Word Vector/Conditional Random Field-based Chinese Spelling Error Detection for SIGHAN-2015 Evaluation

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

Abstract

In order to detect Chinese spelling errors, especially for essays written by foreign learners, a word vector/conditional random field (CRF)-based detector is proposed in this paper. The main idea is to project each word in a test sentence into a high dimensional vector space in order to reveal and examine their relationships by using a CRF. The results are then utilized to constrain the time-consuming language model rescoring procedure. Official SIGHAN-2015 evaluation results show that our system did achieve reasonable performance with about 0.601/0.564 ac-curacies and 0.457/0.375 F1 scores in the detection/correction levels.

1 Introduction

Chinese spelling check could be treated as an abnormal word sequence detection and correction problem. Convention approaches to do this job often heavenly rely on a language models (LM) trained from a large text corpus (for example Chinese Gigaword[1]) to find potential errors and provide suitable candidate words (Bengio 2003, Wang 2013) to replace them. These approaches usually could be successfully applied to examine essays written by Chinese element or junior school students.

However, for essays written by foreign learners, conventional LM methods may not be so helpful. Because, the writing behaviors of foreign learners are usually different with native Chinese writers. They may embedded spelling errors into rarely used word sequences (low LM scores, but are somehow grammar or syntactic corrected). For example:

- 然後你們工廠應該要蓋起來比較高高厚厚的床比。(“床比” should be “牆壁”)
- 小孩不過不知道那個好那個不好，也不知道那作法是適合，難怪常常看到他們用部是對的。(“部” should be “不”)
- 王大明今天六點半起來就洗澡穿上就去廚房吃早飯他等公車十分就坐上，他坐著坐著到學校來了。(“穿上” should be “穿衣”)

They may also produce some semantic errors (but are all grammar and syntactic corrected and therefore with high LM scores). This type of errors are difficult, if not impossible, to detect using only LM models trained from conventional Chinese text corpora. For example:

- 吃了碗飯以後，我們兩個人馬上去看電影。(“碗” should be “晚”)
- 他戴著眼鏡跟襪子入睡了。(“襪子” should be “穿著襪子”)
- 我跟我的同學學數學。我們對號碼有興趣。(“號碼” should be “數字”)

In order to properly deal with those errors, it is necessary to understand foreign learners’ writing behaviors. Therefore, this paper focus on how to automatically learn the behaviors of foreign learners. Our major idea is to transform the problem into a machine learning task. To this end, the vector representations of the words were first constructed and then CRF-based approach was adopted to detect the errors.

2 Overview of the proposed system

The block diagram of our system is shown in Fig. 1. There are four main components including (1) a misspelling correction rules frontend, (2) a CRF-based parser, (3) a word vector/CRF-based spelling error detector and (4) a 120k tri-gram LM.

Basically, our approach is to utilize the error detection results to guide and speed up the time-consuming LM rescoring procedure. It iteratively exchanges potential error words with their confusable ones and examine the modified sentence using the tri-gram LM. The final goal is to produce a modified sentence with maximum LM

[1] http://www.aclclp.org.tw/use_asbc_c.php

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 46–49,
Beijing, China, July 30-31, 2015. ©2015 Association for Computational Linguistics and Asian Federation of Natural Language Processing

score. By this way, potential Chinese spelling errors could be detected and corrected.

Since, the details of our parser, LM modules and character replacement procedure could be found in (Wang 2013), only the newly added word vector/CRF-based error detection module will be further described in the following subsections.

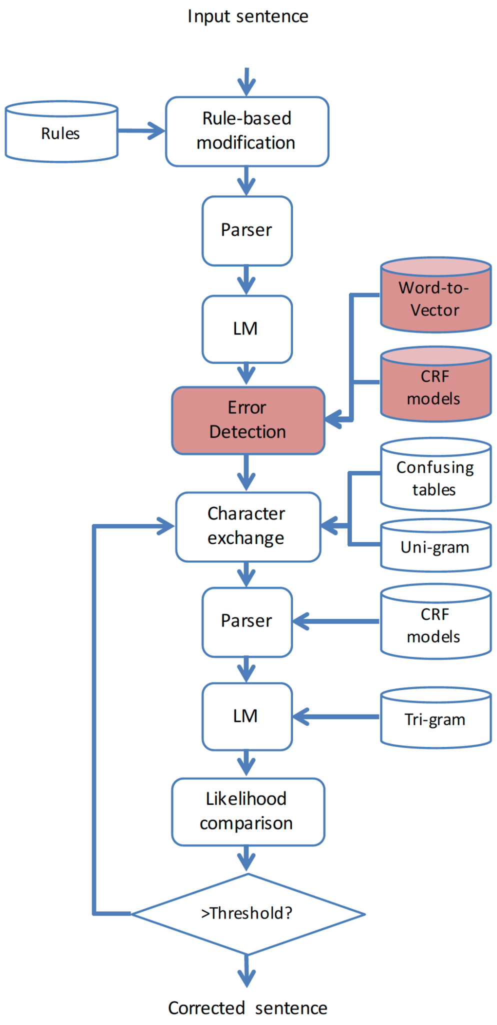

Fig. 1: The schematic diagram of the proposed Chinese spelling checker. The are four modules including a rule-based frontend, a CRF-based parser, a tri-gram LM and a word vector/CRF-based spelling error detector. Among them, the spelling error detector is newly added for SIGHAN-2015 evaluation.

3 Word Vector/CRF-based Spelling Error Detector

Fig. 2 shows the block diagram of the word vector/CRF-based Chinese spelling error detection module. Its two main modules, i.e., word2vec and CRF will be discussed in the following subsections.

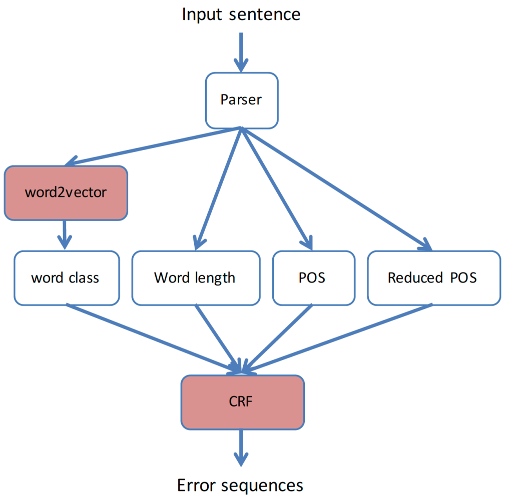

Fig. 2: The schematic diagram of the proposed Chinese spelling error detector. The input features of the CRF includes word classes tagged by word2vec, length, POS and reduced POS provided by parser module.

3.1 Word vector representation

The word to vector algorithm proposed by Tomas Mikolov (Mikolov 2013a, 2013b) is adopted in this paper to encode words. It uses the CBOW (continuous bag of words, as shown in Fig. 3) representations to project each word into a high dimensional vector space.

These representations have been shown to be capable to capture deep linguistic information beyond surface words (Mikolov 2013). Therefore, CBOW is used here to reveal the prosperities and relationship between normal and abnormal word sequences.

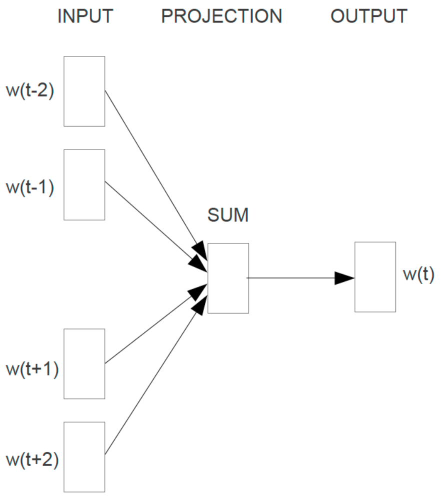

Fig. 3: The CBOW word to vector encoding architecture that predicts the current word based on the context.

3.2 CRF Chinese spelling error detector

To detect potential spelling errors, the word vectors and parser outputs are further combined into a feature sequences for CRF error detector. CRF then learns from a set of labels samples (ground-truth) to tell between correct and incorrect word spellings instances.

Fig. 4 shows a typical example of the extracted feature sequences of a training sample. Here each word is transformed into a 5 dimensional vector including (1) the length of the word, its (2) POS and (3) reduced POS tags, (4) the word class index and the ground-truth (correct or error spelling) labels.

聽起來	3	D	ADV	436	c
是	1	SHI	Vt	441	c
一	1	Neu	DET	136	c
份	1	Nf	M	162	c
很	1	Dfa	ADV	441	c
好	1	VH	Vi	398	c
的	1	DE1	T	390	c
公司	2	Nc	N	609	c
。	1	PM	M	-2	c
又	1	Caa	C	551	w
意思	2	Na	N	77	c
又	1	Caa	C	551	c
很多	2	Neqa	DET	441	c
錢	1	Na	N	270	c
。	1	PM	PM	-2	c

Fig. 4: A typical example of a training sample (from left to right) the word segmentation results and the corresponding input features (word length, POS, reduced POS and word class index) and ground-truth labels.

4 Evaluation Results

4.1 System setting

Basically, the parser, 120K tri-gram LM and word vector representation were all trained using Sinica Balanced Corpus version 4.0[2]. There is in total about 4.4 billion words in the corpus.

For the parser, its F-measure of the word segmentation is 96.72% and 97.67% for the original and manually corrected corpus. The accuracy of the 47-type POS tagging is about 94.24%. To build the word vector representation, a window of 17 (8+1+8) words was used. Each word was first projected into a 200 dimensional CBOW vector and then further clustered into one of 1024 classes.

On the other hand, to build the CRF-based spelling error detector, Bake-off 2014 and SIGHAN-2015 development corpora were utilized. There are in total 106,815 words in the training set. Among them, only 4,537 words are incorrect. For the test set, there are 11,808 words including 498 errors.

4.2 Error detection frontend results

First of all, Fig. 5 shows a typical output of the word vector/CRF-based spelling error detector. It is worth to note that the last column in Fig. 5 shows the correct scores reported by the CRF. If the scores are less than 0.5, the corresponding words will be treated as good ones, otherwise spelling errors will be reported. For example, the last word “阿” has a very low score 0.0048 and is therefore will be labelled as an error.

```
但是  2    Cbb   C     441   0.9999
我    1    Nh    N     738   0.9998
不能  2    D     ADV   441   0.9833
去    1    D     ADV   738   0.9945
參加  2    VC    Vt    723   0.9985
，    1    PM    PM    -2    0.9998
因為  2    Cbb   C     441   0.9999
我    1    Nh    N     738   0.9999
有一點 3   Dfa   ADV   738   0.9997
事情  2    Na    N     441   0.9687
阿    1    T     T     820   0.0048
！    1    PM    PM    -2    0.9999
```

Fig. 5: A typical example of the CRF outputs. The last column shows the scores given by the CRF’s correct spelling nodes.

Moreover, Table 1 show the evaluation results of the error detection frontend on Bake-off 2014 and SIGHAN-2015 development corpora. From the table, it can be found that the detection results for the training set is quite good. But for test set, there is serious bias issue. This may due to the over-fitting problem since there are unbalanced numbers of correct and incorrect spelling word samples in the training set. To alleviate the difficulties, we will try to lower detector’s decision threshold for the following LM rescoring procedure to cover more hypotheses.

4.3 Overall detection and correction results

Finally, three system configurations (Run1~3) were tested to explore different LM rescoring space. i.e., using three different CRF score thresholds including 0.999, 0.98 and 0.95.

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

Among them, the search space of Run1 is very restricted and Run3 is much larger than others.

Table 2 show the official evaluation results given by the SIGHAN-2015 evaluation organizer. From Table 2, it can be found that Run1 had lowest false positive and recalls rates in both measures. On the other hand, Run3 had highest recall rates and F1 scores but produced many more false alarms.

In summary, these results show that our approach had achieved reasonable performance. But the settings of our systems (even Run3) were still too conservative. Therefore, there are still some rooms to further lower the threshold in order to improve the F1 scores.

		Acc.	Pre.	Rec.	F1
Training	C		99.92	99.98	99.95
	W		99.21	97.47	98.33
	All	99.90	99.90	99.90	99.90
Test	C		98.23	99.03	98.63
	W		54.10	38.98	45.31
	All	97.32	97.32	97.32	97.32

Table 1: Detection results of the proposed word vector/CRF-based error detector on Bake-off 2014 and SIGHAN-2015 corpora. The table shows the accuracy (Acc.), precision (Pre.), recall (Rec.) and F1 score for both the training and test sets (C: correct, W: wrong words).

Run	F/P	Detection Level			
		Acc.	Pre.	Rec.	F1
1	0.050	0.605	0.837	0.261	0.398
2	0.065	0.609	0.812	0.283	0.420
3	0.132	0.601	0.717	0.336	0.457
Run	F/P	Correction Level			
		Acc.	Pre.	Rec.	F1
1	0.050	0.578	0.802	0.207	0.329
2	0.065	0.580	0.776	0.227	0.351
3	0.132	0.564	0.663	0.261	0.375

Table 2: Official evaluation results of the proposed systems for SIGHAN-2015 Chinese spelling check task. The table shows the false positive (F/P) rate, accuracy (Acc.), precision (Pre.), recall (Rec.), and F1 score for both the detection and correction levels.

5 Conclusions

In this paper, a word vector/CRF-based Chinese spelling error detector have been newly added to improve our spelling check system. Evaluation results show that our systems had achieved reasonable performance. Especially, configuration Run3 achieves about 0.601/0.564 accuracies and 0.457/0.375 F1 scores in the detection/correction level, respectively.

Experimental results also showed that our error detector frontend suffered serious overfitting problem. Beside, the time consuming LM scoring procedure should be replaced with a candidate word predictor (for example the CBOW structure shown in Fig. 3). These two issues will be further studied in the future. Finally, our latest traditional Chinese parser is available on-line at http://parser.speech.cm.nctu.edu.tw.

Acknowledgments

This work was supported by the Ministry of Science and Technology, Taiwan, under the projects with contract MOST 103-2221-E-027-079 and MOST 103-2221-E-009-125-MY2.

References

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

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 (June 2011).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013b). Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

Yih-Ru Wang, Yuan-Fu Liao, Yeh-Kuang Wu and Liang-Chun Chang (2013). Traditional Chinese Parser and Language Model-Based Chinese Spelling Checker. Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing (SIGHAN'13), Nagoya, Japan, 14 October, 2013, pp. 69-73.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee (2013). Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013. Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing (SIGHAN'13), Nagoya, Japan, 14 October, 2013, pp. 35-42.

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 2006, pp. 108-117.