Extended HMM and Ranking models for Chinese Spelling Correction

Jinhua Xiong, Qiao Zhao, Jianpeng Hou,
Qianbo Wang, Yuanzhuo Wang and Xueqi Cheng
CAS Key Laboratory of Network Data Science and Technology
Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
xjh@ic.ac.cn, zhangqiao@software.ict.ac.cn

Abstract

Spelling correction has been studied for many decades, which can be classified into two categories: (1) regular text spelling correction, (2) query spelling correction. Although the two tasks share many common techniques, they have different concerns. This paper presents our work on the CLP-2014 bake-off. The task focuses on spelling checking on foreigner Chinese essays. Compared to online search query spelling checking task, more complicated techniques can be applied for better performance. Therefore, we proposed a unified framework for Chinese essays spelling correction based on extended HMM and ranker-based models, together with a rule-based model for further polishing. Our system showed better performance on the test dataset.

1 Introduction

The number of people learning Chinese as a Foreign Language (CFL) is booming in recent decades, and the number is expected to become even larger for the years to come[1]. Therefore spelling correction tool to support such learners to correct and polish their essays becomes very valuable. Spelling correction has been studied for many years on regular text and web search query. Although the two tasks share many common techniques, they have different concerns. Compared to web search query spelling correction which normally need to give corrections instantly, complicated techniques can be applied to spelling correction on essays, in order to improve the performance. Spelling correction on Chinese essays of CFL learners faces the following challenges:

(1) There is no word boundary between Chinese word, which may result in the error on splitting, and the error may accumulate.

(2) The number of error type is more than other case, because CFL learners are prone to different kinds of error which we can not imagine as a native speaker. Meanwhile, more errors can be caused by various Chinese input methods. As illustrated in Table 1, some errors can be found only in the essays of CFL learners, e.g. the 3rd and the last errors.

Error Types	Misspelled	Corrections
Homophone	聯合國公布	聯合國公佈
Near-homophone	好碼差不多一樣	號碼差不多一樣
Similar-shape	列如：家庭會變冷漠	例如：家庭會變冷漠
Others errors	每個禮拜1、3、5	每個禮拜一、三、五
	受了都少苦	受了多少苦

Table 1. Examples of spelling error

(3) Chinese language is continuously evolving, for example, traditional Chinese and simple Chinese may have different choices for the same word. In some cases, it is very difficult to distinguish them. Therefore, online high quality corpus is needed for decision-making.

To address the above challenges, we present a unified framework, named HANSpeller, to combine different methods for Chinese essays spelling detecting and correction. The contribu-

[1] http://www.cipsc.org.cn/clp2014/webpage/en/four_bakeoffs/Bakeoff2014cfp_ChtSpellingCheck_en.htm

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 133–138,
Wuhan, China, 20-21 October 2014

tion of our approach is as follows: (1) A HMM-based approach is used to segment sentences and generate candidates for sentences spelling correction. (2) Under this framework, all kinds of error types can be easily integrated for candidates generating. We collected some error types which only may be found in CFL learner essays, and add them into candidates generating process. And then ranking-based approach is used for choosing candidates for correction. (3) In order to address the evolving feature of Chinese, we not only collect high quality Taiwan web pages and also use search engine results to help decision-making on candidates.

The rest of the paper is organized as follow. In Section 2, we introduce related work on spelling checking. Then our unified framework approach is discussed in detail in Section 3. Section 4 presents the detailed experiment on the task. Section 5 concludes the paper and discusses future work.

2 Related work

Chinese essays spelling correction as a special kind of spelling correction research effort has been promoted by efforts such as the SIGHAN bake-offs (Wu et al., 2013).

Spelling correction was first proposed for English (Peterson, 1980). And it can be mainly divided into single word and context-sensitive spelling correction technology.

For the single word spelling error, it commonly uses dictionary-based method. It matches the original word with all the words in dictionaries to determine whether the word has spelling errors.

For the context-sensitive spelling errors, there are two major kinds of processing methods: Rule-based methods and Statistics-based methods. Rule-based methods use some rules generated fromrelevant grammars, the collocation of words, syntactic knowledge, etc, for spelling correction. Mangu and Bill (1997) proposed a transition-based learning method for spelling correction. Their methods generated three types of rules from training data, which constructs a high performance and concise system for English. A statistics-based method first finds related candidates, and then ranks the candidates based on the statistical model. Atwell and Elliott (1987) used n-gram and part-of-speech language models for spelling correction. Cucerzan and Brill (2004) presented an iterative process for query spelling check, using a query log and trust dictionary. And the noisy channel mode is used to select the best correction. Ahmad and Kondrak (2005) also learned a spelling error model from search query logs to improve the quality of query spelling check. Li et al. (2006) applied distributional similarity based models for query spelling correction. Gao et al. (2010) presented a large scale ranker-based system for search query spelling correction, the ranker uses web scale language models and many kinds of features for better performance, including: surface-form similarity, phonetic-form similarity, entity, dictionary, and frequency features. Microsoft (2010) provides Microsoft web n-gram services. Google (2010) has developed a Java API for Google spelling check service.

As for Chinese spelling correction, an early work was by (Chang, 1995), which used a character dictionary of similar shape, pronunciation, meaning, and input-method-code to deal with the spelling correction task. The system replaced each character in the sentence with the similar character in dictionary and calculated the probability of all modified sentences based on language model.

Zhang et al. (2000) introduced a method that can handle not only Chinese character substitution, but also insertion and deletion errors. They distinguished the way of matching between the Chinese and English, thus largely improved the performance over the work of (Chang 1995).

Huang et al. (2007) used a word segmentation tool (CKIP) to generate correction candidates, and then to detect Chinese spelling errors.

Hung et al. (2008) introduced a method which used the manually edited error templates to correct errors.

Zheng et al. (2011) found the fact that when people typed Chinese Pinyins, there are several wrong types. Then they introduced a method based on a generative model and the typed wrong types to correct spelling errors.

Liu et al. (2011) pointed out visually and phonologically similar characters are major factors for errors in Chinese text. And by defining appropriate similarity measures that consider extended Cangjie codes, visually similar characters can be quickly identified.

Note that all spelling correction methods require lexicons and/or language corpora. And Chinese essays spelling correction has some different concerns with query spelling correction. In our approach, we adopt the method based on statistics combining with lexicon and rule-based methods.

3 The Unified Framework for Chinese Spelling Correction

In this section we present a unified framework, named HANSpeller, for Chinese spelling correction based on extended HMM and ranking models. The major idea of our approach is to model the spelling correction process as a ranking and decision-making problem. Generally speaking, our approach has four major steps: Firstly the spelling correction process generates lots of candidates for sentences being checked; and then a ranking algorithm is applied to rank top-k correction candidates for later decision; the third step conducts rule-based analysis for specific correction task, e.g. the correction rule of the usage of three confusable words "的", "地" and "得". Finally, the system makes decision whether to output the correction or not based on the previous output and global constrains.

The system architecture is illustrated in Figure 1. This framework provides a unified approach for spelling correction tasks, which can tailored to different scenarios and can be regarded as a language independent framework. To move to another language scenario, you only need to collect some language related corpus, but you don't need to be a language expert.

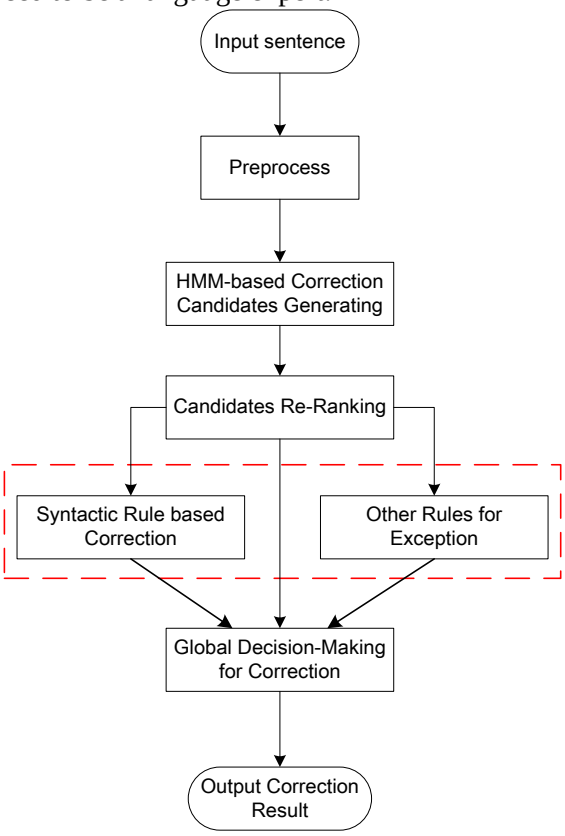

Figure1. The Unified Framework (HANSpeller) for Chinese Spelling Correction

3.1 Generating Candidates

Generating candidates of spelling correction task is the basic part for the whole task, because it determines the upper bound of precision and recall rate of the approach. The spelling correction problem can be typically formulated under the framework of noisy channel model. According to such a model, the spelling correction task is to find the correction with the highest probability of yielding the misspelled input sentence. Formally, given an "observed" sentence S which might contain error characters, we need to find the corrected sentence $\hat{C}$ with the highest probability of different replacement C. Symbolically, it is represented by:

$$\hat{C} = arg\,max_c\, p(C|S) \qquad (1)$$

By applying Bayes' Rule, we can rewrite Formula 1 as:

$$\hat{C} = arg\,max \frac{p(S|C)p(C)}{p(S)}$$
$$\approx arg\,max\,(p(S|C)p(C) \qquad (2)$$

where p(S|C), called the "error model", represents the chance that a correct Chinese character could be written to the wrong one, while p(C) is the n-gram language model which evaluates the quality of the corrected Chinese sentence.

To solve the above problem, the HMM approach can be used. And the spelling correction can then be ranked by multiplying the error model and language model.

The above one step method for Chinese essays spelling correction faces the following challenges: (1) For high quality spelling correction, the training of HMM is not a trivial task. (2) The long-span dependency in sentences makes first-order hidden Markov model not enough to catch context information. (3) Too many candidates make the algorithm not efficient enough, and right corrections may be concealed by the wrong corrections.

To address the above issues, some extensions have been made on HMM-based spelling correction approach. Firstly, the HMM-based method is used only for candidates generating, not for finally output correction generating. And all kinds of possible error transformations can be integrated into the framework of HMM approach, so as to get high recall rate. Secondly, higher-order hidden Markov model is used to capture long-span context dependency. Thirdly, a pruning dynamic programming algorithm is adopted to dynamical-

ly select the best correction candidates for each round of sentence segmentation and correction.

3.2 Ranking Candidates

In the candidates generation phase, top-k best candidates for a sentence are generated, but the HMM-based framework does not have the flexibility to incorporate a wide variety of features useful for spelling correction, such as the online search results and CKIP Parser results, which can significantly improve the precision of spelling correction.

Given the original sentence, our system first creates a list of candidate sentences. The candidates in the list will be re-ranked at this stage based on the confidence score generated by a ranker, herein by a SVM classifier. We choose the top-2 candidates in the re-ranked candidate list to make the final decision.

We use a lot of features in the re-ranking phase. The features can be grouped into the following categories:

1) **Language model features**, which calculate the n-gram probability of a candidate sentence.
2) **Dictionary features**, which check whether parts of a candidate sentence can match to one or more words or idioms in the dictionaries.
3) **Edit Distance features**, which compute the edit number and its weight, from the original sentence to the candidate sentence.
4) **Segmentation features**, which use the results of the maximum matching segmentation algorithm and that of CKIP Parser segmentation.
5) **Online Resources features**, which use Bing or other search engine's search results, when submitting the spelling correction part and the corresponding part of the original sentence to the search engine.

3.3 Rule-based Correction for Errors

As illustrated in Figure 1, the third step conducts rule-based analysis for specific correction task. One of most common errors is the usage of three confusable words "的", "地" and "得". To correct such common errors, syntactic analysis is needed. For other errors, some other specific rules can be developed for them.

The following sentence contains an error of Chinese syntax:

今天/我/穿著/剛/買/地/新/衣服。

Here the character "地" should be corrected to another character "的". To deal with such kind of errors, sentence parsing must be done before the syntactic rules are applied to check and correct such errors. We have summarized three rules according to Chinese grammar as follows:

1) The Chinese character "的" is the tag of attributes, generally used in front of subjects and objects. Words in front of "的" are generally used to modify, restrict things behind "的".
2) The Chinese character "地" is adverbial marker, usually used in front of predicates(verbs, adjectives). Words in front of "地" are generally used to describe actions behind "地".
3) The Chinese character "得" marks the complement, generally used behind predicates. The part follows "得" is generally used to supplement the previous action.

3.4 Decision-making on Corrections

Through the above processing steps, top candidates for each sub-sentence have been generated. To make the final decision on spelling correction, global constrains should be considered, including the whole error rate of the corpus, which error type should be paid more weight than others, which sub-sentence corrects should be output, etc. Combining the above constrains together, the system determines the final decision for spelling corrections.

4 Experiment and Evaluation

4.1 Experimental Setting

The following corpora are used to train our model, including Taiwan Web as Corpus, SogouW dictionary, a traditional Chinese dictionary of words and idioms, a pinyin mapping table and a cangjie code table of common words. The details of them are described below.

1) Taiwan Web Pages as Corpus

As we known, Taiwan web pages contain high quality traditional Chinese text, so we gathered pages from the Web under .tw domain to build the corpus, containing around 3.2 million web pages. And then the content extracted from these pages is used to build traditional Chinese n-gram model, where n is from 2 to 4.

2) SogouW Dictionary

SougoW dictionary[2] is built from the statistical analysis of Chinese Internet corpus by Sogou Search Engine. It contains about 150,000 high-frequency words of the Chinese Internet. But words in the corpus are simple Chinese characters; it is then translated into traditional Chinese by Google translating service.

3) Chinese Words and Idioms Dictionary

As introduced in [Chiu *et al.* 2013], we also obtained the Chinese words [3]and Chinese idioms[4] published by Ministry of Education of Taiwan, which are built from the dictionaries and related books. There are 64,326 distinct Chinese words and 48,030 distinct Chinese idioms. And we combine these two dictionaries with SogouW dictionary to build our trie tree dictionary.

4) Pinyin and Canjie Code Tables

We collected more than 10000 pinyins of words commonly used in Taiwan to build the homophone and near-homophone words table, which will be used in candidate generation phase. In addition, cangjie code can be used to measure the form/shape similarity between Chinese characters. Therefore, we collected cangjie codes to build the table of Similar-form characters.

5) Segmentation Resources

Besides using the Maximum Matching Method for Chinese word segmentation, we also use the CKIP Parser results to help ranking the candidates. For example, the segmentation of "特續下滑" is "特/續/下滑" while "持續下滑" is "持續/下滑". Thus the segmentation results of wrong candidate sentence will have more words than the correct one.

6) Online Resource

In addition to the above, we use the Bing search results as one feature in candidates ranking phase, which improve the performance obviously. For example, the sentence "根據聯合國公布的數字" has several candidate sentences, one of them may be "根據聯合國公佈的數字". If we use Bing to search the error correction part and the corresponding part of the original sentence "聯合國公佈" and "聯合國公布", the search results will be obviously enough to identify the correct candidate sentence, because the first one is more popular than the second one on the web corpus.

[2] http://www.sogou.com/labs/dl/w.html

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

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

4.2 Evaluation Results

At the CLP-2014 bake-off, the evaluation task is to correct errors in sentences. It is divided into two related subtasks. One is error detection and the other is error correction. There are 1062 sentences with/without spelling errors. The evaluation metrics, including false positive rate, accuracy rate, precision rate, recall rate and F1-score, is provided by the Chinese Spelling Check Task group. The confusion matrix as follow is to help to calculate the related indicators.

Confusion Matrix		System Results	
		Positive (Error)	Negative (No Error)
Gold Standard	Positive	TF	FN
	Negative	FP	TN

Table 2. Confusion Matrix

Each index calculation is as follows:

False Positive Rate (FPR) = FP / (FP + TN)

Accuracy (A) =

(TP + TN) / (TP + TN + FP + FN)

Precision (P) = TP / (TP + FP)

Recall (R) = TP / (TP + FN)

F1-score = 2 * P * R / (P + R)

Our system showed good performance on the evaluation test. Among all 13 teams, our performance ranks second place. The two submitted test results are illustrated in Table 3. Meanwhile, since such an open test is an extremely challenging task, there is still much room for further improvement.

	RUN1		RUN2	
	Detection Level	Correction Level	Detection Level	Correction Level
FPR	0.1525		0.1563	
A	0.6149	0.5829	0.613	0.581
P	0.7148	0.676	0.7098	0.6706
R	0.3823	0.3183	0.3823	0.3183
F1	0.4982	0.4328	0.4969	0.4317

Table 3. Evaluation at CLP-2014 Bake-off

5 Conclusion and Future Work

This paper proposed a unified framework (HANSpeller) for Chinese essays spelling correction based on extended HMM and ranker-

based models. The rule-based strategy is used for further correction polishing, and for final decision on whether outputs the correction or not. Our approach has been evaluated at CLP-2014 bake-off on Chinese spelling correction task, and made good performance with ranking second among 13 teams.

Some interesting future works on Chinese spelling correction include: (1) collecting and considering more error types in the candidates generating process, (2) how to better dealing with the difference between traditional and simple Chinese.

Acknowledgments

This work was supported by the National Basic Research Program of China (Grant No. 2014CB340406), the National High Technology Research and Development Program of China (Grant No. 2014AA015204) and the NSFC for the Youth (Grant No. 61402442).

Reference

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.

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 Transactions on Asian Language Information Processing, 10(2), 10:1-39.

James L. Pterson. 1980. Computers programs for detecting and correcting spelling errors. Communications of the ACM, pp 23: 676-687

Lidia Mangu and Eric Bill. 1997. Automatic rule acquisition for spelling correction. In Proceeding of the 14th International Conference on Machine Learning, pp 187-194, San Francisco, CA

Eric Atwell and Stephen Elliott. 1987. Dealing with ill-formed English text. In the Computational Analysis of English: A Corpus-Based Approach, London.

Ahmad, F., and Kondrak, G. 2005. Learning a spelling error model from search query logs. In HLT_EMNLP, pp 955-962.

Li, M., Zhu, M., Zhang, Y. and Zhou, M. 2006. Exploring distributional similarity based models for query spelling correction. Proceedings of ACL 2006, pp 2025-1032.

Jianfeng Gao, Xiaolong Li, Daniel Micol, Chris Quirk, and Xu Sun. 2010. A large scale ranker-based system for search query spelling correction. The 23rd International Conference on Computational Linguistics 2010 (COLING 2010). pp 358-366.

Microsoft Microsoft web n-gram services. 2010. http://research.microsoft.com/web-ngram

Google. 2010. A Java API for Google spelling check service. http//code.google.com/p/google-api-spellingjava/

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

Lei Zhang, Ming Zhou, Chang-Ning Huang, and Hui-Hua Pan. 2000b. Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm. In Proceedings of the 38the Annual Meeting on Association for Computational Linguistics, pp 248-254, Morristwon NJ.

S. Cucerzan and E. Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In Proceedings of EMNLP, volume 4, pp 293-300.

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.

Y. Zheng, C. Li, and M. Sun. 2011. Chime: An efficient error-tolerant Chinese pinyin input method. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence.