HANSpeller++: A Unified Framework for Chinese Spelling Correction

Shuiyuan Zhang[123], **Jinhua Xiong**[12], **Jianpeng Hou**[123], **Qiao Zhang**[123], **Xueqi Cheng**[12]
[1] CAS Key Laboratory of Network Data Science and Technology
[2]Institute of Computing Technology, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
shuiyuanzhang@gmail.com,xjh@ict.ac.cn

Abstract

Increased interest in China from foreigners has led to a corresponding interest in the study of Chinese. However, the learning of Chinese by non-native speakers will encounter many difficulties, Chinese spelling check techniques for Chinese as a Foreign Language(CFL) learners is highly desirable. This paper presents our work on the SIGHAN-2015 Chinese Spelling Check task. The task focuses on spelling checking on Chinese essays written by CFL learners. We propose a unified framework called HANSpeller++ based on our previous HANSpeller for Chinese spelling correction. The framework consists of candidate generating,candidates re-ranking and final global decision making. Experiments show good performance on the test data of the task.

1 Introduction

The number of people learning Chinese as a Foreign Language (CFL) is booming in recent decades. Chinese is rated as one of the most difficult languages to learn for people whose native language is English, together with Arabic, Japanese and Korean. There are many difficulties when learning Chinese such as confusing four tones, many words that change their meanings based on what other words are around them. When CFL learners write Chinese essays, they are prone to generate more and diversified spelling errors than native language learners. Therefore, spelling correction tools to support such learners become very necessary and valuable.

As for spelling correction on Chinese essays of CFL learners, we are facing more challenges because of the uniqueness of Chinese language:

(1) Chinese characters number in the tens of thousands, many of them have same pronunciation or similar shape, it is easy to confuse these characters.

(2) There are no natural delimiters such as spaces between Chinese words, which may result in the error on word splitting, and accumulate the errors by the splitting.

(3) Chinese corpora for spelling correction, especial for public available ones, are rare, compared with English corpora. Such situation impedes more works on this practical topic.

(4) There are many different versions including simple Chinese and traditional Chinese. It is very difficult to distinguish them for CFL learners.

(5) The number of error types is more than that of other cases, because CFL learners are prone to different kinds of errors which we can not imagine as a native speaker.

To address the above challenges, we present a unified framework for Chinese essays spelling error detecting and correction. Our method combines different methods to improve performance. The main contributions compared with our previous work (Xiong et al., 2014) are:

(1) A HMM-based approach is used to segment sentences and generate candidates for sentences spelling correction. Furthermore, some error types which can be found in CFL learners essays frequently are added to the candidates generating process.

(2) A two stage filter process help to re-rank the candidates efficiently and accurately. The first stage filter enable us to filter out a lot of wrong candidates efficiently, and the second filter process help us to choose the most promising candidates accurately.

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 38–45,
Beijing, China, July 30-31, 2015.

In order to address evolving features of Chinese language, We crawl many web pages from some famous Taiwan websites as corpus, these high quality corpus is used to build the n-gram language model; and the online search resources are also used in the ranking stage, which can also improve the performance significantly.

The rest of the paper is organized as follows. We start with discussing related work in Section 2, followed by introducing our unified framework approach in Section 3, where we focus on the basic processes of our method. In Section 4, we present the detailed setup of the experimental evaluation and the results of the experiments. Finally, in Section 5, we come to conclude the paper and explore future directions.

2 Related work

In recent years, a lot work has been done in the spelling correction field. Chinese essays spelling correction as a special kind of spelling correction research effort has been promoted by efforts such as the SIGHAN bake-offs (Yu et al., 2014) (Wu et al., 2013) (Liu et al., 2011). Spelling correction aims at identifying the misspellings and choosing the optimal words as suggested corrections, and it can be mainly divided into single word spelling correction and context-sensitive spelling correction.

Single word spelling error commonly uses dictionary-based method. (Angell et al., 1983) introduced an automatic correction of misspellings using a trigram similarity measure. This method replace a word by that word in a dictionary which is the nearest neighbour of the misspelling.

For the context-sensitive spelling errors, there are two major kinds of processing methods:Rule-based methods and Statistics-based methods.

(Mangu and Brill, 1997) proposed a transition-based learning method for spelling correction. Their methods generated three types of rules from training data, which constructed a high performance and concise system for English.

(Mays et al., 1991) proposed a context based spelling correction method. This method statistic errors and is able to detect and correct some of these errors when they occur again in sentences.

(Golding and Roth, 1999) introduced an algorithm combining variants of Winnow and weighted-majority voting for context-sensitive spelling correction. When dealing test set which comes from a different corpus, this method can combines supervised learning on the training set with unsupervised learning on the test set.

With the development of Internet, online spelling correction service became available. (Suzuki and Gao, 2012) proposed a transliteration based character method using an approach inspired by the phrase-based statistical machine translation framework and get a good performance on online spelling correction.

Also,there are some online resources can be used for spelling checking. (Microsoft, 2010) provides web n-gram service on real-world web-scale data. (Google, 2013) provides Google books n-gram viewer,it displays how some phrases have occurred in a corpus of books.

As to Chinese Spelling correction, the situation is quite different. Chinese is a character based language, there are many potentially confusing aspects to this language. The nature of Chinese makes the correction much more difficult than that of English.

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.

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

Some hybrid approach is applied to the Chinese spelling correction. (Jin et al., 2014) integrated three models including n-gram language model, pinyin based language model and tone based language model to improve the performance of Chinese checking spelling error system.

In our system, we need to detect and correct spelling errors on Chinese essays written by CFL learners. It has some different concerns with query text or query spelling correction. Noting that spelling correction methods require lexicons and/or language corpora, we adopt the method based on statistics combined with lexicon and rule-based methods.

3 A Unified Framework for Chinese Spelling Correction

For this Chinese Spelling Check task, we propose a unified framework called HANSpeller++. The main improvement of HANSpeller++ is the candidate re-ranking module. For some features used in the re-ranking process will cost a lot time to generate, we introduce a new two stage filter model to re-rank the candidates efficiently and accurately.

The framework coverts this task to 2 main parts, the first part is to generate possible candidates for a given input sentence, the second part is to choose the most promising candidate to output.Figure 1 shows the architecture of the unified framework.

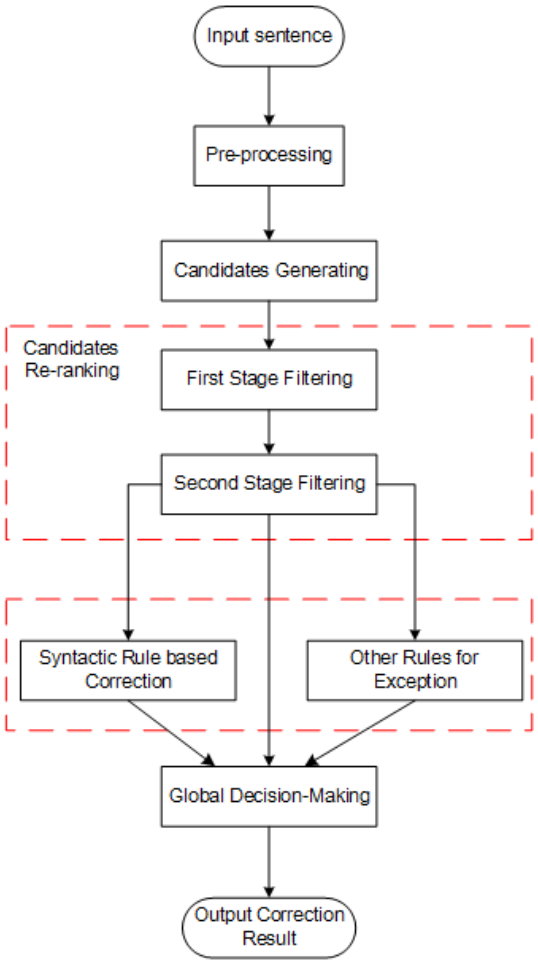

Figure 1: A unified framework for Chinese spelling correction(HANSpeller++).

It separates the Chinese spelling correction system into four major steps. First is to preprocess the input sentence to some sub sentences, then use the extended HMM model to generate top-k candidates for these sub sentences. We then use a two stage filter method to re-rank the correction candidates for later decision. Rule-based correction method is then used to consider some situation such as the usage of three confusable words "的", "地" and "得". Finally, we use global decision method to output the original sentence directly or the most promising candidate based on some constraint and the performance in previous step.

This framework provides a unified approach for spelling correction tasks, which can be regarded as a language independent framework and can be tailored to different scenarios. To move to another scenario, you need to prepare a language related corpus, but you do not need to be an expert of that language.

3.1 Data Preprocessing

Data provided by organizer is in the form of long sentences, and contains some non-Chinese characters.In our framework, sub sentence is the basic unit of the error correction process. We split long sentences into sub sentences by punctuation, and remove non-Chinese characters determined by its unicode code.

The policy of this task is an open test. We also use CLP-2014 CSC Datasets and SIGHAN-2013 CSC Datasets as our training data. The training data include real mistake by CFL learners and its correction, we treat this as confusion pair. Character-based confusion pair and word-based confusion pair are extracted from the whole training data, these 2 confusion pair sets will be used in the candidates generating process.

3.2 Candidates Generating

Generating candidates is the basic part for the whole task, for it determines the upper bound of recall rate of the approach.

Figure 1 shows the flow chart of the candidates generating module.

We first initialize a fixed size priority queue for a certain input sub sentence, this queue is used to store intermediate sub sentences.

For each character of sentences in the priority queue, we try to replace it by its candidate character. The possible candidate character include its homophone, near-homophone, similar shape character and confusion pair. Confusion pair set is extracted from the given training data, we collect the wrong character written by CFL learners and its corresponding correct character as a confusion pair.

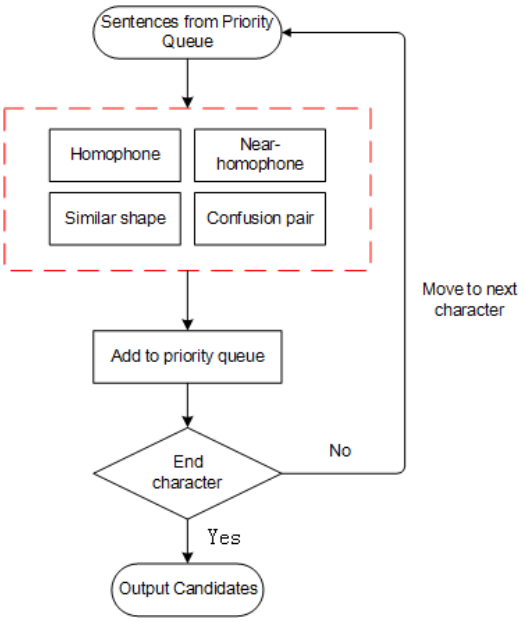

Figure 2: Flow chart of candidates generating module.

Different weight will be set to these different replacement type. Candidates generated by character replacement will be enqueued to the priority queue. When the queue is full candidate with low priority will be discarded, "priority" is defined as Follows.

Let $S = w_1 w_2 w_3 \dots w_N$ be a sub sentence needed to be corrected, where each item w_i is a character.$C = \tilde{w}_1 \tilde{w}_2 \tilde{w}_3 \dots \tilde{w}_{|r|} \dots w_N$ is a candidate generated by replacing the r-th character. The priority of this candidate defined as $P(C|S)$. According to noisy channel model, probability $P(C|S)$ can be expressed as Equation 1.

$$P(C|S) = \frac{P(S|C)P(C)}{P(S))} \quad (1)$$

As $P(S)$ is always same for candidates of the same raw input, Equation 1 can be simplified as Equation 2.

$$log(P(C|S)) \propto log(P(S|C) + log(P(C)) \quad (2)$$

Conceptually, Equation 2 can be calculated approximately by using edit distance and n-gram language model. Priority finally defined as Equation 3.

$$priority = \alpha * log(P(C)) + \beta * edit\_dist \quad (3)$$

3.3 Candidates Re-Ranking

In the candidates generating phase, a lot candidates for a sentence are generated. But at most one candidate for a input sentence is correct,the goal of this re-ranking module is to discard a lot of wrong candidates. We convert this ranking problem to a classification problem, the right candidates are regarded as positive samples while the wrong candidates are regarded as negative samples.

A lot of features can be used in the classifier, but some features are too time-consuming. For a given sub sentence, we may get hundreds of candidates, it will waste a lot time to extract all features for these candidates. In view of this situation, we proposed a two stage filter method. The main purpose of this method is to pre-filter the candidates using a fast model with some simple features, a more accurate model with more features will be used for candidates after filtration.

In the first stage, we train a simple but fast logistic regression classifier with some simple features, generating these features will not be too time-consuming. Then the candidates in the list will be filtered up to 20 at this stage based on the probability score generated by the trained classifier. Features used in this stage list below.

- **Language model features**: which calculates the n-gram text probability of candidate sentences and the original sentence.
- **Dictionary features**: which counts the number and proportion of phrases and idioms in candidates after segmentation according to our dictionaries.
- **Edit distance features**: which compute the edit number and its weight, from the original sentence to candidate sentences. Here different edit operations are given different edit weights.
- **Segmentation features**: which uses the results of the Maximum Matching Algorithm and the CKIP Parser segmentation.

In the second stage, We add some time-consuming features to obtain a more accurate model. For the candidate count decreases a lot after the first filter stage, these time-consuming features are acceptable. We choose top-5 candidates after this stage. Features used in the second stage list below.

- **Web based 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.

- **Translation features**: which use Yandex to compare English translation of the original sentence and each candidate sentence. Right candidate sentence tend to have more fluent English translation.

- **Microsoft Web N-Gram Service probability**: which compute the English translation N-gram probability by using Microsoft Web N-Gram Service.Traditional Chinese corpora for spelling correction, especially for public available ones, are rare. Microsoft Web N-Gram Service provide N-gram probability on real-world web-scale data, so we take advantage of this service by using English translation of each candidate.

In this two stage filter method, a wide variety of features are taken into account in order to obtain the candidate sentences accordance with the actual quality of candidates as much as possible. The first stage filter enhances the overall speed, and the second stage filter can help to improve the performance of final spelling correction. After this re-ranking module, top-5 candidates for a sub sentence will be output to the final global decision.

3.4 Rule-based Correction

After candidates re-ranking, some common errors are still difficult to be distinguished, such as the usage of three confusable words "的", "地", "得". In order to correct such errors, syntactic analysis is necessary to develop. The following sentence contains an error of Chinese syntax:

今天/我/穿着/刚/买/地/新/衣服。

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

The Chinese character "的" is the tag of attributes, which generally used in the front of subjects and objects. Words in front of "的" are generally used to modify, restrict things behind "的".

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".

The Chinese character "得" makes the complement, generally used behind predicates. The part follows "得" is generally used to supplement the previous action.

Another common error is the usage of "他", "她", "它". In the following sentence the character "他" should be corrected to another character "她", for it refers to the word "媽媽" which is a female.

媽 媽/不 會/說/中 文， 而 且/他/不 要/一 個 人/在/家裡。

We collect some simple rules that map keyword to one of the character "ta",such as "姐姐" maps to "她","父亲" maps to "他". When a gender specific word shows in the previous sub sentence, we use the keyword map as the basis for the character "ta".

There is also another situation that the character "ta" shows exactly in front of a gender specific word, such as "他女朋友","她男朋友".

The usage of "ta" is far more complex, we only deal with some obvious cases using simple rules. More complicated situation can be processed by using syntactic analysis.

In addition, some other specific rules are also needed to improve the final performance, which can be concluded from the training data and corpus.

3.5 Global Decision Making

Through the above processing steps, We get top-5 candidates for each sub-sentence. To make the final decision on spelling correction, some global constrains should be considered.

First, we filter out some candidates ,If the n-gram prob of the raw sentence is close to the most promising candidate, the raw sentence will be output. The closeness is measured relatively.

Then the rest candidates is sorted based on a combination of factors. The probability score in the second filter stage is a key factor, for it consider many useful features. Replacement type in the candidates generating process is another factor that can influence the decision making. We set different weights for different types of spelling errors by experience. For example, the confusion pair replacement need to be paid more weight than

others, as these replacement are really happen frequently in the training data, and we assume the test data is consistent with the training data.

Also, we use some global constraints to limit the number of errors. If there are more than 2 errors in a sub sentence, this candidate will be dropped. If there are more than 3 sub sentence errors in a long sentence, this long sentence will not be modified. These rules will increase the precision rate.

Finally, the precision rate and recall rate is balanced by controlling the number of error sentences.

In this task, we regulate some constraints and weights to get our final runs, this step has a great influence on the final performance.

4 Experiments

4.1 Resources

The following corpora are used in our experiment,including Taiwan Web as Corpus, 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.

- **SIGHAN Datasets**

  We extract confusion set from the given training data, but the given training data is not enough, so we also use CLP-2014 CSC Datasets and SIGHAN-2013 CSC Datasets as our training data. Character-based confusion pair and word-based confusion pair are extracted from the whole training data, these 2 confusion pair sets will be used in the candidates generating process.

- **Taiwan Web Pages as Corpus**

  we try to find Taiwan webs whose pages contain high quality traditional Chinese text, to build the corpus. We gathered pages from the artificial selected Webs under .tw domain to build the corpus. And then the content extracted from these pages is used to build traditional n-gram language model, where n is from 2 to 4.

- **Chinese Words and Idioms Dictionary**

  As introduced in (Chiu et al., 2013), we also obtained the Chinese words and Chinese idioms 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.

- **Pinyin and Cangjie Code Table**

  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.

- **Web based Resources**

  We use some web based resources to improve the performance. These resources include CKIP online parser, Bing search service, Yandex translate service and Microsoft Web N-Gram Service. In order to improve efficiency, these resources are only used in the second stage of candidate re-ranking process.

4.2 Evaluation

The criteria for judging correctness is divided into two levels. One is detection level and the other is correction level. For detection level, all locations of incorrect characters in a given passage should be completely identical with the gold standard. For correction level, all locations and corresponding corrections of incorrect characters should be completely identical with the gold standard.

$$FalsePositiveRate = \frac{FP}{FP + TN} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

Team	False Positive Rate	Accuracy	Precision	Recall	F1
CAS*	0.1309	0.7009	0.8027	0.5327	0.6404
NCTU+NTUT	0.1327	0.6018	0.7171	0.3364	0.4579
NTOU	0.5727	0.4227	0.422	0.4182	0.4201

Table 1: Top 3 performance in Detection Level.

Team	False Positive Rate	Accuracy	Precision	Recall	F1
CAS*	0.1309	0.6918	0.7972	0.5145	0.6254
NCTU+NTUT	0.1327	0.5645	0.6636	0.2618	0.3755
NTOU	0.5727	0.39	0.3811	0.3527	0.3664

Table 2: Top 3 performance in Correction Level.

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

Table 3: Confusion Matrix.

The evaluation metrics, including false positive rate, accuracy rate, precision rate, recall rate and F1-score, are used in this task. Formula of these indicators are listed in Equation 4-8. Table 3 is confusion matrix which help to calculate the related indicators.

There are 1100 sentences with/without spelling errors on the evaluation test. Detection level results illustrated in Table 1, correction level results illustrated in Table 2. Our performance ranks first place among all participating teams, which means that our method is feasible. Meanwhile, since such an open test is an extremely challenging task, there is still much room for further improvement.

5 Conclusion

This paper propose a unified framework called HANSpeller++ based on our previous HANSpeller. Candidate generating, candidates re-ranking and final global decision making are included in this framework, some rule-based strategies are used to improve the performance. Our approach has been evaluated at SIGHAN-2015 Chinese Spelling Check task, and achieved a good result.

Some interesting future works on Chinese spelling correction include: (1) Some more valuable features can be added in the re-ranking process. (2) Using machine learning method to make global decision is worth trying. (3) Implementing an online toolkit and service for Chinese spelling correction is a stimulator of this empirical research topic.

Acknowledgments

This research was supported by the National High Technology Research and Development Program of China (Grant No. 2014AA015204), the National Basic Research Program of China (Grant No. 2014CB340406), the NSFC for the Youth (Grant No. 61402442) and the Technology Innovation and Transformation Program of Shandong (Grant No.2014CGZH1103).

References

Richard C Angell, George E Freund, and Peter Willett. 1983. Automatic spelling correction using a trigram similarity measure. *Information Processing & Management*, 19(4):255–261.

Chao-Huang Chang. 1995. A new approach for automatic chinese spelling correction. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, volume 95, pages 278–283. Citeseer.

Hsun-wen Chiu, Jian-cheng Wu, and Jason S Chang. 2013. Chinese spelling checker based on statistical machine translation. In *Sixth International Joint Conference on Natural Language Processing*, page 49.

Andrew R Golding and Dan Roth. 1999. A winnow-based approach to context-sensitive spelling correction. *Machine learning*, 34(1-3):107–130.

Google. 2013. Ngram viewer. `https://books.google.com/ngrams`.

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

Peng Jin, Xingyuan Chen, Zhaoyi Guo, and Pengyuan Liu. 2014. Integrating pinyin to improve spelling errors detection for chinese language. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01*, pages 455–458. IEEE Computer Society.

C-L Liu, M-H Lai, K-W Tien, Y-H Chuang, S-H Wu, and C-Y Lee. 2011. Visually and phonologically similar characters in incorrect chinese words: Analyses, identification, and applications. *ACM Transactions on Asian Language Information Processing (TALIP)*, 10(2):10.

Lidia Mangu and Eric Brill. 1997. Automatic rule acquisition for spelling correction. In *ICML*, volume 97, pages 187–194. Citeseer.

Eric Mays, Fred J Damerau, and Robert L Mercer. 1991. Context based spelling correction. *Information Processing & Management*, 27(5):517–522.

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

Hisami Suzuki and Jianfeng Gao. 2012. A unified approach to transliteration-based text input with online spelling correction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 609–618. Association for Computational Linguistics.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese spelling check evaluation at sighan bake-off 2013. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*, pages 35–42. Citeseer.

Jinhua Xiong, Qiao Zhao, Jianpeng Hou, Qianbo Wang, Yuanzhuo Wang, and Xueqi Cheng. 2014. Extended hmm and ranking models for chinese spelling correction. *CLP 2014*, page 133.

Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2014. Overview of sighan 2014 bake-off for chinese spelling check. *CLP 2014*, page 126.