A Hybrid Chinese Spelling Correction Using Language Model and Statistical Machine Translation with Reranking

Xiaodong Liu, Fei Cheng, Yanyan Luo, Kevin Duh and Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{xiaodong-l,fei-c,yanyan-l,kevinduh,matsu}@is.naist.jp

Abstract

We describe the Nara Institute of Science and Technology (NAIST) spelling check system in the shared task. Our system contains three components: a word segmentation based language model to generate correction candidates; a statistical machine translation model to provide correction candidates and a Support Vector Machine (SVM) classifier to rerank the candidates provided by the previous two components. The experimental results show that the k-best language model and the statistical machine translation model could generate almost all the correction candidates, while the precision is very low. However, using the SVM classifier to rerank the candidates, we could obtain higher precision with a little recall dropping. To address the low resource problem of the Chinese spelling check, we generate 2 million artificial training data by simply replacing the character in the provided training sentence with the character in the confusion set.

1 Introduction

Spelling check, which is an automatic mechanism to detect and correct human spelling errors in every written language, has been an active research area in the field of Natural Language Processing (NLP). However, spelling check in Chinese is very different from that in English or other alphabetical languages. First because there are no word delimiters between the Chinese words; moreover, the average length of a word is very short: usually one to four characters. Therefore, error detection is a hard problem since it must be done within a context, say a sentence or a long phrase with a certain meaning, and cannot be done within one word. For instance, in the words "自己"(self-control) and "自己"(oneself), the character "己" or "已" cannot be detected as an error without the context. Other challenge in the Chinese spelling check is that there is no commonly available data set for this task and the related resource is scarce.

The SIGHAN 2013 shared task is to provide a common evaluation data set to compare the error detection and correction rates between different systems. The evaluation includes two sub-tasks: 1) error detection and 2) error correction.

In this paper, we present a system that combines the correction candidates produced by the language model based method and the statistical machine translation approach, and then uses an SVM classifier to rerank the correction candidates. To address the low resource problem, firstly, we generate around 2 million artificial sentences following a simple rule, which replaces each character in the provided 700 sentences with the character in the confusion set to generate a new training corpurs; secondly, we use unlabeled data corpus, the Chinese Gigaword, to train a language model[1] to estimate the real Chinese texts.

The paper is organized as follows. We first briefly discuss the related work in Section 2 and overview of our system structure in Section 3. Subsections 3.1, 3.2 and 3.3 describe the components of our system respectively. In Section 4, we discuss the experiment setting and experimental results. Finally, we give the conclusions in the final section.

2 Related work

In Chinese spelling check, the confusion sets are collections of candidate error characters, and play a crucial role.

Chang (1995) manually edited confusion sets from 4 viewpoints, i.e., shape, pronunciation, meaning and input keystroke sequence. Then by

[1] We use the SRI Language Modeling Toolkit adopting the interpolated Kneser-Ney smoothing method.

Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), pages 54–58,
Nagoya, Japan, 14 October 2013.

substituting each character in the input sentence with the characters in the corresponding confusion set, they use a language model to generate a plausibility score to evaluate each possible substituted sentence. Because of the importance of confusion sets, some researchers attempted to automatically extend confusion sets by using different Chinese input methods. Intuitively, the characters with similar input key sequences are similar in shape. Zhang (2000) proposed a method to automatically generate confusion sets based on the Wubi method by replacing one key in the input key sequences of a certain character. Lin et al. (2002) used the Cangjie input method to extend confusion sets automatically.

Over the last few years, more and more models using NLP techniques were introduced into the Chinese spell check task. Huang et al. (2007) proposed a method which used a word segmentation tool to detect Chinese spelling errors. They used CKIP word segmentation toolkit to generate correction candidates (CKIP, 1999). By incorporating a dictionary and confusion sets, the system can detect whether a segmented word contains error or not. Hung et al. (2008) proposed a system which was based on manually edited error templates (short phrases with one error). For the cost of editing error templates manually, Cheng et al. (2008) proposed an automatic error template generation system. The basic assumption is that the frequency of a correct phrase is higher than the corresponding error template. Wu et al. (2010) proposed a system which implemented a translate model and a template module. Then the system merged the output of the two single models and reached a balanced performance on precision and recall.

3 System Architecture

Our system includes three components, as shown in Figure 1. Given a sentence with or without error characters, our procedure contains several steps: 1) we simultaneously generate the correction character candidates using the word segmentation based language model and the statistical machine translation model; and then 2) the SVM classifier reranks the candidates to output the most probable sentence. Each component in our system is described in Section 3.1, Section 3.2 and Section 3.3.

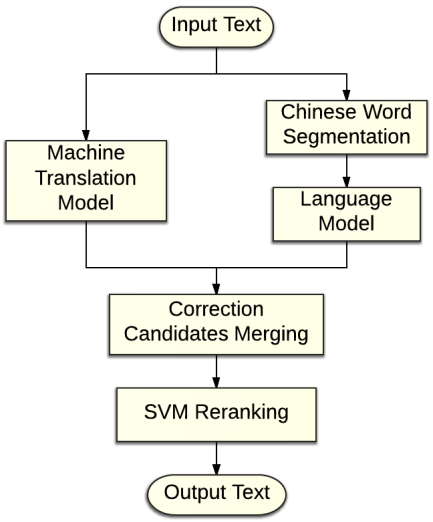

Figure 1: System structure.

3.1 Language Model Based Method

To generate the correction candidates, firstly we segment the sentence into words and then find all possible corrections based on the confusion set and a Chinese dictionary.

In this study, we use the character based Chinese word segmentation model[2] (Xue, 2003), which outperforms the word based word segmentation model in out-of-vocabulary recall. The model is trained on the Academia Sinica corpus, released under the Chinese word segmentation bake-off 2005[3] and the feature templates are the same in Sun (2011).

For example, given the following Chinese sentence (here, the Chinese character in red indicates an error character):

”我看過許多勇敢的人，不怕措折地奮鬥。”.

Firstly, we segment the sentence into words separated by a slash as follows.

”我/看過/許多/勇敢/的/人/，/不怕/措折/的/奮鬥/。”.

Secondly, we build a lattice, as shown in Figure 2, based on the following rules:

1. If a word only contains a single Chinese character, add all the candidates in the confusion set.

2. If a word contains more than one Chinese character and it is not in the dictionary, then

[2]The CRFsuite package is used in our experiment: http://www.chokkan.org/software/crfsuite/

[3]http://www.sighan.org/bakeoff2005/

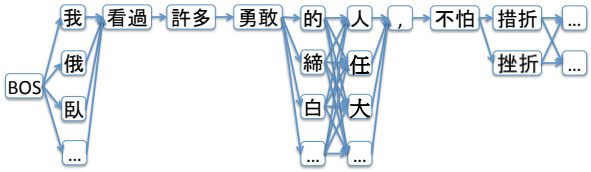

Figure 2: An example of generated candidates lattice.

replace all the characters in the word with candidates in the confusion set. If the generated word is in the dictionary, add it as a candidate.

3. If a word contains more than one Chinese character and it is in the dictionary, do nothing.

Finally, the forward algorithm (Rabiner, 1989) is used to find the k-best sentences using the n-gram language model.

3.2 Statistical Machine Translation Model

As an alternative, we also employ the statistical machine translation model as a new way to detect and correct character errors (Wu et al., 2010), which is widely used by the statistical machine translation community (Brown et al., 1993).

We treat each sentence with error as a source language. Our goal is to find the best correction sentence. Formally, given a sentence S which might contain error characters in it as a source sentence, the output is the sentence $\hat{C}$ in the target language with the highest probability of different replacement C. Symbolically, it is represent by:

$$\hat{C} = \arg\max_{c} p(C|S) \quad (1)$$

Using Bayes Rule, we can rewrite Formula 1 as:

$$\begin{aligned} \hat{C} &= \arg\max \frac{p(S|C)p(C)}{p(S)} \\ &\approx \arg\max p(S|C)p(C) \end{aligned} \quad (2)$$

Here, $p(S|C)$[4] is called "error model", which is the chance that a correct Chinese character could be written wrong, while $p(C)$ is the n-gram language model which evaluates the quality of the corrected Chinese sentence.

[4]We use GIZA++ to train the error model and Moses to decode.
https://code.google.com/p/giza-pp/
http://www.statmt.org/moses/

3.3 SVM Reranking

Support vector machines (SVMs) are supervised learning models used for classification and regression analysis (Burges et al., 1998). The goal of the Chinese spelling error detection task is to detect whether there are any errors in a given sentence, which we can treat as a binary classification problem: if the current character is an error character, the result is 0, otherwise, the result is 1. The probability output of the SVM classifier [5] can also be regarded as a confident score of how possible the current character is an error.

Given the original input text and the outputs of the other models, the system creates a candidate list for each character in the input text. Each character in the candidate list will be reranked based on the confidence score generated by the SVM classifier. The top character in the reranked candidate list will be treated as the correct character of our system. An example of SVM reranking is shown in Figure 3.

We denote a character token c_0 with a context sequence: $...c_{-2}c_{-1}c_0c_{+1}c_{+2}...$ and $c_{s:e}$ as a character sequence that starts at the position s and ends at position e. Our system creates the following features for each candidate.

- Character features: c_{-1}, c_0, c_{+1}, $c_{-1:0}$, $c_{0:+1}$.
- The pointwise mutual information (Gerlof, 2009) between two characters: $PMI(c_{-1}; c_0)$, $PMI(c_0; c_{+1})$.
- The identity of the character sequence if it exists in the dictionary and the n-gram word list. For instance: 2-character window $c_{-1:0}$, 3-character window $c_{-2:0}$, 4-character window $c_{-3:0}$, 5-character windows $c_{-4:0}$

However, the Chinese spelling check shared task provided a sample data with only 700 sentences. We split 80% as training data and 20% as test data and use 5-fold cross-validation to evaluate the SVM reranking results.

4 Experiments

4.1 Data Sets

We used two data sets in our experiments. The first data set is provided by the shared task, which

[5]LIBLINEAR with L2-regularized L2-loss support vector classification is used and optimized the cost parameter (C=3) on the sample data cross-validation result. http://www.csie.ntu.edu.tw/ cjlin/liblinear/.

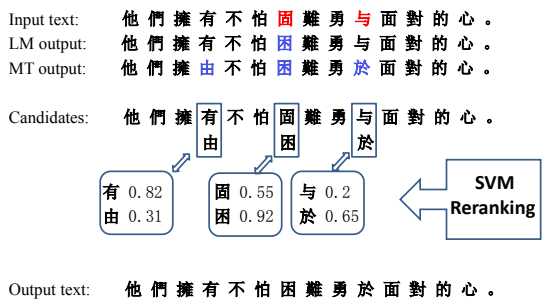

Figure 3: An example of SVM reranking.

includes similar shape confusion sets, similar pronunciation confusion sets, 350 sentences with error characters and 350 sentences without error characters. The second one includes the Chinese Gigaword Second Edition[6], the Chinese word segmentation bake-off 2005 corpus and a free traditional Chinese dictionary[7].

Since only 700 sample sentences are released, it is hard to estimate the error model using Formula 2. A better way is to extend the training corpus to estimate the translation probability. In our experiments, we replace each character in the provided sample sentence with the character in the confusion set to generate a new training instance. Guided by this procedure, around 2 million sentences are generated to train the "error model". However, it is too large for the SVM training. So we limited the candidate samples selecting 20-best sentences ranked by the language model.

4.2 Experiment Setting

For comparison, we combined the outputs of the translation model component and the language model component in three different ways:

1. **NAIST-Run1**: Union of the output candidates of the language model and the statistical machine translation model, and then reranked by SVM.

2. **NAIST-Run2**: Intersection of the output candidates of the language model and the statistical machine translation model, and then reranked by SVM.

[6]Released by LDC. Here we only used the traditional Chinese news to train the language model. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2005T14

[7]CC-CEDICT, which is a free dictionary, is released by Creative Commons Attribution-Share Alike 3.0 License. http://www.mdbg.net/chindict/chindict.php?page=cedict

Submission	LocAcc	CorAcc	CorPrec
NAIST-Run1	0.508	0.467	0.5765
NAIST-Run2	0.261	0.254	0.653
NAIST-Run3	0.487	0.453	0.6155

Table 2: **Final results on sub-task 2.** LocAcc, CorAcc and CorPrec denote location accuracy, correction accuracy and correction precision respectively.

3. **NAIST-Run3**: Only use the output of the language model and then reranked by SVM.

Here, we assume that union of the candidates might get a higher recall (NAIST-Run1), while the intersection of the candidates might get a higher precision (NAIST-Run2).

4.3 Experimental Results

In the final test, there are two data sets. Each task corpus contains 1000 sentences.

As shown in Table 1, NAIST-Run1 obtained the highest detection recall and NAIST-Run2 got the highest detection precision. However, NAIST-Run3 obtained the highest error location recall, the highest detection F-score and the error location F-score. We think the main reason is that the rate of sentences with error characters is much lower, around 5%, while NAIST-Run1 tends to find more correction candidates.

The final results of the error correction sub task are shown in Table 2. As we expect in Section 4.2, NAIST-Run2 obtained the correction precision, while NAIST-Run1 obtained both the highest location accuracy and the highest correction accuracy.

To evaluate the importance of the SVM reranking, we do another set of experiments on the 700 sample sentences with 5-fold cross-validation. We could obtain 34.7% of the error location precision and 69.1% of the error location recall using the language model based approach. After the reranking by the SVM, the error location precision increased to 70.2%, while the error location recall dropped to 67.0%. From this observation, the SVM reranking plays a crucial role for detection and correction of Chinese spelling errors.

5 Conclusion

We proposed a hybrid system which combines the language model and the statistical machine trans-

Submission	FAR	DAcc	DPr	DRe	DF-score	ELAcc	ELPr	ELRe	ELF-score
NAIST-Run1	0.2929	0.746	0.5504	0.8367	0.664	0.645	0.3289	0.5	0.3968
NAIST-Run2	0.0543	0.812	0.7979	0.5	0.6148	0.764	0.5426	0.34	0.418
NAIST-Run3	0.2243	0.777	0.5985	0.78	0.6773	0.698	0.3964	0.5167	0.4486

Table 1: **Final results on sub-task 1.** FAR denotes the false-alarm rate. DAcc, DPr, Dre and DF-score indicate detection accuracy, detection precision, detection recall and detection f-score respectively. ELAcc, ELPr, ELRe and ELF-score denote error location accuracy, error location precision, error location recall and error location f-score respectively.

lation model to generate almost all the correction candidates. To improve the precision of the Chinese spelling check, we employ SVM to rerank the correction candidates, where we could obtain a higher precision with a little recall dropping. We also proposed a simple approach to generate many artificial samples, which improved the recall of the statistical machine translation model. Our final test results reveal that our approach is competitive to other systems.

Acknowledgments

We would like to thank Keisuke Sakuchi, Komachi Mamoru and Lis Kanashiro for valuable discussions and comments.

References

Brown, Peter F and Pietra, Vincent J Della and Pietra, Stephen A Della and Mercer, Robert L. 2003. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics.*. 19-2, pp. 263–311.

Xue, Nianwen. 2003. Chinese word segmentation as character tagging *Computational Linguistics and Chinese Language Processing*. 8-1, pp. 29–48.

Sun, Weiwei. 2011. A Stacked Sub-Word Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. pp. 1385–1394, Portland, Oregon, USA.

Wu, Shih-Hung and Chen, Yong-Zhi and Yang, Pingche and Ku, Tsun and Liu, Chao-Lin. 2010. Reducing the false alarm rate of Chinese character error detection and correction. *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing*. pp. 54–61.

Burges, Christopher JC. 1998. A tutorial on support vector machines for pattern recognition *Data mining and knowledge discovery*. 2-2, pp. 121–167.

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

Zhang, Lei and Huang, Changning and Zhou, Ming and Pan, Haihua. 2000. Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. pp. 248–254.

Huang, Chuen-Min and Wu, Mei-Chen and Chang, Ching-Che. 2007. Error detection and correction based on Chinese phonemic alphabet in Chinese text. *Modeling Decisions for Artificial Intelligence*. pp. 463–476.

Hung, Ta-Hung and Wu, Shih-Hung. 2008. Chinese Essay Error Detection and Suggestion System. *Taiwan E-Learning Forum*.

Hung, Ta-Hung and Wu, Shih-Hung. AutoTag. Academia Sinaca.

Chen, Yong-Zhi and Wu, Shih-Hung and Yang, Ping-Che and Ku, Tsun. 2008. Improve the detection of improperly used Chinese characters in students' essays with error model. *International Journal of Continuing Engineering Education and Life Long Learning*. 21-1, pp. 103–116.

Chen, Yong-Zhi and Wu, Shih-Hung and Yang, Ping-Che and Ku, Tsun. 2009. Chinese confusion word set for automatic generation of spelling error detecting template. *The 21th Conference on Computational Linguistics and Speech Processing, Taichung, Taiwan, September*. pp. 1–2

Lin, Yih-Jeng and Huang, Feng-Long and Yu, Ming-Shing. 2002. A Chinese spelling error correction system. *Proceedings of the Seventh Conference on Artificial Intelligence and Applications (TAAI)*.

Bouma, Gerlof. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference*. pp. 31–40.

Rabiner, Lawrence. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*. 77-2, pp. 257–286.