Chinese Spelling Checker Based on Statistical Machine Translation

Hsun-wen Chiu **Jian-cheng Wu** **Jason S. Chang**
Department of Institute of Information Systems and Applications
National Tsing Hua University
{chiuhsunwen, wujc86, jason.jschang}@gmail.com

Abstract

Chinese spelling check is an important component for many NLP applications, including word processor and search engines. However, compared to checkers for alphabetical languages (e.g., English or French), Chinese spelling checkers are more difficult to develop, because there are no word boundaries in Chinese writing system, and errors may be caused by various Chinese input methods. In this paper, we proposed a novel method to Chinese spelling checking. Our approach involves error detection and correction based on the phrasal statistical machine translation framework. The results show that the proposed system achieves significantly better accuracy in error detecting and more satisfactory performance in error correcting.

1 Introduction

Chinese spelling check is a task involving automatically detecting and correcting typos, roughly corresponding to misspelled words in English. Liu et al. (2011) show that people tend to unintentionally generate typos that sound similar (e.g., "*措折 cuo zhe" and "挫折 cuo zhe"), or look similar (e.g., "*固難 gu nan" and "困難 kun nan"). On the other hand, some typos found on the Web (such as forums or blogs) are used deliberately for the purpose of speed typing or just for fun. Therefore, spelling check is an important component for many applications such as computer-aided writing and corpus cleanup.

The methods of spelling check can be broadly classified into two types: rule-based methods (Ren et al., 2001; Jiang et al., 2012) and statistical methods (Hung and Wu, 2009; Chen and Wu, 2010). Rule-based methods use knowledge resources such as a dictionary to identify a word as a typo if the word is not in the dictionary, and provide similar words in the dictionary as suggestions. However, simple rule-based methods have their limitations. Consider the sentence "心是很重要的。 xin shi hen zhong yao de" which is correct. However, the two single-character words "心 xin" and "是 shi" are likely to be regarded as an error by a rule-based model for the longer word "心事 xin shi" with identical pronunciation.

Data driven, statistical spelling check approaches appear to be more robust and performs better. Statistical methods tend to use a large monolingual corpus to create a language model to validate the correction hypotheses. Considering "心是 xin shi", the two characters "心 xin" and "是 shi" are a bigram which has high frequency in a monolingual corpus, so we may determine that "心是 xin shi" is not a typo after all.

In this paper, we propose a model, which combines rule-based with statistical approaches to detect errors and generate the most appropriate corrections in Chinese text. Once, an error is identified by the rule-based detection model, we use statistic machine translation (SMT) model (Koehn, 2010) to provide the most appropriate correction. Rule-based models tend to ignore context, so that we use SMT to deal with this problem. Our model treats spelling correction as a kind of translation, where typos are translated into correctly spelled words according to the translation probability and language model probability. Consider the same case "心是很重要的。 xin shi hen zhong yao de". The string "心是 xin shi" will not be incorrectly replaced with "心事 xin shi" because we would consider "心是 xin shi" is highly probable according to the language model.

The rest of the paper is organized as follows. We present the related work in the next section. Then we describe the proposed model for automatically detecting the spelling errors and correcting the found errors in Section 3. Section 4

Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), pages 49–53,
Nagoya, Japan, 14 October 2013.

and Section 5 present the experimental data and evaluation results. And we conclude in Section 6.

2 Related Work

Chinese spelling check is a task involving automatically detecting and correcting typos in a given Chinese sentence. Previous work typically takes the approach of combining a confusion set and a language model. Rule-based approach depends on dictionary knowledge and a confusion set, a collection set of a certain character consists of visually and phonologically similar characters. On the other hand, statistical-based methods usually use a language model, which is generated from a reference corpus. Statistical language model assigns a probability to a sentence of words by means of ngram probability to compute the likelihood of a corrected sentence.

Chang (1995) proposed a system that replaces each character in the sentence based on the confusion set and estimates the probability of all modified sentences according to a bigram language model built from a newspaper corpus, then comparing the probability before and after substitution. They used a confusion set consists of pairs of character with similar shape that are collected by comparing the original text and its OCR results. Similarly, Zhuang et al. (2004) proposed an effective approach using OCR to recognize possible confusion set. In addition, Zhuang et al. (2004) also used a multi-knowledge based statistical language model, the n-gram language model, and Latent Semantic Analysis. However, the experiments by Zhuang et al. (2004) seem to show that the simple n-gram model performs the best.

In recent years, Chinese spelling checkers have incorporated word segmentation. The method proposed by Huang et al. (2007) incorporates Sinica Word Segmentation System (Ma and Chen, 2003) to detect typos. With a character-based bigram language model and the rule-based methods of dictionary knowledge and confusion set, the method determines whether the word is a typo or not. There are many more systems that use word segmentation to detect errors. For example, in Hung and Wu (2009), the given sentence is segmented using a bigram language model. In addition, the method also uses confusion set and common error templates manually edited and provided by Ministry of Education in Taiwan. Chen and Wu (2010) modified the system proposed by Hung and Wu (2009), by combining statistic-based methods and a template matching module generated automatically to detect and correct typos based on language model.

In a work closer to our method, Wu et al. (2010) adopts the noise channel model, a framework used both in spell checkers and machine translation systems. The system combined statistic-based method and template matching with the help of a dictionary and a confusion set. They also used word segmentation to detect errors, but they did not use an existing word segmentation as Huang et al. (2007) did, because it might regard a typo as a new word. They used a backward longest first approach to segment sentences with an online dictionary sponsored by MOE, and a templates with a confusion set. The system also treat Chinese spelling check as a kind of translation, they combine the template module and translation module to get a higher precision or recall.

In our system, we also treat Chinese spelling checking problem as machine translation like Wu et al. (2010), with a different way of handling word segmentation to detect typos and translation model where typos are translated into correctly spelled words.

3 Method

In this section, we describe our solution to the problem of Chinese spelling check. In the error detection phase, the given Chinese sentence is segmented into words. (Section 3.1) The detection module then identifies and marks the words, which may be typos. (Section 3.2) In the error correction phase, we use a statistical machine translation (SMT) model to translate the sentences containing typos into correct ones (Section 3.3). In the rest of this section, we describe our solution to this problem in more details.

3.1 Modified Chinese Word Segmentation System

Unlike English text in which sentences are sequences of words delimited by spaces, Chinese texts are represented as strings of Chinese characters (called Hanzi) with word delimiters. Therefore, word segmentation is a pre-processing step required for many Chinese NLP applications. In this study, we also perform word segment to reduce the search space and the probability of false alarm. After segmentation, sequences of two or more singleton words are considered likely to contain an error. However, over-segmented might lead to falsely identified errors, which we will describe in Section 3.2. Considering the sen-

Replaced character	氣		份	
Translations	汽份	泣份	氣分	氣忿
	器份	契份	氣憤	氣糞
	企份	憩份	氣奮	氣氛

Table 1. Sample "translations" for "氣份 qi fen".

tence "除了要有超世之才，也要有堅定的意志 chu le yao you chao shi zhi cai, ye yao you jian ding de yi zhi", the sentence is segmented into "除了/要/有/超世/之/才/，/也/要/有/堅定/的/意志." The part "超世之才 chao shi zhi cai" of the sentence is over-segmented and runs the risk of being identified as containing a typo. To solve the problem of over-segmentation, we used additional lexicon items and reduce the chance of generating false alarms.

3.2 Error Detection

Motivated by the observation that a typo often causes over-segmentation in the form of a sequence of single-character words, so we target the sequences of single-character words as candidates for typos. To identify the points of typos, we take all n-grams consist of single-character words in the segmented sentence into consideration. In addition to a Chinese dictionary, we also include a list of web-based ngrams to reduce the false alarm due to the limited coverage of the dictionary.

When a sequence of singleton word is not found in the dictionary, or in the web-based character ngrams, we regard the ngram as containing a typo. For example, "森林 的 芳 多 精 sen lin de fang duo jing" is segmented into consecutive singleton words: bigrams such as "的 芳 de fang", and "芳 多 fang duo" and trigrams such as "的 芳 多 de fang duo" and "芳 多 精 fang duo jing" are all considered as candidates for typos since those ngrams are not found in the reference list.

3.3 Error Correction

Once we generate a list of candidates of typos, we attempt to correct typos, using a statistical machine translation model to translate typos into correct word. When given a candidate, we first generate all correction hypotheses by replacing each character of the candidate typo with similar characters, one character at a time.

Take the candidate "氣份 qi fen" as example, the model generates all translation hypotheses, according to a visually and phonologically confusion set. Table 1 shows some translation hypotheses. The translation hypotheses are then validated (or pruned from the viewpoint of SMT) using the dictionary.

Translations	Freq.	LM prob.	tp
氣憤	48	-4.96	-1.20
氣氛	473	-3.22	-1.11

Table 2. Translations for "氣份 qi fen".

The translation probability *tp* is a probability indicates how likely a typo is translated into a correct word. *tp* of each correction translation is calculated using the following formula:

$$tp = \log_{10}\left(\frac{freq(trans)}{freq(trans) - freq(candi)}\right) * \gamma$$

where *freq(trans)* and *freq(candi)* are the frequency of the translation and the candidate correspondingly, and γ is the weight of different error types: visual or phonological. *tp* is set to 0 if *freq(trans)* = 0.

Take "氣份 qi fen" from "不/一樣/的/氣/份 bu/yi yang/de/qi/fen" for instance, the translations with non-zero *tp* after filtering are shown in Table 2. Only two translations are possible for this candidate: "氣憤 qi fen" and "氣氛 qi fen".

We use a simple, publicly available decoder written in Python to correct potential spelling errors found in the detection module. The decoder reads in a Chinese sentence at a time and attempts to "translate" the sentence into a correctly spelled one. The decoder translates monotonically without reordering the Chinese words and phrases using two models — translation probability model and the language model. These two models read from a data directory containing two text files containing a translation model in GIZA++ (Och and Ney, 2003) format, and a language model in SRILM (Stolcke et al., 2011) format. These two models are stored in memory for quick access.

The decoder invokes the two modules to load the translation and language models and decodes the input sentences, storing the result in output. The decoder computes the probability of the output sentences according to the models. It works by summing over all possible ways that the model could have generated the corrected sentence from the input sentence. Although in general covering all possible corrections in the translation and language models is intractable, but a majority of error instances can be "translated"

effectively by using the translation model and the language model.

4 Experimental Setting

To train our model, we used several corpora including Sinica Chinese Balanced Corpus, TWWaC (Taiwan Web as Corpus), a Chinese dictionary, and a confusion set. We describe the data sets in more detail below.

Sinica Corpus
"Academia Sinica Balanced Corpus of Modern Chinese", or Sinica Corpus for short, is the first balanced Chinese corpus with part-of-speech tags (Huang et al., 1996). Current size of the corpus is about 5 million words. Texts are segmented according to the word segmentation standard proposed by the ROC Computational Linguistic Society. We use the corpus to generate the frequency of bigram, trigram and 4-gram for training translation model and to train the n-gram language model.

TWWaC (Taiwan Web as Corpus)
We use TWWaC for obtaining more language information. TWWaC is a corpus gathered from the Web under the .tw domain, containing 1,817,260 Web pages, 30 billions Chinese characters. We use the corpus to generate the frequency of all character n-grams for n = 2, 3, 4 (with frequency higher than 10).

Words and Idioms in a Chinese Dictionary
From the dictionaries and related books published by Ministry of Education (MOE) of Taiwan, we obtained two lists, one is the list of 64,326 distinct Chinese words[1], and the other one is the list of 48,030 distinct Chinese idioms[2]. We combine the lists into a Chinese dictionary for validating words with lengths of 2 to 17 characters.

Confusion Set
After analyzing erroneous Chinese word, Liu et al. (2011) found that more than 70% of typos were related to the phonologically similar character, about 50% are morphologically similar and almost 30% are both phonologically and morphologically similar. We use the ratio as the weight for the translation probabilities. In this study, we used two confusion sets generated by Liu et al. (2011) and provided by SIGHAN 7 Bake-off 2013: Chinese Spelling Check Shared Task as a full confusion set based on loosely similar relation.

In order to improve the performance, we expanded the sets slightly and also removed some loosely similarly relations. For example, we removed all relations based on non-identical phonologically similarity. After that, we added the similar characters based on similar phonemes in Chinese phonetics, such as "ㄣ，ㄥ en, eng", "ㄤ，ㄢ ang, an", "ㄕ，ㄙ shi, si" and so on. We also modify the similar shape set to a more strongly similar set. The characters are checked automatically by comparing corresponding Cangjie code (倉頡碼). Two characters which differ from each other by at most one symbol in Cangjie code are considered as strongly similar and are retained. For example, the code of "徵 zheng" and "微 wei" are strongly similar in shape, since in their corresponding codes "竹人山土大" and "竹人山山大", differ only in one place.

5 Evaluation Results

In Bake-off 2013, the evaluation includes two sub-tasks: error detection and error correction. For the error detection, sub-task1, there are 1000 sentences with/without spelling errors. And sub-task2 for the error correction, there are also containing 1000 sentences but all with errors. The evaluation metrics, which computes false-alarm rate, accuracy, precision, recall, and F-Score is provided by SIGHAN 7 Bake-off 2013: Chinese Spelling Check Shared Task. In this paper, we describe Run3 system and results.

On sub-task1, evaluation results as follows:

Evaluation metrics	Score
False-Alarm Rate	0.0514
Detection Accuracy	0.861
Detection Precision	0.8455
Detection Recall	0.6567
Detection F-Score	0.7392
Error Location Accuracy	0.82
Error Location Precision	0.6695
Error Location Recall	0.52
Error Location F-Score	0.5854

Table 3. Evaluation metrics of Sub-task1.

We obtain higher detection accuracy, error location accuracy, and error location F-Score, which put our system in first place among 13 systems evaluated. On sub-task2, our system obtained

[1] Chinese Dictionary http://www.edu.tw/files/site_content/m0001/pin/yu7.htm?open

[2] Idioms http://dict.idioms.moe.edu.tw/cydic/index.htm

location accuracy, correction accuracy, and correction precision of 0.454, 0.443, and 0.6998, respectively.

6 Conclusions and Future Work

Many avenues exist for future research and improvement of our system. For example, new terms can be automatically discovered and added to the Chinese dictionary to improve both detection and correction performance. Part of speech tagging can be performed to provide more information for error detection. Named entities can be recognized in order to avoid false alarms. Supervised statistical classifier can be used to model translation probability more accurately. Additionally, an interesting direction to explore is using Web ngrams in addition to a Chinese dictionary for correcting typos. Yet another direction of research would be to consider errors related to a missing or redundant character.

In summary, we have introduced in this paper, we proposed a novel method for Chinese spelling check. Our approach involves error detection and correction based on the phrasal statistical machine translation framework. The error detection module detects errors by segmenting words and checking word and phrase frequency based on a compiled dictionary and Web corpora. The phonological or morphological spelling errors found are then corrected by running a decoder based on statistical machine translation model (SMT). The results show that the proposed system achieves significantly better accuracy in error detecting and the evaluation results show that the method outperforms other system in Chinese Spelling Check Shared Task.

References

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

Yong-Zhi Chen and Shih-Hung Wu. 2010. *Improve the detection of improperly used Chinese characters with noisy channel model and detection template*. Master thesis, Chaoyang University of Technology.

Chu-Ren Huang, Keh-jiann Chen and Li-Li Chang. 1996. Segmentation standard for Chinese natural language processing. *Proceedings of the 1996 International Conference on Computational Linguistics (COLING 96)*, Vol. 2, pp. 1045-1048.

Chuen-Min Huang, Mei-Chen Wu and Chang Ching-Che. 2007. Error detection and correction based on Chinese phonemic alphabet in Chinese text. *Proceedings of the 4th International Conference on Modeling Decisions for Artificial Intelligence (MDAI IV)*, pp. 463-476.

Ta-Hung Hung and Shih-Hung Wu. 2009. *Automatic Chinese character error detecting system based on n-gram language model and pragmatics knowledge base*. Master thesis, Chaoyang University of Technology.

Ying Jiang, Tong Wang, Tao Lin, Fangjie Wang, Wenting Cheng, Xiaofei Liu, Chenghui Wang and Weijian Zhang. 2012. A rule based Chinese spelling and grammar detection system utility. *2012 International Conference on System Science and Engineering (ICSSE)*, pp. 437-440.

Philipp Koehn. 2010. Statistical Machine Translation. United Kingdom: Cambridge University Press.

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, pp. 39.

Wei-Yun Ma and Keh-Jiann Chen. 2003. Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff. *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, Vol. 17, pp. 168-171.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, Vol. 29, number 1, pp. 19-51.

Fuji Ren, Hongchi Shi and Qiang Zhou. 2001. A hybrid approach to automatic Chinese text checking and error correction. *2001 IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 3, pp. 1693-1698.

Andreas Stolcke, Jing Zheng, Wen Wang and Victor Abrash. 2011. SRILM at Sixteen: Update and Outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*.

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

Li Zhuang, Ta Bao, Xiaoyan Zhu, Chunheng Wang and Satoshi Naoi. 2004. A Chinese OCR spelling check approach based on statistical language models. *2004 IEEE International Conference on Systems, Man and Cybernetics*, Vol. 5, pp. 4727-4732.