機器翻譯為本的中文拼字改錯系統

Chinese Spelling Checker Based on Statistical Machine Translation

邱絢紋　Hsun-wen Chiu
吳鑑城　Jian-cheng Wu
張俊盛　Jason S. Chang
國立清華大學資訊系統與應用研究所
Department of Institute of Information Systems and Applications
National Tsing Hua University
chiuhsunwen@gmail.com, wujc86@gmail.com, jason.jschang@gmail.com

Abstract

Chinese spell check is an important component for many NLP applications, including word processors, search engines, and automatic essay rating. However, 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 Chinese writing system, and errors may be caused by various Chinese input methods. Chinese spell check involves 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 alike (e.g., *固難 [gu nan] and 困難 [kun nan]).

The methods for spell check can be broadly classified into two types: rule-based methods (Ren et al., 2001; Jiang et al., 2012) and statistical methods (Hung & Wu, 2009; Chen, 2010). Rule-based methods use knowledge resources such as a dictionary to identify a word as a typo. Statistical methods tend to use a large monolingual corpus to create a language model to validate the correction hypotheses. Consider the sentence “心是很重要的。” [xin shi hen zhong yao de] which is correct. However, “心” and “是” are likely to be regarded as an error by a rule-based model for the word “心事” with identical pronunciation. In statistical methods, “心” and “是” are a bigram which has high frequency in a monolingual corpus, so we may determine that “心是” is not a typo after all. In this paper, we propose a model that combines rule-based and statistical approaches. Probable errors, proposed by the rule-based detection module, are verified using statistical machine translation (SMT) model. Our model treats spell check and correction as a kind of translation, where typos are translated into correctly spelled words according to the translation probability and the language model probability.

We describe three modules for solving the problem of Chinese spell check: word segmentation, error detection, and error correction. The first module segments the input sentence into word tokens in an attempt to reduce the search space and the probability of false alarm. The second module detects probable errors in the segmented tokens. Any sequences of two or more singleton words are considered likely to contain an error. However, over-segmentation might lead to falsely identified errors. For example, phrases like “超世之才” [chao shi zhi cai] tend to be over-segmented to “超世/之/才” which might lead to false alarms. So we use additional lexicon items and reduce the chance of generating false alarms.

In addition, we use n-grams consist of single-character words to distinguish between correct token sequences and typos: when a sequence of singleton words is not found in the reference list of dictionary entries plus 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]: bigrams such as "的 芳", and "芳 多" and trigrams such as "的 芳 多" and "芳 多 精" are all considered as candidates for typos since those ngrams are not found in the reference list.

The third and final module is the error corrector. we use a SMT model to translate the sentences containing typos into correct ones. 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. The translation probability *tp* is a probability indicating how likely a typo is translated into a correct word. *tp* of each correction translation is calculated using the following formula:

$$tp = \log_{10}\left(\frac{freq(trans)}{freq(trans) - freq(candi)} * \gamma\right) \quad \begin{matrix} if\ trans\ in\ ngrams \\ otherwise = 0 \end{matrix}$$

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.

We use a simple, publicly available decoder written in Python which translates monotonically without reordering the Chinese words and phrases using translation probability and language models. To train our model, we used several corpora including Sinica Chinese Balanced Corpus, TWWaC (Taiwan Web as Corpus), a Chinese dictionary (MOE, 1997), and a confusion set (Liu et al., 2011). The decoder reads a translation model in GIZA++ format, and a language model in SRILM format. We used the official dataset from SIGHAN 7 Bake-off 2013: Chinese Spell Check to evaluate our systems.

The results produced by the proposed system were evaluated using precision rate, recall rate and F-score. We evaluated the results of detection as well as correction for many systems with different language resources and settings. The results show that using Web corpus achieves higher precision than dictionary or compiled corpus in detection systems. Using dictionary leads to the highest recall but slightly lower precision. By combining dictionary and Web corpus, we achieve the best precision, recall, and F-score. By restricting the sound confusion to identical sounds and the shape confusion to strongly similar shapes, we can improve precision dramatically. We can further improve the precision and recall, by using different weights in modeling probability of sound and shape based hypotheses which obtain precision rate of .95, recall rate of .56, and F-score of .70 in correction. Because typos are more often related to sound confusion than shape confusion, so giving higher weight to sound confusion indeed leads to further improvement in both precision and recall. In order to test whether we can reduce false alarms further, we tested our systems on a dataset with additional 350 sentences without typos. The best performing system obtain 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 high precision rate in different situations.

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 missing or unnecessary characters.

In summary, we have introduced in this paper, we proposed a novel method for Chinese spell 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 more satisfactory performance in error correcting than the state-of-the-art systems. The experiment results show that the method outperforms previous work.

References

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

Ta-Hung Hung ＆ 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, ＆ Weijian Zhang (2012). A rule based Chinese spelling and grammar detection system utility. *2012 International Conference on System Science and Engineering (ICSSE)*, pages 437 - 440, 30 June - 2 July 2012.

Chao-Lin Liu, Min-Hua Lai, Kan-Wen Tien, Yi-Hsuan Chuang, Shih-Hung Wu, **&** 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, pages 39, June 2011.

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

Fuji Ren, Hongchi Shi, ＆ 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, pages 1693 - 1698, 07 - 10 Oct. 2001.