View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by CiteSeerX

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 4, December 2013, pp. 17-30
© The Association for Computational Linguistics and Chinese Language Processing

Integrating Dictionary and Web N-grams for Chinese Spell Checking

Jian-cheng Wu*, Hsun-wen Chiu+, and Jason S. Chang*,+

Abstract

Chinese spell checking is an important component of many NLP applications, including word processors, search engines, and automatic essay rating. Nevertheless, compared to spell checkers for alphabetical languages (*e.g.*, English or French), Chinese spell checkers are more difficult to develop because there are no word boundaries in the Chinese writing system and errors may be caused by various Chinese input methods. In this paper, we propose a novel method for detecting and correcting Chinese typographical errors. Our approach involves word segmentation, detection rules, and phrase-based machine translation. The error detection module detects errors by segmenting words and checking word and phrase frequency based on compiled and Web corpora. The phonological or morphological typographical errors found then are corrected by running a decoder based on the statistical machine translation model (SMT). The results show that the proposed system achieves significantly better accuracy in error detection and more satisfactory performance in error correction than the state-of-the-art systems.

Keywords: Chinese Spelling Detection, Chinese Spelling Correction, Chinese Similar Characters, Ngram, Language Model, Machine Translation.

1. Introduction

Chinese spell checking is a task involving automatically detecting and correcting typographical errors (typos), roughly corresponding to misspelled words in English. In this paper, we define typos as Chinese characters that are misused due to shape or phonological similarity. Liu *et al*. (2011) shows 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

* Department of Computer Science, National Tsing Hua University
E-mail: { wujc86; jason.jschang}@gmail.com
+ Department of Institute of Information Systems and Applications, National Tsing Hua University
E-mail: chiuhsunwen@gmail.com

or blogs) are used deliberately for the purpose of speed typing or just for fun. Therefore, spell checking is an important component for many applications, such as computer-aided writing and corpus cleanup.

The methods of spell checking can be classified broadly into two types: rule-based methods (Ren *et al*., 2001; Jiang *et al*., 2012) and statistical methods (Hung & Wu, 2009; Chen & 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 to provide similar words in the dictionary as suggestions. Simple rule-based methods, however, have their limitations. Consider the sentence "心是很重要的。" [xin shi hen zhong yao de] which is correct. Nevertheless, 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 spell checking approaches appear to be more robust and perform 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 with 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 that combines rule-based and 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 the statistic machine translation (SMT) model (Koehn, 2010) to provide the most appropriate correction. Rule-based models tend to ignore context, so 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] would not be incorrectly replaced with "心事" [xin shi] because we would consider "心是" [xin shi] to be 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 and Section 5 present the experimental data, results, and performance analysis. We conclude in Section 6.

2. Related work

Chinese spell checking 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. A rule-based approach depends on dictionary knowledge and a

confusion set, a collection set of certain characters consisting 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. A statistical language model assigns a probability to a sentence of words by means of n-gram 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 before comparing the probability before and after substitution. They used a confusion set consisting of pairs of characters with similar shape that were collected by comparing the original text and its OCR results. Similarly, Zhuang *et al*. (2004) proposed an effective approach using OCR to recognize a 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. Nevertheless, the experiments by Zhuang *et al*. (2004) seem to show that the simple n-gram model performs the best.

In recent years, Chinese spell checkers have incorporated word segmentation. The method proposed by Huang *et al*. (2007) incorporates the Sinica Word Segmentation System (Ma & Chen, 2003) to detect typos. With a character-based bigram language model and the rule-based methods of dictionary knowledge and confusion sets, 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 a confusion set and common error templates manually edited and provided by the Ministry of Education in Taiwan (MOE, 1996). 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 the language model.

Closer to our method, Wu *et al*. (2010) adopted the noise channel model, a framework used both in spell checkers and in machine translation systems. The system combined a 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 existing word segmentation, as Huang *et al*. (2007) did, because that 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 (MOE, 2007), and a templates with a confusion set provided by Liu *et al*. (2009). The system also treated Chinese spell checking as a kind of translation by combining the template module and translation module to get a higher precision or recall.

In our system, we also treat the Chinese spell checking problem as machine translation, but we use a different method 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 spell checking. 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 that may be typos. (Section 3.2) In the error correction phase, we use the 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 detail.

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 segmentation to reduce the search space and probability of false alarms. After segmentation, sequences of two or more singleton words are considered likely to contain an error. Nevertheless, over-segmentation might lead to falsely identified errors, which we will describe in Section 3.2. Considering the sentence "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 lexical items to 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, we target the sequences of single-character words as candidates for typos. To identify the points of typos, we take all n-grams consisting of single-character words in the segmented sentence into consideration. In addition to a Chinese dictionary, we also include a list of web-based n-grams to reduce false alarms due to the limited coverage of the dictionary.

When a sequence of singleton words is not found in the dictionary or in the web-based character n-grams, we regard the n-gram 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 n-grams 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 words. 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 then are validated (or pruned from the viewpoint of SMT) using the dictionary.

Table 1. Sample "translations" for the candidate "氣份" [qi fen].

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	氣吩
	訖份	氮份	氣扮	氣汾
	迄份	粥份	氣芬	氣氛

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

$$tp(candi,\ trans) = \log_{10}\left(\frac{freq(trans)}{freq(trans) - freq(candi)} * \gamma\right)^{\substack{if\ trans\ in\ ngrams \\ otherwise = 0}} \quad (1)$$

where *freq(trans)* is the frequency of translation, *freq(candi)* is the frequency of the candidate, and γ is the weight of different error types: visual or phonological.

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].

Table 2. Translations for "氣份" [qi fen] with corresponding translation probability and language model probability (log).

Translations	Frequency	LM probability	tp
氣憤	48	-4.96	-1.20
氣氛	473	-3.22	-1.11

We use a simple, publicly available decoder written in Python to correct potential spelling errors found in the detection module. The decoder reads one 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 — the 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 & 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, a majority of error instances can be "translated" effectively via the translation model and the language model.

4. Experimental Setting

Our systems were designed to provide wide coverage spell checking for Chinese. As such, we trained our systems using a dictionary, a compiled corpus, and Web scale n-grams. We evaluated our systems on the sentence level. Finally, we used an annotated dataset to provide human judges the ability to evaluate the quality of error detection and correction. In this section, we first present the details of data sources used in training (Section 4.1). Then, Section 4.2 describes the test data. Section 4.3 describes the systems evaluated and compared. The evaluation metrics for the performance of the systems are reported in Section 4.4.

4.1 Data Sources

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," is the first balanced Chinese corpus with part-of-speech tags (Huang *et al*., 1996). The 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. Each segmented word is tagged with its part of speech. We used the corpus to generate the frequency of bigrams, trigrams, and 4-grams for training the translation model and to train the n-gram language model.

TWWaC (Taiwan Web as Corpus)

We used TWWaC for obtaining more language information. TWWaC is a corpus gathered from the Web under the .tw domain, containing 1,817,260 Web pages that consist of 30 billion 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). Table 3 shows the information of n-grams in Sinica Corpus and TWWaC.

Table 3. The information of n-grams in Sinica corpus and TWWaC.

N-gram	Sinica Corpus Types	TWWaC Types
2-gram	66,778	2,848,193
3-gram	45,382	13,745,743
4-gram	12,294	17,191,359

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 (MOE, 1997)[1], and the other one is the list of 48,030 distinct Chinese idioms[2]. We combined the lists into a Chinese dictionary for validating words with lengths of 2 to 17 characters.

Confusion Set

After analyzing erroneous Chinese words, 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 used 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 modified the similar shape set, so we checked the characters by comparing the characters in Cangjie codes (倉頡碼) and required strong shape similarly. Two characters differing from each other by at most one symbol in Cangjie code were considered as strongly similar and were retained. For example, the code of “徵” [zheng] and “微” [wei]

[1] Chinese Dictionary http://www.edu.tw/files/site_content/m0001/pin/biau2.htm?open

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

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.

4.2 Test Data

We used the official dataset from SIGHAN 7 Bake-off 2013: Chinese Spelling Check to evaluate our systems. This dataset contains two parts: 350 sentences with errors and 350 sentences without errors, extracted from student essays that covered various common errors. The dataset was released in XML format with the information of sentences, wrong position, typos, and correction. A sample is shown below:

<DOC Nid="00001">
<P>我看過許多勇敢的人，不怕措折地奮鬥，這種精神值得我們學習。</P>
<TEXT>
<MISTAKE wrong_position=13>
<WRONG>措折</WRONG>
<CORRECT>挫折</CORRECT>
</MISTAKE>
</TEXT>
</DOC>

We found that all of the sentences with errors contain exactly one typo and that most errors were either similar in pronunciation or shape. Therefore, the confusion set was suitable for error correction. We generated the sentence with/without error and the correct answer from XML format. In this data, more than 80% of errors were characters with identical pronunciation, almost 20% of errors were characters with similar shape, and 40% of errors involved both phonological and visual similarity. Hence, we focused on detecting and correcting these two common types of errors in our study.

4.3 Systems Compared

Recall that we propose a system to detect and correct typos in Chinese based broadly on statistical machine translation. We experimented with different resources as kinds of language models to detect typos: dictionary entries, a compiled corpus, and Web corpus. The four detection systems evaluated are:

—Dictionary (**DICT**): A dictionary is used to detect unregistered words as errors.

—Corpus (**CORP**): A word list from a reference corpus is used to detect unseen words as errors.

—Web corpus (**WEB**): A character n-gram of Web corpus is used to detect unseen n-grams as errors.

—Dictionary and Web corpus (**DICT+WEB**): A dictionary combining a character n-gram of Web corpus is used to detect unregistered words as errors.

To correct typos, we used a character confusion set to transform the detected typos and generate the "*translation*" hypotheses with translation probability. These hypotheses were pruned using a Chinese dictionary before running the MT decoder in order to reduce the load on the decoder. The scope of this confusion set and the weights associated with translation probability clearly influenced the performance of our system. We evaluated and compared four different confusion set and weight settings. The four correction systems evaluated are:

—Full confusion set (**FULL+WT**): A broad confusion set with loosely similar relations in character sound and shape was used to generate mapping from a detected typo to its correction. Different weights were used in modeling probability for sound and shape based mapping.

—Confusion set with identical sound (**SND+WT**): A broad confusion set with identical sounds and loosely similar shape relations was used to generate mapping. Different weights were used in modeling probability for sound and shape based mapping.

—Restricted confusion set with identical sound and strong similarly shape (**SND+SHP**): A broad confusion set with identical sounds and strongly similar shape relations was used to generate mapping. Sound and shape were given the same weight.

—Restricted confusion set with different weights (**SND+SHP+WT**): A broad confusion set with identical sounds and strongly similar shape relations was used to generate mapping. Different weights were used in modeling probability for sound and shape based mapping.

4.4 Evaluation Metrics

To assess the effectiveness of the proposed system, we used test data to experiment with our system. We also exploited several language resources, including TWWaC, Sinica Corpus, a Chinese dictionary, and the confusion set, in the proposed system to detect errors and correct errors. The Chinese Word Segmentation System produces the word segmentation result with the help of a Chinese dictionary to improve the proposed system. To evaluate our system, we used the precision rate and recall rate, which are defined as follows:

$$Precision = C / S \tag{2}$$

$$Recall = C / N \tag{3}$$

where N is the number of error characters, S is the number of characters translated by the proposed system, and C is the number of characters translated correctly by the proposed system. We also compute the corresponding F-score as:

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

5. Evaluation Results

In this section, we report the results of the experimental evaluation using the methodology described in the previous section. We evaluated detection, as well as correction, for many systems with different language resources and settings. During this evaluation, we tested our systems on 350 sentences containing at least one typo, provided in SIGHAN Bake-off 2013: Chinese Spelling Check. Table 4 shows the precision, recall, and F-score for four detection systems, while Table 5 shows the same metrics for four correction systems.

Table 4. The comparison of the detection with different references.

System	Precision	Recall	F-score
DICT	.91	.52	.66
CORPUS	.90	.46	.61
WEB	.93	.47	.63
WEB+DICT	**.95**	**.56**	**.71**

Table 5. The comparison of the correction experiment.

System	Precision	Recall	F-score
FULL+WT	.53	.51	.52
SND+WT	.74	.57	.65
SND+SHP	.90	.55	.68
SND+SHP+WT	**.95**	**.56**	**.70**

As can be seen in Table 4, using the Web corpus (**WEB**) achieves higher precision than the dictionary (**DICT**) or compiled corpus (**CORPUS**) with slightly lower recall. Using the dictionary (**DICT**) leads to the highest recall but slightly lower precision. By combining the dictionary and Web corpus (**WEB+DICT**), we achieve the best precision, recall, and F-score.

Table 5 shows that using the full confusion set with loosely similar sound and shape relation leads to the lowest recall and precision in error correction (**FULL**). By restricting the sound confusion to identical sound and the shape confusion to strongly similar shape, we can improve precision dramatically, with a small increase in recall (**SND** and **SND+SHP**).

We can further improve the precision and recall by applying different weights in modeling the probability of sound and shape based hypotheses (**SND+SHP+WT**). Since typos are more often related to sound confusion than shape, giving higher weight to sound confusion

indeed leads to further improvement in both precision and recall. Previous works typically have used only a language model to correct errors, but we compute language model probability and translation probability, resulting in more effective error correction. For this reason, we were placed among the top scoring systems in the SIGHAN Bake-off 2013.

In order to test whether the system can produce false alarms as rarely as possible, when handling the sentences with typos, we tested our systems on a dataset with an additional 350 sentences without typos. The best performing system (**SND+SHP+WT**) obtained a precision rate of .91, recall rate of .56, and F-score of .69 in correction. The results show that this system is very robust, maintaining a high precision rate in different situations.

The recall of our system is limited by the dictionary that we used to correct a typo. For example, the typo "七彈場" [qi tan chang], which is detected by the model, is not corrected to "漆彈場" [qi tan chang] because it is a new term and not found in the Chinese dictionary we used. To correct such errors, we could use Web-based character n-grams, which are more likely to contain such new terms or productive compounds not found in a dictionary.

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. A supervised statistical classifier can be used to model translation probability more accurately. Additionally, an interesting direction to explore is using Web n-grams 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 proposed a novel method for Chinese spell checking. 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 then are corrected by running a decoder based on the statistical machine translation model (SMT). The results show that the proposed system achieves significantly better accuracy in error detection and more satisfactory performance in error correction than the state-of-the-art systems. The experimental results show that the method outperforms previous works.

References

Chang, C.-H. (1995). A new approach for automatic Chinese spelling correction. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, 278 - 283.

Chen, Y.-Z. (2010). *Improve the detection of improperly used Chinese characters with noisy channel model and detection template*. Master thesis, Chaoyang University of Technology.

Huang, C.-R., Chen, K.-j., & Chang, L.-L. (1996). Segmentation standard for Chinese natural language processing. In *Proceedings of the 1996 International Conference on Computational Linguistics (COLING 96)*, *2*, 1045 - 1048.

Huang, C.-M., Wu, M.-C., & Chang C.-C. (2007). Error detection and correction based on Chinese phonemic alphabet in Chinese text. In *Proceedings of the 4th International Conference on Modeling Decisions for Artificial Intelligence (MDAI IV)*, 463 - 476.

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

Jiang, Y., *et al*. (2012). A rule based Chinese spelling and grammar detection system utility. *2012 International Conference on System Science and Engineering (ICSSE)*, 437 - 440, 30 June - 2 July 2012.

Koehn, P. (2010). *Statistical Machine Translation*. United Kingdom: Cambridge University Press.

Liu, C.-L., Tien, K.-W., Lai, M.-H., Chuang, Y.-H., & Wu, S.-H. (2009). Phonological and logographic influences on errors in written Chinese words. In *Proceedings of the Seventh Workshop on Asian Language Resources (ALR7), the Forty Seventh Annual Meeting of the Association for Computational Linguistics (ACL'09)*, 84 - 91.

Liu, C.-L., Lai, M.-H., Tien, K.-W., Chuang, Y.-H., Wu, S.-H., &Lee, C.-Y. (2011). Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications. *ACM Trans. Asian Lang, Inform. Process*, *10*(2), Article 10, pages 39, .

Ma, W.-Y., & Chen, K.-J. (2003). Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff. In *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, *17*, 168 - 171.

MOE. (1997). *MOE word frequency table*, Taiwan: Ministry of Education.

MOE. (2007). *MOE Dictionary new edition*. Taiwan: Ministry of Education.

MOE. (1996). *Common errors in Chinese writings*. Taiwan: Ministry of Education.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19 - 51.

Ren, F., Shi, H., & Zhou, Q. (2001). A hybrid approach to automatic Chinese text checking and error correction. *2001 IEEE International Conference on Systems, Man, and Cybernetics*, *3*, 1693 - 1698, 07 - 10 Oct. 2001.

Stolcke, A., Zheng, J., Wang, W., & Abrash, V. (2011). SRILM at Sixteen: Update and Outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 2011.

Wu, S.-H., Chen, Y.-X., Yang, P.-c., Ku, T., & Liu, C.-L. (2010). Reducing the false alarm rate of Chinese character error detection and correction. In *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2010)*, 54 - 61, 28 - 29 Aug. 2010.

Zhuang, L., Bao, T., Zhu, X., Wang, C., & Naoi, S. (2004). A Chinese OCR spelling check approach based on statistical language models. *2004 IEEE International Conference on Systems, Man and Cybernetics*, 5, 4727 - 4732, 10 - 13 Oct. 2004.