An Improved Graph Model for Chinese Spell Checking*

Yang Xin[1,2], **Hai Zhao**[1,2,†], **Yuzhu Wang**[1,2] **and Zhongye Jia**[1,2]
[1]Center for Brain-Like Computing and Machine Intelligence,
Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
xuechen.xy@gmail.com, zhaohai@cs.sjtu.edu.cn,
hfut0830@sjtu.edu.cn, jia.zhongye@gmail.com

Abstract

In this paper, we propose an improved graph model for Chinese spell checking. The model is based on a graph model for generic errors and two independently-trained models for specific errors. First, a graph model represents a Chinese sentence and a modified single source shortest path algorithm is performed on the graph to detect and correct generic spelling errors. Then, we utilize conditional random fields to solve two specific kinds of common errors: the confusion of "在" (at) (pinyin is 'zai' in Chinese), "再" (again, more, then) (pinyin: zai) and "的" (of) (pinyin: de), "地" (-ly, adverb-forming particle) (pinyin: de), "得" (so that, have to) (pinyin: de). Finally, a rule based system is exploited to solve the pronoun usage confusions: "她" (she) (pinyin: ta), "他" (he) (pinyin: ta) and some others fixed collocation errors. The proposed model is evaluated on the standard data set released by the SIGHAN Bake-off 2014 shared task, and gives competitive result.

1 Introduction

Spell checking is a routine processing task for every written language, which is an automatic mechanism to detect and correct human spelling errors. Given sentences, the goal of the task is to return the locations of incorrect words and suggest the correct words. However, Chinese spell checking (CSC) is very different from that in English or other alphabetical languages from the following ways.

Usually, the object of spell checking is words, but "word" is not a natural concept in Chinese, since there are no word delimiters between words in Chinese writing. An English "word" consists of Latin letters. While a Chinese "word" consists of characters, which also known as "漢字" (Chinese character) (pinyin[1] is 'han zi' in Chinese). Thus, essentially, the object of CSC is misused characters in a sentence. Meanwhile, sentences for CSC task are meant to computer-typed but not those handwritten Chinese. In handwritten Chinese, there exist varies of spelling errors including non-character errors which are probably caused by stroke errors. While in computer-typed Chinese, a non-character spelling error is impossible, because any illegal Chinese characters will be filtered by Chinese input method engine so that CSC never encounters "out-of-character (OOC)" problem. Thus, the Chinese spelling errors come from the misuse of characters, not characters themselves.

Spelling errors in alphabetical languages, such as English, are always typically divided into two categories:

- The misspelled word is a non-word, for example "come" is misspelled into "cmoe";

*This work was partially supported by the National Natural Science Foundation of China (No. 60903119, No. 61170114, and No. 61272248), the National Basic Research Program of China (No. 2013CB329401), the Science and Technology Commission of Shanghai Municipality (No. 13511500200), the European Union Seventh Framework Program (No. 247619), the Cai Yuanpei Program (CSC fund 201304490199 and 201304490171), and the art and science interdiscipline funds of Shanghai Jiao Tong University (A study on mobilization mechanism and alerting threshold setting for online community, and media image and psychology evaluation: a computational intelligence approach).

†Corresponding author.

[1] Pinyin is the official phonetic system for transcribing the sound of Chinese characters into Latin script.

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 157–166,
Wuhan, China, 20-21 October 2014

- The misspelled word is still a legal word, for example “come” is misspelled into “cone”.

While in Chinese, if the misspelled word is a non-word, the word segmenter will not recognize it as a word, but split it into two or more words with fewer characters. For example, if “你好世界” in Example 1 of Table 1 is misspelled into “你好世節”, the word segmenter will segment it into “你好/世/節” instead of “你好/世節”. For non-word spelling error, the misspelled word will be mis-segmented.

Name	Example 1	Example 2
Golden	你好/世界	好好/地/出去/玩
Misspelled	你好/世/節	好好/的/出去/玩
Pinyin	ni hao shi jie	hao hao de chu qu wan
Translation	hello the world	enjoy yourself outside

Table 1: Two examples for Chinese spelling error. Both examples have the same pinyin.

Thus CSC cannot be directly applied those edit distance based methods which are commonly used for alphabetical languages. CSC task has to deal with word segmentation problem first, since misspelled sentence could not be segmented properly by word segmenter.

There also exist Chinese spelling errors which are unrelated with word segmentation. For example, “好好地出去玩” in Example 2 of Table 1 is misspelled into “好好的出去玩”, but both of them have the same segmentation. So it is necessary to perform further specific process.

In this paper, based on our previous work (Jia et al., 2013b) in SIGHAN Bake-off 2013, we describe an improved graph model to handle the CSC task. The improved model includes a graph model for generic spelling errors, conditional random fields (CRF) for two special errors and a rule based system for some collocation errors.

2 Related Work

Over the past few years, there were many methods proposed for CSC task. (Sun et al., 2010) developed a phrase-based spelling error model from the clickthrough data by means of measuring the edit distance between an input query and the optimal spelling correction. (Gao et al., 2010) explored the ranker-based approach which included visual similarity, phonological similarity, dictionary, and frequency features for large scale web search. (Ahmad and Kondrak, 2005) proposed a spelling error model from search query logs to improve the quality of query. (Han and Chang, 2013) employed maximum entropy models for CSC. They trained a maximum entropy model for each Chinese character based on a large raw corpus and used the model to detect the spelling errors.

Two key techniques, word segmentation (Zhao et al., 2006a; Zhao and Kit, 2008b; Zhao et al., 2006b; Zhao and Kit, 2008a; Zhao and Kit, 2007; Zhao and Kit, 2011; Zhao et al., 2010) and language model (LM), are also popularly used for C-SC. Most of those approaches can fall into four categories. The first category consists of the methods that all the characters in a sentence are assumed to be errors and an LM is used for correction (Chang, 1995; Yu et al., 2013). (Chang, 1995) proposed a method that replaced each character in the sentence based on a confusion set and computed the probability of the original sentence and all modified sentences according to a bigram language model generated from a newspaper corpus. The method based on the motivation that all the typos were caused by either visual similarity or phonological similarity. So they manually built a confusion set as a key factor in their system. Although the method can detect misspelled words well, it was very time consuming for detection, generated too much false positive results and was not able to refer to an entire paragraph. (Yu et al., 2013) developed a joint error detection and correction system. The method assumed that all characters in the sentence may be errors and replaced every character using a confusion set. Then they segmented all new generated sentences and gave a score of the segmentation using LM for every sentence. In fact, this method did not always perform well according to (Yu et al., 2013).

The second category includes the methods that all single-character words are supposed to be errors and an LM is used for correction, for example (Lin and Chu, 2013) . They developed a system which supposed that all single-character words may be typos. They replaced all single-character words by similar characters using a confusion set and segmented the newly created sentences again. If a new sentence resulted in a better word segmentation, spelling error was reported. Their system gave good detection recall but low false-alarm rate.

The third category utilizes more than one approaches for detection and an LM for correction. (Hsieh et al., 2013) used two different systems for error detection. The first system detected error characters based on unknown word detection and LM verification. The second one solved error detection based on a suggestion dictionary generated from a confusion set. Finally, two systems were combined to obtain the final detection result. (He and Fu, 2013) divided typos into three categories which were character-level errors (CLEs), word-level errors (WLEs) and context-level errors (CLEs), and three different methods were used to detect the different errors respectively. In addition to using the result of word segmentation for detection, (Yeh et al., 2013) also proposed a dictionary-based method to detect spelling errors. The dictionary contained similar pronunciation and shape information for each Chinese character. (Yang et al., 2013) proposed another method to improve the candidate detections. They employed high confidence pattern matching to strengthen the candidate errors after word segmentation.

The last category is formed by the methods which use word segmentation for detection and different models for correction (Liu et al., 2013; Chen et al., 2013; Chiu et al., 2013). (Liu et al., 2013) used support vector machine (SVM) to select the most probable sentence from multiple candidates. They used word segmentation and machine translation model to generate the candidates respectively. The SVM was used to rerank the candidates. (Chen et al., 2013) not only applied LM, but also used various topic models to cover the shortage of LM. (Chiu et al., 2013) explored statistical machine translation model to translate the sentences containing typos into correct ones. In their model, the sentence with the highest translation probability which indicated how likely a typo was translated into its candidate correct word was chosen as the final correction sentence.

3 The Revised Graph Model

The graph model (Jia et al., 2013b) of SIGHAN Bake-off 2013 is inspired by the idea of shortest path word segmentation algorithm which is based on the following assumption: a reasonable segmentation should maximize the lengths of all segments or minimize the total number of segments (Casey and Lecolinet, 1996). A directed acyclic graph (DAG) is thus built from the input sentence similar. The spelling error detection and correction problem is transformed to a single source shortest path (SSSP) problem on the DAG.

Given a dictionary $\mathbb{D}$ and a similar characters $\mathbb{C}$, for a sentence S of m characters $\{c_1, c_2, \ldots, c_m\}$, the original vertices V of the DAG in (Jia et al., 2013b) are:

$$\begin{aligned} V = &\{w_{i,j} | w_{i,j} = c_i \ldots c_j \in \mathbb{D}\} \\ &\cup \{w^k_{i,j} | w^k_{i,j} = c_i \ldots c'_k \ldots c_j \in \mathbb{D}, \\ &\qquad \tau \leq j - i \leq T, \\ &\qquad c'_k \in \mathbb{C}[c_k], k = i, i+1, \ldots, j\} \\ &\cup \{w_{-,0}, w_{n+1,-}\}. \end{aligned}$$

where $w_{-,0}$ = "<S>" and $w_{n+1,-}$ = "</S>" are two special vertices represent the start and end of the sentence.

However, the graph model cannot be applied to continuous word errors. Take the following sentence as an example, "健康" (health) (pinyin: jian kang) is misspelled into "建缸" (pinyin: jian gang). Because the substitution strategy does not simultaneously substitute two continuous characters.

- 然後，我是計劃我們到我家一個附近的 '建缸' (pinyin: jian gang) 中心去游泳。

  Translation after correction: And then, we plan to go swimming near my house.

For example, the substitution of "建缸" (pinyin: jian gang) may be "碱缸" (pinyin: jian gang), "建鋼" (pinyin: jian gang), "建行" (pinyin: jian hang) and so on, none of which is the desired correction. So we revise the construction method of the graph model. Considering efficiency, we only deal with the continuous errors with 2 characters. The revised V are:

$$\begin{aligned} V = &\{w_{i,j} | w_{i,j} = c_i \ldots c_j \in \mathbb{D}\} \\ &\cup \{w^k_{i,j} | w^k_{i,j} = c_i \ldots c'_k \ldots c_j \in \mathbb{D}, \\ &\qquad \tau \leq j - i \leq T, \\ &\qquad c'_k \in \mathbb{C}[c_k], k = i, i+1, \ldots, j\} \\ &\cup \{w^l | w^l = c'_l c'_{l+1} \in \mathbb{D}, \\ &\qquad c'_l, c'_{l+1} \in \mathbb{C}\} \\ &\cup \{w_{-,0}, w_{n+1,-}\}. \end{aligned}$$

With the modified DAG G, the "建缸" (pinyin: jian gang) is substituted as "健康" (health) (pinyin: jian kang), "峴港" (Danang) (pinyin: xian gang), "潛航" (submerge) (pinyin: qian hang) and so on, which have already contained the desired correction.

4 The Improved Graph Model

The graph model based on word segmentation in (Jia et al., 2013b) includes the revised graph model in section 3 still has its limitations. For a sentence, in the graph construction stage, the substitution is only applied to the situation that the number of words after segmenting has to be decreased, which means there exists new longer word after segmentation. In addition, if the segmentation result of a sentence is a single character, the graph model does not work, because a single character will not be substituted. For example in the following two sentences, the “他” (he) (pinyin: ta) in the first sentence should be corrected into “她” (she) (pinyin: ta) and the “的” (of)(pinyin: de) in the second sentence should be corrected into “地” (-ly, adverb-forming particle) (pinyin: de), however, the graph model does not work for this case.

- 雖然我不在我的國家，不能見到媽媽，可是我要給‘他’ (him) (pinyin: ta)打電話！

  Translation after correction: Though I’m not in my country so that I cannot see my mum, I would like to call her!

- 我們也不要想太多；我們來好好‘的’ (of) (pinyin: de)出去玩吧！

  Translation after correction: We would not worry too much, just enjoy ourselves outside now!

The graph model is also powerless for the error situation that the wrong character was segmented into a legal word. Take the following sentence as an example, the word “心裡” (in mind, at heart) (pinyin: xin li) will be not separated after the building the graph, so “裡” (pinyin: li) could not be corrected into “理” (pinyin: li).

- 我對心‘裡’ (pinyin: li)研究有興趣。

  Translation after correction: I’m interested in psychological research.

For the sake of alleviating the above limitations of the graph model, we utilize CRF model to deal with two kinds of errors, and a rule based system is established to cope with the pronoun errors: “她” (she) (pinyin: ta), “他” (he) (pinyin: ta) and collocation errors.

4.1 CRF Model

Two classifiers using CRF model are respectively trained to tackle the common character usage confusions: 在” (at) (pinyin: zai), 再” (again, more, then) (pinyin: zai) and “的” (of)(pinyin: de), “地” (-ly, adverb-forming particle) (pinyin: de), “得”(so that, have to) (pinyin: de). We assume that the correct character selection is related with its neighboring two words and part-of-speech (POS) tags. The classifiers are trained on a large five-gram token set which is extracted from a large POS tagged corpus. The feature selection algorithm is according to (Zhao et al., 2013; Wang et al., 2014; Jia et al., 2013a). The feature set for CRF model is as follows:

$$w_{j,-2}, pos_{j,-2}, w_{j,-1}, pos_{j,-1}, w_{j,0}, pos_{j,0},$$
$$w_{j,1}, pos_{j,1}, w_{j,2}, pos_{j,2}$$

where j is the token index to indicate its position, $w_{j,0}$ is the current candidate character and $pos_{j,0}$ is its POS tag. ICTCLAS (Zhang et al., 2003) is adopted for POS tagging.

A set of feature strings that we used are presented in Table 2. The labels for “的” (of) (pinyin: de), “地” (-ly, adverb-forming particle) (pinyin: de), “得”(so that, have to) (pinyin: de) are 1, 2, 3 and “在” (at) (pinyin: zai), “再” (again, more, then) (pinyin: zai) are 1, 2.

4.2 The Rule Based System

To effectively handle pronoun usage errors for “她” (she) (pinyin: ta), and “他” (he) (pinyin: ta) and other collocation errors, we design a rule based system extracted from the development set.

The Table 3 is the rules we set for solving the pronoun usage errors, where the $prefix[i]$ is the current word $w[i]$’s prefix in a sentence. For the others rules, we divide them into five categories, which are presented in Table 4 – Table 8. In Table 4, we only present several typical rules in Rule 3. The negation symbol “¬” in the Table 6 and Table 7 means that the word in corresponding position is not the one in the brackets. Each rule in the tables is verified by the Baidu[2] search engine. If the error situation is legally emerged in the search result, we will not correct the error any more.

[2] http://www.baidu.com/

Feature	Example1	Example2
$w_{j,-2}$	“來”	“和”
$w_{j,-1}$	“好好”	“你”
$w_{j,1}$	“出”	“一起”
$w_{j,-2}$,$w_{j,-1}$	“來”,“好好”	“和”,“你”
$w_{j,-2}$,$w_{j,-1}$,$w_{j,1}$	“來”,“好好”,“出”	“和”,“你”,“一起”
$w_{j,1}$,$w_{j,2}$	“出”,“去”	“一起”,“。”
$pos_{j,-2}$	*v*	*p*
$pos_{j,-1}$	*z*	*r*
$pos_{j,1}$	*v*	*s*
$pos_{j,-2}$,$pos_{j,-1}$	*v*,*z*	*p*,*r*
$pos_{j,-1}$,$pos_{j,1}$	*z*,*v*	*r*,*s*
$pos_{j,1}$,$pos_{j,2}$	*v*,*v* o	*s*,*w*
$pos_{j,-2}$,$pos_{j,-1}$,$pos_{j,1}$	*v*,*z*,*v*	*p*,*r*,*s*
$w_{j,-1}$,$pos_{j,1}$	“好好”,*v*	“你”,*s*
$pos_{j,-1}$,$w_{j,1}$	*z*,“出”	*r*,“一起”
$pos_{j,-2}$,$pos_{j,-1}$,$w_{j,1}$	*v*,*z*,“出”	*p*,*r*,“一起”

Table 2: Feature strings for sentences “我們來好好地出去玩吧！” and “我只要和你在一起。”.

$prefix[i]$ does not contain	$prefix[i]$ contains	$w[i]$	corrected $w[i]$
(媽 and 爸) or (她 and 他) or (母 and 父) or (女 and 男) or (太太 and 先生)	她 or 媽 or 母 or 女 or 妹 or 姊 or 姐 or 婆 or 阿姨 or 太太	他	她
她 or 媽 or 母 or 女 or 妹 or 姊 or 姐 or 婆 or 阿姨 or 太太	他 or 爸 or 父 or 男 or 哥 or 先生	她	他

Table 3: Specific rules for the pronouns “她、他” confusion.

$w[i]$	$pos[i+1]$	corrected $w[i]$
阿	w	啊
馬 or 碼	w	嗎
門	r, n	們
把	r, n	吧

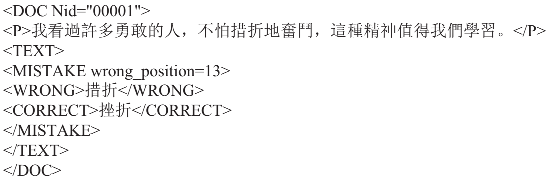
Table 4: Rule 1. The correction related with right neighbored POS tag.

5 Experiments

5.1 Data Sets and Resources

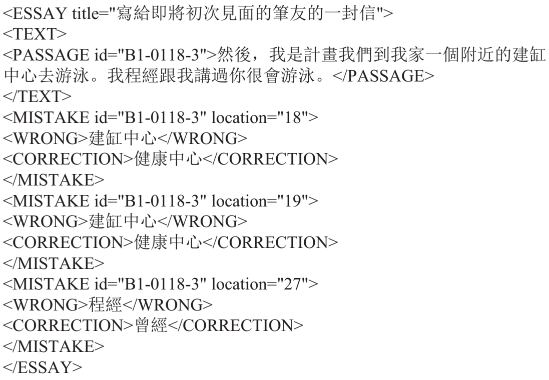
The proposed method is evaluated on the data sets of SIGHAN Bake-off shared tasks in 2013 and 2014. In Bake-off 2013, the sentences were collected from 13 to 14-year-old students' essays in formal written tests (Wu et al., 2013). In Bake-off 2014, the sentences were collected from Chinese as a foreign language (CFL) learners' essays selected from the National Taiwan Normal University (NTNU) learner corpus[3]. All the data sets are in traditional Chinese.

In Bake-off 2013, the essays were manually annotated with different labels (see Figure 1). There is at most one error in each sentence. However, the development set in Bake-off 2014 is enlarged and the error types (see Figure 2) are more diverse. More than one error might be in each sentence. And there exists continuous errors as in Figure 2.



```
<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>
```

Figure 1: A sample of annotated essay in Bake-off 2013.



```
<ESSAY title="寫給即將初次見面的筆友的一封信">
<TEXT>
<PASSAGE id="B1-0118-3">然後，我是計畫我們到我家一個附近的建缸
中心去游泳。我程經跟我講過你很會游泳。</PASSAGE>
</TEXT>
<MISTAKE id="B1-0118-3" location="18">
<WRONG>建缸中心</WRONG>
<CORRECTION>健康中心</CORRECTION>
</MISTAKE>
<MISTAKE id="B1-0118-3" location="19">
<WRONG>建缸中心</WRONG>
<CORRECTION>健康中心</CORRECTION>
</MISTAKE>
<MISTAKE id="B1-0118-3" location="27">
<WRONG>程經</WRONG>
<CORRECTION>曾經</CORRECTION>
</MISTAKE>
</ESSAY>
```

Figure 2: A sample of annotated essay in Bake-off 2014.

Statistical information on data sets is shown in Table 9. Three development sets are named as

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

$w[i]$	$suffix[i]$ contains	corrected $w[i]$
帶	帽，眼鏡，皮帶，手環	戴
負，府	費，錢，經濟，薪水	付
做，座	車，巴士，飛機，捷運，船，高鐵	坐

Table 5: Rule 2. The correction related with the current word's suffix.

$w[i-1]$	$w[i]$	$w[i+1]$	corrected $w[i]$
知	到	–	道
¬(內，肝，腎)	臟	–	髒
–	總	於	終
–	俄	¬(羅)	餓
改	以	改	一
¬(很)	多	很	都
心	理	¬(学，研)	裡
¬(一，二，這，兩，幾，草，壓)	根	¬(部，本，據，源，基，治，除	跟

Table 6: Rule 3. The correction related with neighbored words.

$w[i-2]$	$w[i-1]$	$w[i]$	$w[i+1]$	$w[i+2]$	corrected $w[i]$
林	依	神	–	–	晨
鋼	鐵	依	–	–	衣
游	泳	世	–	–	池
星	期	路	–	–	六
西	門	丁	–	–	町
–	–	很	不	得	恨
–	–	仍	在	了	扔
–	–	打	出	租	搭
–	–	機	程	車	計
–	–	¬(少)	子	化	少

Table 7: Rule 4. The correction related with two neighbored words.

$w[i-1]$	$w[i]$	$w[i+1]$	$w[i+2]$	$w[i+3]$	corrected $w[i]$ and $w[i+1]$
–	自	到	–	–	知道
–	式	式	–	–	試試
–	蘭	滿	–	–	浪漫
–	令	令	–	–	冷冷
–	排	排	–	–	拜拜
–	柏	柏	–	–	伯伯
–	莎	增	–	–	沙僧
–	旅	管	–	–	旅館
–	棒	組	–	–	幫助
–	想	心	–	–	相信
–	名	性	–	–	明星
–	頂	頂	大，有	名	鼎鼎
–	白	花	商	店	百貨
為	是	嗎	–	–	什麼

Table 8: Rule 5. Two words are simultaneously corrected.

DEV13, DEV14C and DEV14B and the test set is named as TEST14 respectively. In the DEV14B, there are 4624 errors, in which the statistics information of the three common character usage confusions in section 4 is shown in Table 10, so it is necessary to deal with them respectively.

The dictionary $\mathbb{D}$ used in SSSP algorithm is *SogouW*[4] dictionary from *Sogou inc.*, which is in simplified Chinese. The *OpenCC*[5] converter is used for simplified-to-traditional Chinese converting. Similar character set $\mathbb{C}$ provided by (Liu et al., 2010) is used to substitute the original words in the graph construction stage. The LM is built on the Academia Sinica corpus (Emerson, 2005) with IRSTLM toolkit (Federico et al., 2008). The CRF model is achieved by training and tuning on the Academia Sinica corpus with the toolkit *CRF++ 0.58*[6]. For Chinese word segmentation, the *ICTCLAS2011*[7] is exploited.

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

[5] http://code.google.com/p/opencc/

[6] https://code.google.com/p/crfpp/downloads/list

[7] http://www.ictclas.org/ictclas_download.aspx

Name			Data Size (lines)	Character number (k)
Development set	Bake-off 2013		700	29
	Bake-off 2014	C1	342	16
		B1	3004	149
Test set			1062	53

Table 9: Statistical information of data sets.

Error Type	Number	Percent (%)
在， 再	101	2.18
的， 地， 得	398	8.61
她， 他	101	3.98

Table 10: Three common character usage confusions in the DEV14B.

5.2 The Improved Graph Model

We treat the graph model without filters in Bake-off 2013 as our baseline in Bake-off 2014. The edge function is the linear combination of similarity and log conditional probability:

$$\omega^L = \omega_s - \beta \log P$$

where $\omega_0 \equiv 0$ which is omitted in the equation, and ω_s for different kinds of characters are shown in Table 11. The LM is set to bigram according to (Yang et al., 2012). Improved Kneser-Ney method is used for LM smoothing (Chen and Goodman, 1999).

Type	ω_s
same pronunciation same tone	1
same pronunciation different tone	1
similar pronunciation same tone	2
similar pronunciation different tone	2
similar shape	2

Table 11: ω_s used in ω^L.

We utilize the correction precision ($\mathcal{P}$), correction recall ($\mathcal{R}$) and F1 score ($\mathcal{F}$) as the metrics. The computational formulas are as follows:

- Correction precision:

$$\mathcal{P} = \frac{\text{number of correctly corrected characters}}{\text{number of all corrected characters}};$$

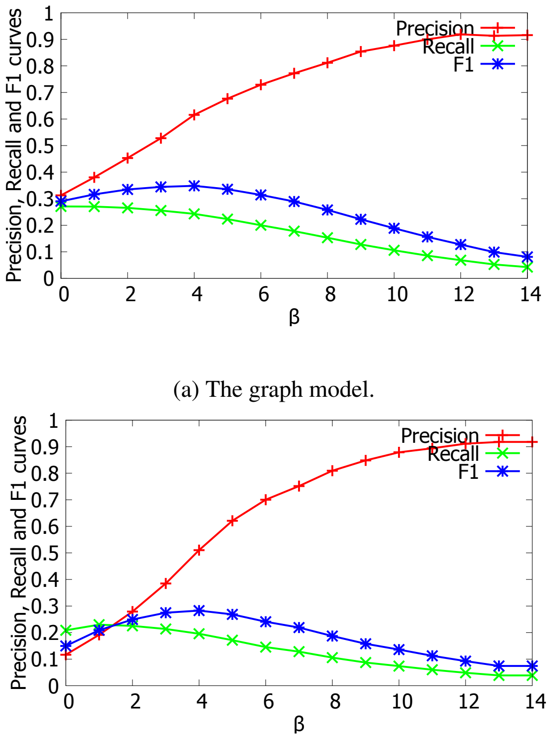
- Correction recall:

$$\mathcal{R} = \frac{\text{number of correctly corrected characters}}{\text{number of wrong characters of gold data}};$$

- F1 macro:

$$\mathcal{F} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P} + \mathcal{R}}.$$

We firstly use the revised graph model in section 3 to tackle the continuous word errors. The results achieved by the graph model and its revision on DEV14B with different β are shown in Figure 3 respectively. We can see that the result with the revised graph model is not improved, and even worse than the baseline. Therefore, for the improved graph model in Bake-off 2014, we remain use the graph model in Bake-off 2013 without any modification.



(a) The graph model.

(b) The revised graph model.

Figure 3: The results of the graph model and its revision on DEV14B.

To observe the performance of the improved graph model in detail, on the three development sets: DEV13, DEV14C, DEV14B, we report the results from the following settings:

1. *CRF*. We use the CRF model to process the common character usage confusions: “在” (at) (pinyin: zai), “再” (again, more, then) (pinyin: zai) and “的” (of) (pinyin: de),

	Dev13			Dev14C			Dev14B		
Model	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Graph (baseline)	0.802	0.6	0.686	0.790	0.238	0.366	0.729	0.2	0.314
+CRF	0.623	0.6	0.611	0.75	0.38	0.504	0.631	0.282	0.389
+CRF+Rule_Post	0.512	0.614	0.558	0.723	0.421	0.532	0.699	0.461	0.555
+CRF+Rule_Pre	0.526	0.614	0.567	0.75	0.38	0.504	0.706	0.479	0.571
+CRF+Rule_Pre+Rule_Post	0.51	0.611	0.556	0.723	0.421	**0.532**	0.706	0.484	**0.574**

Table 14: The results with different models.

“地” (-ly, adverb-forming particle) (pinyin: de), “得”(have to, get, obtain) (pinyin: de) on all development sets. The results achieved by the CRF model are shown in Table 12.

Development set	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Dev13	0.060	0.014	0.023
Dev14C	0.718	0.162	0.264
Dev14B	0.549	0.072	0.128

Table 12: The results of CRF model.

2. *Rule*. The rule based system is carried out on the development sets to solve the fixed collocation errors. The results achieved by the rule based system are shown in Table 13.

Development set	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Dev13	0.111	0.034	0.052
Dev14C	0.583	0.076	0.135
Dev14B	0.766	0.253	0.380

Table 13: The results of the rule based system.

3. *Graph+CRF*. In this setting, the graph model with different β in ω^L is performed on the *CRF* results. For each development set, an optimal β could be found to obtain the optimal performance.

4. *CRF+Graph+Rule_Post*. Based on the results of the *Graph+CRF* model, we add the rule based system. Similarly, the optimal β could be found.

5. *CRF+Rule_Pre+Graph*. Different from the third setting, we firstly utilize the rule based system on the development sets, and then use the graph model with different β in ω^L.

6. *CRF+Rule_Pre+Graph+Rule_Post*. Based on the results of *CRF+Rule_Pre+Graph* model, we add the rule based system at last.

In Table 14, we compare different improved graph models on the development sets, in which we set β as 6 in ω^L. We could find that though the results of the improved graph model on Dev13 are relatively declined, the results both on the Dev14C and Dev14B are improved. The results in Table 14 prove that CRF model and the rule based system are effective to cover the shortage of the graph model.

5.3 Results

In Bake-off 2014, we submit 3 runs, using the *CRF+Rule_Pre+Graph* model and the weight function ω^L, of which the β is set as 0, 6, and 10, respectively. The results on Test14 are listed in Table 15.

Metric	Run1	Run2	Run3
False Positive Rate	0.5951	0.2279	0.1921
Detection Accuracy	0.3117	0.5471	0.5367
Detection Precision	0.2685	0.5856	0.5802
Detection Recall	0.2185	0.322	0.2655
Detection F1-Score	0.2409	0.4156	0.3643
Correction Accuracy	0.2938	0.5377	0.5311
Correction Precision	0.2349	0.5709	0.5696
Correction Recall	0.1827	0.3032	0.2542
Correction F1-Score	0.2055	0.3961	0.3516

Table 15: Official results of Bake-off 2014.

6 Conclusion

In this paper we present an improved graph model to deal with Chinese spell checking problem. The model includes a graph model and two independently-trained models. To begin with, the graph model is utilized to solve generic spell checking problem and SSSP algorithm is adopted as the model implementation. Furthermore, a CRF model and a rule based system are used to cover the shortage of the graph model. The effectiveness of the proposed model is verified on the data released by the SIGHAN Bake-off 2014 shared task and our system gives competitive results according to official evaluation..

References

Farooq Ahmad and Grzegorz Kondrak. 2005. Learning a spelling error model from search query logs. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages pp. 955–962, Vancouver, British Columbia, Canada, October.

Richard G Casey and Eric Lecolinet. 1996. A survey of methods and strategies in character segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):690–706.

Chaohuang Chang. 1995. A new approach for automatic Chinese spelling correction. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, pages pp. 278–283, Seoul, Korea.

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Kuanyu Chen, Hungshin Lee, Chunghan Lee, Hsinmin Wang, and Hsinhsi Chen. 2013. A study of language modeling for Chinese spelling check. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages pp. 79–83, Nagoya, Japan, October.

Hsunwen Chiu, Jiancheng Wu, and Jason S. Chang. 2013. Chinese spelling checker based on statistical machine translation. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages pp. 49–53, Nagoya, Japan, October.

Thomas Emerson. 2005. The second international Chinese word segmentation Bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages pp. 123–133, Jeju Island, Korea.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Proceedings of 9th Annual Conference of the International Speech Communication Association*, pages pp. 1618–1621, Brisbane, Australia.

Jianfeng Gao, Xiaolong Li, Daniel Micol, Chris Quirk, and Xu Sun. 2010. A large scale ranker-based system for search query spelling correction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages pp. 358–366, Beijing, China, August.

Dongxu Han and Baobao Chang. 2013. A maximum entropy approach to Chinese spelling check. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages pp. 74–78, Nagoya, Japan, October.

Yu He and Guohong Fu. 2013. Description of HLJU Chinese spelling checker for SIGHAN Bakeoff 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages pp. 84–87, Nagoya, Japan, October.

Yuming Hsieh, Minghong Bai, and Kehjiann Chen. 2013. Introduction to CKIP Chinese spelling check system for SIGHAN Bakeoff 2013 evaluation. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages pp. 59–63, Nagoya, Japan, October.

Zhongye Jia, Peilu Wang, and Hai Zhao. 2013a. Grammatical error correction as multiclass classification with single model. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages pp. 74–81, Sofia, Bulgaria, August.

Zhongye Jia, Peilu Wang, and Hai Zhao. 2013b. Graph model for Chinese spell checking. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages pp. 88–92, Nagoya, Japan, October.

Chuanjie Lin and Weicheng Chu. 2013. NTOU Chinese spelling check system in SIGHAN Bakeoff 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages pp. 102–107, Nagoya, Japan, October.

Chaolin Liu, Minhua Lai, Yihsuan Chuang, and Chiaying Lee. 2010. Visually and phonologically similar characters in incorrect simplified Chinese words. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages pp. 739–747, Beijing, China, August.

Xiaodong Liu, Kevin Cheng, Yanyan Luo, Kevin Duh, and Yuji Matsumoto. 2013. A hybrid Chinese spelling correction using language model and statistical machine translation with reranking. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages pp.54–58, Nagoya, Japan, October.

Xu Sun, Jianfeng Gao, Daniel Micol, and Chris Quirk. 2010. Learning phrase-based spelling error models from clickthrough data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages pp. 266–274, Uppsala, Sweden, July.

Peilu Wang, Zhongye Jia, and Hai Zhao. 2014. Grammatical error detection and correction using a single maximum entropy model. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages pp. 74–82, Baltimore, Maryland, June.

Shihhung Wu, Chaolin Liu, and Lunghao Lee. 2013. Chinese spelling check evaluation at SIGHAN Bakeoff 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages pp. 35–42, Nagoya, Japan, October.

Shaohua Yang, Hai Zhao, Xiaolin Wang, and Baoliang Lu. 2012. Spell checking for Chinese. In *International Conference on Language Resources and Evaluation*, pages pp. 730–736, Istanbul, Turkey, May.

Tinghao Yang, Yulun Hsieh, Yuhsuan Chen, Michael Tsang, Chengwei Shih, and Wenlian Hsu. 2013. Sinica-IASL Chinese spelling check system at SIGHAN-7. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages pp. 93–96, Nagoya, Japan, October.

Juifeng Yeh, Shengfeng Li, Meirong Wu, Wenyi Chen, and Maochuan Su. 2013. Chinese word spelling correction based on N-gram ranked inverted index list. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages pp. 43–48, Nagoya, Japan, October.

Liangchih Yu, Chaohong Liu, and Chunghsien Wu. 2013. Candidate scoring using web-based measure for Chinese spelling error correction. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages pp. 108–112, Nagoya, Japan, October.

Huaping Zhang, Hongkui Yu, Deyi Xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, pages pp. 184–187, Sapporo,Japan.

Hai Zhao and Chunyu Kit. 2007. Incorporating global information into supervised learning for Chinese word segmentation. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages pp. 66–74, Melbourne, Australia.

Hai Zhao and Chunyu Kit. 2008a. An empirical comparison of goodness measures for unsupervised Chinese word segmentation with a unified framework. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages pp. 9–16, Hyderabad, India.

Hai Zhao and Chunyu Kit. 2008b. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages pp. 106–111, Hyderabad, India.

Hai Zhao and Chunyu Kit. 2011. Integrating unsupervised and supervised word segmentation: The role of goodness measures. *Information Sciences*, 181(1):163–183.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006a. An improved Chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, volume 1082117, pages pp. 162–165, Sydney, Australia, July.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2006b. Effective tag set selection in Chinese word segmentation via conditional random field modeling. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, volume 20, pages pp. 87–94, Wuhan, China.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2010. A unified character-based tagging framework for Chinese word segmentation. *ACM Transactions on Asian Language Information Processing*, 9(2):1–32.

Hai Zhao, Xiaotian Zhang, and Chunyu Kit. 2013. Integrative semantic dependency parsing via efficient large-scale feature selection. *Journal of Artificial Intelligence Research*, 46:203–233.