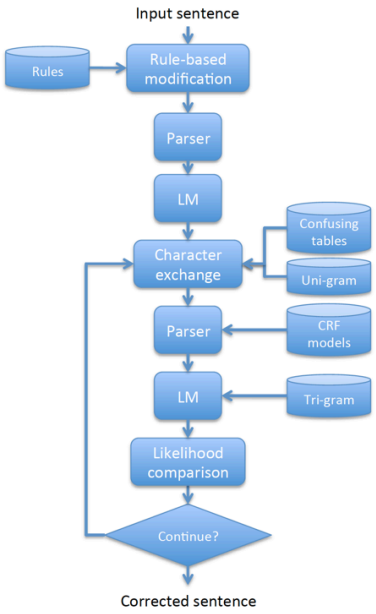

Fig. 1: The schematic diagram of the proposed Chinese spelling checker.

2.1 Rule-based Frontend

Basically, the rule-based spelling error correcting was easy and will not increase the complexity of our parser. In our parser, only the rules with high accuracy and low false alarm were added to the parser. It can also increase the accuracy of our parser.

There are about 600 high frequency error words in our database. Those words are basically collected from Internet. The rule to replace error words is in general as follows:

(1) Direct spelling errors correcting: Most of those cases are frequently error words, some interesting examples of the rules are:

- 倉惶 → 倉皇
- 翹課 → 蹺課
- 百摺裙 → 百褶裙
- 經不起 → 禁不起
- 明查秋毫 → 明察秋毫

It can be seen from those examples that some errors are due to misunderstanding of the meaning of words. Since these errors are often unconsciously replaced with other high-frequency characters, it is usually difficult to detect and corrected using LM.

(2) Errors correcting with constraints: In this case, the word *XX,* usually two characters, will be corrected to *YY*, with some constraints. We need to check if the *XX* is cross word boundary. If p_i is preceding character and P_i succeeding character of *XX*, and the p_i-*XX* can be segment into p_i*X*-*X* or *XX*-P_i can be segment into *X*-*X*P_i. In order to avoiding false alarm, the constraints were checked before the correcting.

For example: 一但 → 一旦, but the preceding character p_i is not 統, or the succeeding character P_i is not 書.

(3) Spelling errors correcting after parsing: Some frequently happened spelling errors were difficult to correction without the word segmentation information. The error words were added in the lexicon of parser in order to get the corrected word segmentation. And, the error words were correcting after parsing.

2.2 CRF-based Chinese PARSER

A block diagram of the proposed Chinese parser is shown in Fig. 2. There are three blocks including (1) text normalization, (2) word segmentation and (3) POS tagging. The last two modules are briefly described as follows.

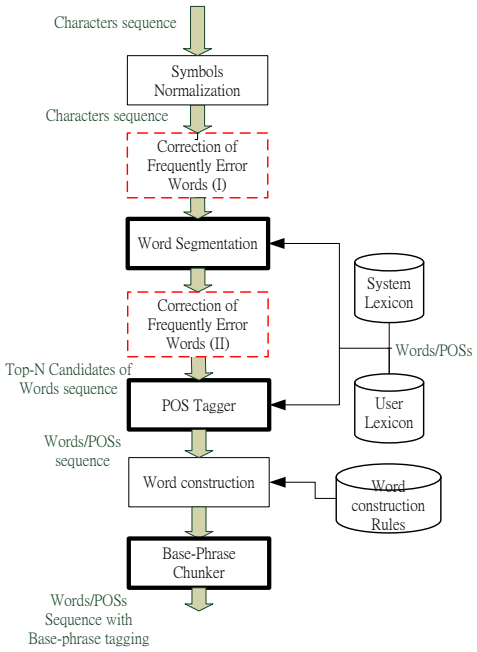

Fig. 2: The schematic diagram of the proposed Chinese parser.

2.2.1 Word Segmentation

Basically, it is based on CRF method and implemented following Zhao's work (Zhao, 2006). Six tags, denoted as B1, B2, B3, M, E and S, are used to represent the activated functions. The information used in feature template is listed in the following:

- C_n: Unicode of the current character (Unicode plain-0 only).
- B_n: radical of the current character ("bushu", 部首).
- SB_n: if B_n is equal to B_{n-1}.
- WL_n: length of the maximum-length word in lexicon that matches the string including the current character. Here, the 87,000-word lexicon released from Sinica (Sinica Chinese Electronic Dictionary[1]) is used as the basic internal lexicon. A user-defined lexicon is allowed to define more words, and in most cases they are named entities.
- WT_n: tags of the maximum-length word comprising the current character (indicating character position in word by B1, B2, B3, M, E, S).
- D/E_n: indicator showing whether the current character is a digit.
- PM_n: 0-1 tags to indicate whether the current character is a punctuation mark (PM).

Moreover, the CRF templates used in the word segmentation are shown in Table 1.

Information	Templates
Character n-gram	Cn-2, Cn-1, Cn, Cn+1, Cn+2, (Cn-2 Cn-1 Cn), (Cn Cn+1 Cn+2), (Cn-1 Cn Cn+1), (Cn-2 Cn-1 Cn Cn+1 Cn+2)
Digits/English	(D/En-1 D/En), (D/En D/En+1), (D/En-1 D/En D/En+1)
Bushu	(Bn BSn), (Bn-1 BSn-1), (Bn+1 BSn+1)
Tag of candidate word	WTn-1, WTn, WTn+1, (WTn-1 WTn), (WTn WTn+1), (WTn-1 WTn WTn+1)
Length of candidate word	WLn-1, WLn, WLn+1, (WLn-1 WLn), (WLn WLn+1), (WLn-1 WLn WLn+1)
Length/tag of candidate word	(WTn WLn), (WTn-1 WLn-1 WTn WLn), (WTn WLn WTn+1 WLn+1)
Repeated word	(LWn SW1n), (LWn SW2n)
PM	PMn-1, PMn, PMn+1

Table 1: List of CRF templates for word segmentation.

The word segmentation module is trained by using Sinica Balanced Corpus version 4.02. Before training the word segmentation CRF, data in the corpus are check to correct in-consist word-segmentation. More than 1% of data in the corpus are corrected manually.

The protocol of consistency check is described here. The unigram and bigram of Sinica Balanced corpus are first generated. Then all pairs of words, excepting the words with POS of "Nf" and "Neu", are checked to see whether they can be combined into a single word. There are about 10% of such word-pairs. For example:

(1) For the case that both a word-pair (e.g. 民意(Na) 代表(Na)) and the combination word (e.g. 民意代表(Na) appear in the corpus, we divide the combination word into two words.

(2) For a word-pair (e.g. /長途(A) 電話(Na)/) whose combination does not appear as a single word in the corpus but is a word entry (e.g. /長途電話(Na)/) in the Sinica lexicon, we keep the two words and remove the combination word from the lexicon.

(3) Most of the bound morphemes (i.e., prefixes and suffixes), named entities, compound words, idioms, and abbreviations in the corpus were checked for consistency.

(4) Some words, especially for function words, have different POSs and can be divided into smaller words, like "就是(T), 就是 (SHI), 就

[1] http://www.aclclp.org.tw/use_ced.php
[2] http://www.aclclp.org.tw/use_asbc.php

[2] http://www.aclclp.org.tw/use_asbc.php

是(Nc), 就(D) 是(SHI), 就是(D), 就是(Cbb)" and "真是(VG), 真是(D), 真(D) 是(SHI)]". Some of them need to be corrected according to the syntactic and/or sematic context in the sentence.

The corpus is divided into two parts: a training set containing 90% of the corpus (about 1 million words including PMs) and a test set containing 10% (about 120K words including PMs). The training set is used to train the word segmentation CRF. The F-measure of the word segmentation is 96.72% for the original database and 97.50% for the manually correct one. The difference between precision and recall rates is less than 0.1%. If all PMs are excluded, the F-measure reduces to 97.01%.

2.2.2 POS Tagger

Here is the features used in the CRF method:

- PM_n: Unicode of the first character of the current word when it is a PM, or "X" if it is not a PM. We note that some PMs, such as "?!" and "…", are formed by string of more than one character.
- WL_n: word length of the current word.
- $LPOS_n$: all possible POSs of the current word if it is in the internal or external lexicons, or "X" if it is not in those lexicons, e.g. the word "一"(one) can be "Cbb_Di_D_Neu".
- FC_n: first character of the current word if it is not in lexicon, or "X" if it is in lexicon.
- LC_n: last character if the word is not in lexicon, or "X" if it is in lexicon.

Table 2 shows the CRF templates used for POS tagging.

Information	Templates
Possible POS n-gram	$LPOS_{n-2}$, $LPOS_{n-1}$, $LPOS_n$, $LPOS_{n+1}$, $LPOS_{n+2}$, ($LPOS_{n-1}$ $LPOS_n$), ($LPOS_{nn}$ $LPOS_{n+1}$), ($LPOS_{n-1}$ $LPOS_{n+1}$)
PM	PM_{n-1}, PM_n, PM_{n+1}
Information of OOV word	(WL_n FC_n), (WL_n LC_n)

Table 2: List of CRF templates for POS tagging.

The POS tagger is trained by using the same training set used in the word segmentation. In the test, the POS tagger processes the top-N output sequences of the word segmentation. It combines the log-likelihood scores of word segmentation and POS tagging to find the best output word sequenc. The accuracy of the 47-type POS tagging is 94.22%. The performance is reasonable except "*Nv*".

2.3 Language Modeling

For constructing the LM, two corpora, the Sinica Balanced Corpus CIRB030 (Chinese Information Retrieval Benchmark, version 3.03), the Taiwan Panorama Magazine[4] and the Wikipedia (zh-version, 2013/04/20), containing 440 million words totally, are parsed.

Some post-processing are done on the parsed text database, including

(1) text normalization,
(2) segment long number (the word with POS '*Neu*') into short number strings,
(3) change the hyphen between number and date (the word with POS '*Nd*') into "至" (to) to make the text readable,
(4) change some variation words (Here, variation word means a word have different written forms).

Finally, A lexicon with 100K words is used to build the LM. The coverage rate of the lexicon is about 97%.

3 Bakeoff 2013 Evaluation Results

3.1 Task

The task is divided into two sub-tasks including (1) error detection and (2) error correction. For the error detection sub-task, the system should return the locations of the incorrect characters. For the error correction sub-task, the system should return the locations of the incorrect characters, and must point out the correct characters. Moreover, one Sample Set (selected from students' essays) and two Similar Character Set (abbrev. Bakeoff 2013 CSC Datasets) are provided for this evaluation. There are two test data sets for the evaluation. Each set contains 1000 Chinese texts selected from students' essays.

3.2 Evaluation Results

Two configurations of our system (Run1 and Run2) were tested. Run1 applied only the rule-based frontend. Run2 utilized the whole system. The performances of the proposed spelling check method are shown in Table 3 and 4.

From Table 3, it can be found that Run1 has very low false alarm and recall rates, but higher accuracy in error detection. The reason is that it only modified few errors with high confidence.

[3] http://www.aclclp.org.tw/use_cir.php
[4] http://www.aclclp.org.tw/use_gh_c.php (in Chinese)

Run2 has much higher false alarm and recall rates, but lower accuracy, since it tried to change as much as possible errors and may introduce overkill. However, in general, Run2 has better F-score than Run1. Furthermore, Table 4 also shows that Run2 has higher location and correction accuracies (although it has lower correction precision than Run1). These results show the benefits of combining CRF-based parser and LM in the second stage of spelling check system.

Error	False-Alarm	Accuracy	Precision	Recall	F-score
Run1	0.0243	0.722	0.6964	0.13	0.2191
Run2	0.8329	0.411	0.3352	0.98	0.4995

(a)

Error Location	Accuracy	Precision	Recall	F-score
Run1	0.711	0.5	0.0933	0.1573
Run2	0.257	0.1596	0.4667	0.2379

(b)

Table 3: Evaluation results of the proposed system on Bakeoff 2013 sub-task 1: (a) detection error rates, (b) location error rates on 1000 test sentences.

	Location Accuracy	Correction Accuracy	Correction Precision
Run1	0.07	0.065	0.5118
Run2	0.485	0.404	0.404

Table 4: Evaluation results of the proposed system on Bakeoff 2013 sub-task 2. There are 1000 test sentences.

3.3 Error Analysis

Here are some examples that show the typical overkill behaviors of the proposed system ("O" original, "M" modified):

O: 人生 是 需要 巨浪 激出 美麗 的 浪花
M:人生 是 需要 巨浪 洗出 美麗 的 浪花
O: 很 難 感受到 快樂 的 人
M: 很 難看 受到 快樂 的 人

In brief, it was found that the most overkill errors are due to the out of vocabulary (OOV) problem. Especially, in the above three cases, the outputs of parser are in fact correct but unfortunately, the LM didn't recognize "激出" and "感受到" and our system gave them high penalties.

4 Conclusions

In this paper, a Chinese spelling check approach that integrating our CRF-based parser and LM originally built for ASR is proposed. Experimental results on the Bakeoff 2013 tasks confirmed the generalization and sophistication of our parser and LM. The work to improve our traditional Chinese parser and LM is still continued. Our latest Chinese parser is available online at http://parser.speech.cm.nctu.edu.tw.

Acknowledgments

This work was supported by the National Science Council, Taiwan, Republic of China, under the project with contract NSC 101-2221-E-009-149-MY2, 101-2221-E-027-129 and under the "III Innovative and Prospective Technologies Project" of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

References

A. L. Berger, Stephen A. D. Della Pietra, and V. J. Della Pietra, 1996, A maximum entropy approach to natural language processing, Computational Linguistics, 22(1): 39-71.

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 Trans. Asian Lang. Inform. Process. 10, 2, Article 10.

Chongyang Zhang, Zhigang Chen, Guoping Hu, 2010, A Chinese Word Segmentation System Based on Structured Support Vector Machine Utilization of Unlabeled Text Corpus, Joint Conference on Chinese Language Processing

H. Zhao, C. N. Huang and M. Li, 2006, An Improved Chinese Word Segmentation System with Conditional Random Field, the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 108-117.

S. H. Chen, J. H. Yang, C. Y. Chiang, M. C. Liu, and Y. R. Wang, 2012, A New Prosody-Assisted Mandarin ASR System, IEEE Trans. on Audio, Speech and Language Processing, Vol. 20, No, 6, pp. 1669-1684.

Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, 2003, A neural probabilistic language model, Journal of Machine Learning Research, No. 3(2), pp. 1137–1155.

Yong-Zhi Chen, Shih-Hung Wu, Ping-che Yang, Tsun Ku, and Gwo-Dong 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, Vol. 21, No. 1, pp.103-116.