A Hybrid Model for Chinese Spelling Check

HAI ZHAO and DENG CAI, Shanghai Jiao Tong University
YANG XIN and YUZHU WANG, Huawei Technologies Co. Ltd.
ZHONGYE JIA, Baosteel Research Institute

Spelling check for Chinese has more challenging difficulties than that for other languages. A hybrid model for Chinese spelling check is presented in this article. The hybrid model consists of three components: one graph-based model for generic errors and two independently trained models for specific errors. In the graph model, a directed acyclic graph is generated for each sentence, and the single-source shortest-path algorithm is performed on the graph to detect and correct general spelling errors at the same time. Prior to that, two types of errors over functional words (characters) are first solved by conditional random fields: 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*), and “得” (*so that*, *have to*) (pinyin: *de*). Finally, a rule-based model is exploited to distinguish pronoun usage confusion: “她” (*she*) (pinyin: *ta*), “他” (*he*) (pinyin: *ta*), and some other common collocation errors. The proposed model is evaluated on the standard datasets released by the SIGHAN Bake-off shared tasks, giving state-of-the-art results.

CCS Concepts: • **Computing methodologies → Natural language processing;**

Additional Key Words and Phrases: Chinese spelling check, hybrid model, graph model, conditional random field, rule-based model

ACM Reference Format:
Hai Zhao, Deng Cai, Yang Xin, Yuzhu Wang, and Zhongye Jia. 2017. A hybrid model for Chinese spelling check. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 16, 3, Article 21 (March 2017), 22 pages.
DOI: http://dx.doi.org/10.1145/3047405

1. INTRODUCTION

Q1 As for every written language, spelling check is a task to detect and correct human spelling errors. Given written sentences with spelling errors, the purpose of the task is to return the locations of incorrect words and suggest the correct ones. Compared

H. Zhao and D. Cai contributed equally as co-first authors. Part of this work was done by Y. Xin, Y. Wang, and Z. Jia when they were affiliated with Shanghai Jiao Tong University. The article is extended from our workshop papers published in CIPS-SIGHAN-2014 [Xin et al. 2014] and SIGHAN-2013 [Jia et al. 2013].

This work was partially supported by the Cai Yuanpei Program (CSC 201304490199 and 201304490171), the National Natural Science Foundation of China (61170114, 61672343, and 61272248), the National Basic Research Program of China (2013CB329401), the Major Basic Research Program of Shanghai Science and Technology Committee (15JC1400103), the Art and Science Interdisciplinary Funds of Shanghai Jiao Tong University (14JCRZ04), and the Key Project of the National Society Science Foundation of China (15-ZDA041).

Q2 Authors' addresses: H. Zhao and D. Cai (corresponding author), (1) Department of Computer Science and Engineering, Shanghai Jiao Tong University, and (2) Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China; emails: thisisjcykcd@gmail.com, zhaohai@cs.sjtu.edu.cn; Y. Xin and Y. Wang, Huawei Technologies Co. Ltd., 2222 Xinjinqiao Road, Shanghai 201206, China; emails: xuechen.xy@gmail.com, wangyuzhu0830@163.com; Z. Jia, Baosteel Research Institute, 655 Fujin Road, Shanghai 201900, China; email: jia.zhongye@gmail.com.

DOI: http://dx.doi.org/10.1145/3047405

Table I. Two Example Sentences for Chinese Spelling Error, One for Each Category

	Example 1 (Nonword Error)	Example 2 (Word Error)
Golden	傳統/美德	好好/地/出去/玩
Misspelled	穿/統/美德	好好/的/出去/玩
Pinyin	*chuan tong mei de*	*hao hao de chu qu wan*
Translation	*traditional virtues*	*enjoy yourself outside*

Note: The sentences in Chinese of each column have the same pinyin.

to English or other alphabetical languages, Chinese has many distinct characteristics [Zhang et al. 2012; Zhang and Zhao 2011; Ma et al. 2010; Li et al. 2009; Zhao 2009]. Chinese spelling check (CSC) is therefore quite different and more challenging in the following ways.

On one hand, the object of spelling check in English is word, but "word" is not a clearly defined unit in Chinese [Huang and Zhao 2007], as there is no explicit word delimiter between words. In English, a word consists of Latin letters, whereas in Chinese, a word consists of characters, which are also known as "漢字" (*character*) (pinyin[1] is *han zi* in Chinese). Thus, essentially, the object of spelling check in Chinese is the characters in a sentence. On the other hand, texts handled by the CSC task are not from handwritten Chinese but from computer-typed Chinese. In handwritten Chinese, due to the characters' own writing complexity as an ideograph, there exist various spelling errors, including noncharacter errors that are caused by misplacing strokes, whereas in computer-typed Chinese, noncharacter spelling errors never occur, namely there is never an "out-of-character" (OOC) problem. Because the Chinese input method engine only allows the legal characters that have been stored in computer to be shown and input [Yang et al. 2012], the characters themselves in Chinese can never be misspelled like in English words. In summary, Chinese spelling errors only come from the misuse of similarly pronounced or written characters, not the writing of characters themselves. For this reason, CSC requires deeper linguistic analysis.

Spelling errors in alphabetical languages, such as English, have two typical categories:

—*Word errors*: The misspelled word is still a legal word, for example, *world* is misspelled as *word*.
—*Nonword errors*: For example, *world* is misspelled as *workd*.

We can distinguish Chinese spelling errors in the similar way, although in each category there exist distinct and more complicated phenomena. In Chinese, if the misspelled word is a nonword, a word segmenter will hardly recognize it as a word but will split its characters into two or more words. For example, if "傳統美德" (*traditional virtues*) in Example 1 of Table I is misspelled as "穿統美德," the word segmenter will segment it into "穿/統/美德" instead of "傳統/美德," as the first two characters cannot form a meaningful word in Chinese. In other words, the word segmenter will almost certainly fail only if a nonword spelling error happens.

Therefore, although word-level information is necessary for spelling check, it is insufficient to perform effective word segmentation before CSC, as the misspelled part cannot be segmented properly by a standard word segmenter, which is supposed to work on a correctly written sentence. As a result, edit distance–based methods for alphabetical languages cannot be directly applied to CSC, which has to deal with the word segmentation problem first. Word segmentation–related errors require information beyond the word level to be handled.

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

Meanwhile, there also exist situations in which the misspelled word is also a legal word. Those spelling errors have little influence in word segmentation. For example, "好好地出去玩" in Example 2 of Table I is misspelled as "好好的出去玩," but both have the same segmentation. Thus, it is necessary to perform a further specific process.

To effectively handle both types of spelling errors in Chinese, we present a hybrid model designed to tackle the CSC task. The hybrid model includes a graph model for generic errors and two independently trained models for specific errors.

As the core of the hybrid model, the graph model is inspired by the idea of the shortest-path word segmentation algorithm. Similar to the shortest-path word segmentation algorithm, a directed acyclic graph (DAG) is built from the input sentence. The spelling error detection and correction problem is then transformed to the single-source shortest-path (SSSP) problem on the DAG. To prevent aggressive corrections, we also adopt filters based on sentence perplexity (PPL) and character mutual information (MI).

The proposed method will be strictly evaluated in datasets released by the latest SIGHAN Bake-off shared tasks.

2. RELATED WORK

In recent years, several methods have been proposed for the CSC task. Generally, most existing works consider adopting two main tools: word segmentation and the language model (LM) for CSC. According to the differences of strategies in use, most approaches fall into four categories.

The first category consists of the methods that all 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 LM generated from a newspaper corpus. The method was based on the observation that all typos were caused by either visual similarity or phonological similarity. Thus, they manually built a confusion set as a key factor in their system. Although the method can detect misspelled words well, some weaknesses needed be improved. For example, it was very time consuming for detection, it generated too many false-positive results, and it was not able to refer to an entire paragraph. Yu et al. [2013] developed a system that did not have a separate error detection. In their system, the correction method itself served as an error detection mechanism. The method assumed that all characters in a 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. However, this method did not always perform well according to the result in Yu et al. [2013].

The second category includes the methods that all single-character words are supposed to be errors, and LM is used for correction. Lin and Chu [2013] developed a system by supposing that all single-character words may be typos. They replaced all single-character words with similar characters using a confusion set and segmented the newly created sentences again. If a new sentence resulted in better word segmentation, a spelling error was reported. Their system performed well in detection recall but not so well in other aspects, especially in the false-alarm rate.

The third category utilizes more than one approach for detection and the LM for correction. Hsieh et al. [2013] used two different systems for error detection. The first system detected error characters according to unknown word detection and LM verification. The second system solved error detection by a suggestion dictionary generated from a confusion set. Finally, the two systems were combined to output the final detection result. In He and Fu [2013], typos were divided into three categories: character-level errors (CLEs), word-level errors (WLEs), and context-level errors (CLEs). They used three different methods to detect the different errors. 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. They generated a dictionary containing similar pronunciation and shape information for each Chinese character, which was used to generate candidate detections. Yang et al. [2013] proposed another method to improve the candidate detections. They employed high-confidence pattern matchers to strengthen the candidate errors after word segmentation.

The last category is formed by the methods that use word segmentation for detection and different strategies for correction [Liu et al. 2013; Chen et al. 2013; Chiu et al. 2013]. Liu et al. [2013] used the support vector machine (SVM) classifier to select the most probable sentence from multiple candidates. They used word segmentation and the machine translation model to generate the candidates. 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 the statistical machine translation model to translate 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. Although there are various attempts for error correction, LM **Q3** is always kept as a simple enough model with relatively good performance, which is also followed to be exploited error correction.

In addition to the preceding four solutions based on word segmentation and LM, there also exist other methods to deal with the CSC task. For example, Sun et al. [2010] developed a phrase-based spelling error model from click-through data by 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] trained a maximum entropy model for each Chinese character based on a large raw corpus and used the model to detect spelling errors in documents.

Our proposed graph model is also related to recent work on Pinyin-to-Chinese conversion [Jia and Zhao 2014], in which the graph construction procedure is similar to ours. However, some quite different modifications aimed at CSC, such as the edge function and the use of a filter, will be explored.

3. SYSTEM OVERVIEW

This article presents a hybrid model for CSC and correction. The model itself is composed of several individual submodels to deal with different types of spelling errors. For nonword errors, a graph word segmentation model is extended to consider additional substitutable characters during the construction of the graph, with the objective to search for a valid composition of a word, which provides a natural way to detect a spelling error from the possible word candidate. In addition, two specific CRF models are trained to deal with the single-character errors caused by two groups of confusing function words, “在, 再” and “的, 地, 得”, respectively, which cannot be discovered by the graph model. Furthermore, for legal word errors, a rule-driven correction system is designed. According to the characteristics of errors, six different categories of rules are defined to detect and correct the errors caused by the misuse of legal words.

The workflow of the whole system is illustrated in Figure 1. In the following sections, we introduce each component in detail.

4. THE GRAPH MODEL

The graph model plays a central role in our entire system, as it handles most spelling errors in reality. Empirical studies have shown that using only an annotated corpus

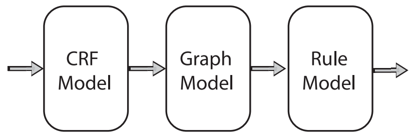


Fig. 1. Workflow of our system.

cannot yield satisfactory performance, as the learning of spelling errors is a serious imbalanced machine learning task with very few but diverse errors spotting in a much larger correct text. This situation drives us to find a better way by integrating useful Chinese language natures. One of our observations is that Chinese spelling errors closely connect to word segmentation because spelling errors may also inevitably cause word segmentation errors. This observation makes sense, as spelling errors may spoil a correct word formation and naturally generate a less likely segmented sentence. Thus, we can roughly summarize that among all possible word segmentations with or without spelling correction, the segmentation with the highest likelihood usually results in the correct sentence (with spelling errors corrected). The proposed graph model is then to solve word segmentation and spelling error checking/correction at the same time through the preceding criterion.

4.1. The Shortest-Path Algorithm for Word Segmentation

Chinese word segmentation has been widely studied [Cai and Zhao 2016; Zhao et al. 2010a, 2013; Zhao and Kit 2008]. The shortest-path word segmentation algorithm is based on the following assumption: a reasonable segmentation should maximize the likelihood of the segmented sentence [Casey and Lecolinet 1996]. In other words, for a character sequence C of m characters $\{c_1, c_2, \dots, c_m\}$, the best segmented sentence $S^* = \{w_1^*, w_2^*, \dots, w_{n^*}^*\}$ should be

$$S^* = \mathop{\arg\max}_{S \in \mathrm{GEN}(C)} \prod_{i=1}^{n} P(w_i | w_1, \dots, w_{i-1}). \tag{1}$$

To keep the preceding optimization problem tractable in practice, a bigram Markov assumption is widely adopted:

$$P(w_i | w_1, \dots, w_{i-1}) = P(w_i | w_{i-1}). \tag{2}$$

Then this optimization problem could be easily transformed into an SSSP problem on a DAG.

A graph $G = (V, E)$ is built to represent the sentence to be segmented. The vertices of G are possible word candidates from the combining of adjacent characters. A dictionary $\mathbb{D}$ is to give all possible legal words. Two special vertices $w_{-,0}$ = “*<START>*” and $w_{n+1,-}$ = “*<END>*” are added to represent two borders of the sentence:

$$V = \{w_{i,j} | w_{i,j} = c_i \dots c_j \in \mathbb{D}\} \cup \{w_{-,0}, w_{n+1,-}\}.$$

The edges are from one word to the next:

$$E = \{<w_{i,j} \rightarrow w_{j+1,k}, \omega>| w_{i,j}, w_{j+1,k} \in V\},$$

where ω is the weight of the edge that should be determined by an LM to indicate the possibility for $w_{j+1,k}$ following $w_{i,j}$ (as in Equation (2)). For example, the Chinese sentence “家書抵萬金” in Table II could be represented by the graph shown in Figure 2.

The graph G is defined as a DAG, and our purpose is to find the optimal segmentation according to Equation (1) that is equal to find the shortest path from “*<START>*” to “*<END>*.”

Table II. Example of Chinese Spelling Error

Golden	家書/抵/萬金
Misspelled	假/書/抵/萬金 or 家/屬地/萬金
Pinyin	*jia shu di wan jin*
Translation	*A letter from home is a priceless treasure*

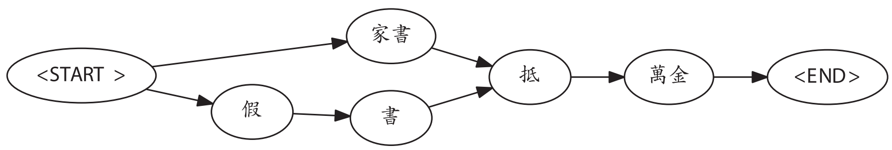


Fig. 2. Sample of a graph for segmentation.

ALGORITHM 1: SSSP Algorithm for Word Segmentation

Input: character sequence S
Input: dictionary $\mathbb{D}$
Output: segmented sentence S^*
Build DAG $G = (V, E)$ from S with $\mathbb{D}$;
Topologically sort G into L;
Init $D[v] \leftarrow -\infty, \forall v \in V$;
Init $B[v] \leftarrow \Phi, \forall v \in V$;
$D[<START>] \leftarrow 0$;
for $u \in L$ **do**
  for v, ω *s.t.* $< u \rightarrow v, \omega > \in E$ **do**
    if $D[v] > D[u] + \omega$ **then**
      $D[v] \leftarrow D[u] + \omega$;
      $B[v] \leftarrow u$;
    end
  end
end
$S^* = \Phi$;
$v \leftarrow <END>$;
while $v \neq \Phi$ **do**
  Insert v into the front of S^*;
  $v \leftarrow B[V]$;
end

The SSSP problem on DAG can be solved by a simple algorithm with time complexity of $O(|V| + |E|)$ [Eppstein 1998], which is shown in Algorithm 1. $B[v]$ denotes the precursor for v along the shortest path from source node to node v and is initialized to a nonexistent node invented for convenience. The segmentation of the preceding example “家書抵萬金” is “家書/抵/萬金” with the SSSP algorithm.

4.2. Integrating Word Segmentation and Spell Checking

The basic idea of using the SSSP algorithm for spelling check stems from the observation that a misspelled word is quite possibly tended to be split into two or more pieces by a word segmenter so that the resulting segmented sentence makes less sense. If those misspelled characters are allowed to be substituted with the correct ones, then the shortest-path word segmenter will choose a segmentation with nodes of word that are spelling corrected. Therefore, we can adopt the SSSP algorithm to solve spelling check

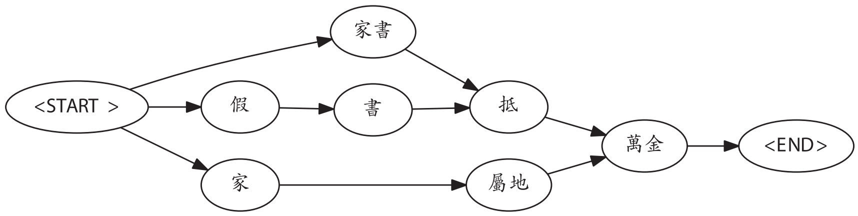


Fig. 3. Sample of a graph for spelling check.

and word segmentation at the same time. For this purpose, a new graph that further takes possible character substitution into consideration is constructed as follows.

First, the vertex set is enlarged by allowing one character substitution in each word. To narrow the search range for substitution candidates, the confusion sets[2] for each character are used as a substitution dictionary $\mathbb{C}$—that is, the substituting character can only come from the confusion set of the original character. The revised vertex set V then is

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

The substitution only happens on those words with lengths between thresholds τ and T.

Second, the edge weights are now determined by both substitution probability and the LM, as in the following equation:

$$\omega = f(\omega_l, \omega_s), \tag{3}$$

where ω_s is a variable that indicates the similarity between the original character and its replacer in a word, and ω_l is the conditional probability derived from the LM (as in Equation (1)). In addition, $f(\cdot, \cdot)$ is a function to score the impact of these two aspects, which will be discussed in detail in Section 6.2.

With the modified DAG G, the SSSP algorithm could perform both word segmentation and spelling check as a joint operation. In this way, the knowledge of a well-trained word segmenter (i.e., the LM) is leveraged, which is essential to resolve some problems requiring sentence-level view.

For example, suppose that the sentence "家書抵萬金" in Table II is misspelled as "假書抵萬金"; the modified graph is shown in Figure 3. The spelling checker may output "家書/抵/萬金" or "家/屬地/萬金," although the latter is not desired.

However, the preceding graph model cannot be applied to continuous character errors. Take the following sentence as an example: "健康" (*health*) (pinyin: *jian kang*) is misspelled as "建缸" (pinyin: *jian gang*) (meaningless character sequence):

—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: *Then I have a plan to let us go swimming in health center near my home.* Q5

[2]In this work, we adopt the confusion sets released in Liu et al. [2011], which collects visually or phonologically similar characters for each individual Chinese character. This dataset is officially provided by SIGHAN Bake-off challenges.

The possible substitutions of “建缸” (pinyin: *jian gang*) may be “缸” (pinyin: *jian gang*), “建鋼” (pinyin: *jian gang*), “建行” (pinyin: *jian hang*), and so on, none of which is the desired correction. Therefore, we have to furthermore revise the construction of the graph model. Considering efficiency and few errors continuously occurring over more than two characters according to our empirical statistics, we only deal with the continuous errors with two characters. The vertex set V now is

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

With the modified G, the incorrect character sequence “建缸” (pinyin: *jian gang*) could be substituted with “健康” (*health*) (pinyin: *jian kang*), “峴港” (*Danang*) (pinyin: *xian gang*), “潛航” (*submerge*) (pinyin: *qian hang*), and so on, and now the desired correction has been successfully included in the candidate set.

5. THE CRF AND RULE-BASED MODELS

Graph model–based word segmentation presented in Section 4 has its limitations. Concretely, the graph model may fail in the following two cases.

First, if a word from the segmentation of a sentence is a single character, the graph model does not work, because substitution is used to turn a meaningless character sequence into words in the vocabulary (dictionary). However, each character has been automatically regarded as a single-character word according to Chinese word segmentation rules. For example, in the following two sentences, “他” (*he*) (pinyin: *ta*) in the first sentence should be corrected to “她” (*she*) (pinyin: *ta*) and “的” (*of*) (pinyin: *de*) in the second sentence should be corrected to “地” (*-ly*, adverb-forming particle) (pinyin: *de*); unfortunately, the graph model does not work for the 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 am outside my motherland and unable to see my mother, I want 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!*

Second, the graph model cannot find the errors that the misused characters have been segmented into a legal word by chance. Take the following sentence as an example. The word “心裡” (*in mind*, *at heart*) (pinyin: *xin li*) will be not separated by any word segmenter, so “裡” (pinyin: *li*) has no chance to be corrected to “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 preceding limitations of the graph model, we adopt a supervised learning approach (CRF) to deal with two kinds of specific errors and a rule-based method to cope with pronoun errors “她” (*she*) (pinyin: *ta*) and “他” (*he*) (pinyin: *ta*), and the fixed collocation errors.

Table III. Feature Template Used in CRF Models

Feature	Example 1	Example 2
w_{-2}	“來”	“和”
w_{-1}	“好好”	“你”
w_0	“地”	“在”
w_0, pos_0	“地,” u	“ 在,” p
w_1	“出”	“一起”
w_{-2}, w_{-1}	“來,” “好好”	“ 和,” “你”
w_{-2}, w_{-1}, w_1	“來,” “好好,” “出”	“ 和,” “你,” “一起”
w_1, w_2	“出,” “去”	“一起,” “ 。”
pos_{-2}	v	p
pos_{-1}	z	r
pos_1	v	s
pos_{-2}, pos_{-1}	v, z	p, r
pos_{-1}, pos_1	z, v	r, s
pos_1, pos_2	v, v	s, w
$pos_{-2}, pos_{-1}, pos_1$	v, z, v	p, r, s
w_{-1}, pos_1	“好好,” v	“你,” s
pos_{-1}, w_1	z, “出”	r, “一起”
pos_{-2}, pos_{-1}, w_1	v, z, “ 出”	p, r, “一起”

Note: Example 1 is the word “地” in the sentence “我們來好好地出去玩吧 !,” and example 2 is the word “在” in the sentence “我只要和你在一起。”.[3]

5.1. The CRF Model

CRFs have been shown to be effective with many natural language processing tasks [Zhao et al. 2006a, 2006b; Zhao and Kit 2007, 2009]. In this work, we utilize two CRF models to respectively tackle two 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*), and “得” (*so that*, *have to*) (pinyin: *de*). The used feature set is presented in Table III with two examples, in which in this table and all following tables, 0 indexes the position of the word currently under consideration, and –2, –1, 1, and 2 index the relative location to the current word position. The CRF models are respectively trained to learn the correct character for two types of character usage confusions.

5.2. The Rule-Based Model

Rule-based models can solve language inference fast and accurately [Zhao et al. 2010b; Shou and Zhao 2012]. To effectively handle pronoun usage errors for “她” (*she*) (pinyin: *ta*), “他” (*he*) (pinyin: *ta*), and other conference or collocation errors. Based on Chinese linguistic knowledge, a series of rules in this work are designed to perform the correction.

Table IV shows the rules for solving pronoun usage errors. Other rules are divided into five categories, which are correspondingly presented in Tables V through IX.[4] In those tables, $prefix_0$ and $suffix_0$ denote the text parts before and after the current word, respectively. The negation symbol “¬” in the tables means that every word in the

[3]For POS tags, *v, p, z, r, s, o, w, u*, and *n* denote *verb, preposition, state word, pronoun, place word, onomatopoeic, punctuation, auxiliary,* and *noun*, respectively. All tables in the rest of this article use this notation.

[4]For simplicity, we only present a part of the rules in Rule 3 in Table VII. The full list for Rule 3 is presented in the appendix.

Table IV. Specific Rules for the Pronouns “ 她、他” Confusion

$prefix_0$ Contains Any of	But None of	w_0	Corrected w_0
爸, 他, 父, 男, 先生	她, 媽, 母, 女, 妹, 姊,姐, 婆, 阿姨, 太太	他	她
她, 媽, 母, 女, 妹, 姊, 姐, 婆, 阿姨, 太太	他, 爸, 父, 男, 哥, 先生	她	他

Table V. Rule 1: Corrections Related to the POS Tag of the Next Word, pos_1

w_0	pos_1	Corrected w_0
阿	w	啊
馬,碼	w	嗎
門	r, n	們
把	r, n	吧

Table VI. Rule 2: Corrections Related to the Suffix After the Current Word, $suffix_0$

w_0	$suffix_0$ Contains Any of	Corrected w_0
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, 船, 高鐵	坐

Table VII. Rule 3: Corrections Related to the Current Word’s Previous and Next Words, w_{-1} and w_1

w_{-1}	w_0	w_1	Corrected w_0
知	到	–	道
¬(內, 肝, 腎)	臟	–	髒
–	總	於	終
–	俄	¬(羅)	餓
改	以	改	一
¬(很)	多	很	都
心	理	¬(學, 研)	裡

Table VIII. Rule 4: Corrections Related to the Current Word’s Neighboring Words, w_{-2}, w_{-1}, w_1, and w_2

w_{-2}	w_{-1}	w_0	w_1	w_2	Corrected w_0
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 IX. Rule 5: Two Words Are Simultaneously Corrected

w_{-1}	w_0	w_1	w_2	w_3	Corrected w_0 and w_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 X. Dataset Statistics Information

Name			Data Size (Lines)	Character Number (K)
Training Set	SIGHAN Bake-off 2013		700	29
	SIGHAN Bake-off 2014	A	342	16
		B	3,004	149
Test Set	SIGHAN Bake-off 2013		2,000	142
	SIGHAN Bake-off 2014		1,062	53

followed brackets does not show in the corresponding position. For example, “他” in the sentence “媽媽他想我” will be corrected to “她” according to the second row of Table IV.

6. EXPERIMENTS

6.1. Datasets and Resources

The proposed models are evaluated on the benchmark datasets of SIGHAN Bake-off shared tasks 2013 and 2014. For SIGHAN Bake-off 2013, sentences were collected from 13- to 14-year-old students’ essays from formal written tests [Wu et al. 2013]. Q8 The training instances are split into two subsets according to the error types. For SIGHAN Bake-off 2014, sentences were collected from Chinese as a foreign language (CFL) learners’ essays selected from the National Taiwan Normal University (NTNU) learner corpus.[5] Both of them are in traditional Chinese. For convenience, the training and test sets of SIGHAN Bake-off 2013 are named TRAIN13 and TEST13, respectively. Q9 The two subsets of the training set of SIGHAN Bake-off 2014 are named TRAIN14A and TRAIN14B, respectively, and the test set is denoted as TEST14. The basic statistics information of both datasets are shown in Table X. A detailed description of the three training sets are summarized as follows:

—TRAIN13: Misused characters can consist of a word with its adjacent character or word, such as “健康” (*health*) (pinyin: *jian kang*) misused by “建康” (meaningless character sequence) (pinyin: *jian kang*).
—TRAIN14A. A misused character is a word itself, such as “在” (*at*) (pinyin: *zai*) misused as “再” (*again*, *more*, *then*) (pinyin: *zai*) and “她” (*she*) (pinyin: *ta*) misused as “他” (*he*) (pinyin: *ta*).
—TRAIN14B. In addition to the preceding two error types, TRAIN14B also includes an error type that two continuous characters are misused, such as “精彩” (*wonderful*)

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

(pinyin: *jing cai*) misused as “警睬” (meaningless character sequence) (pinyin: *jing cai*).

For system details, the dictionary $\mathbb{D}$ used in the SSSP algorithm is SogouW[6] from Sogou Inc. As the original dictionary is in simplified Chinese. The OpenCC[7] converter is then used to convert it to traditional Chinese. A similar character set $\mathbb{C}$ used to substitute characters when constructing the graph in Section 4.2 is provided by Liu et al. [2010]. A bigram LM is built on the Academia Sinica corpus [Emerson 2005] with the IRSTLM toolkit with improved Kneser-Ney smoothing [Chen and Goodman 1999; Federico et al. 2008; Yang et al. 2012]. For Chinese word segmentation, ICTCLAS 2011[8] is exploited.

6.2. Tuning the Graph Model

The hyperparameter settings of machine learning models have a significant impact on performance. As mentioned in Section 6.3.1, in the proposed graph model, a two-variable edge weight function ($f(\cdot, \cdot)$ in Equation (3)) is expected to score the impact of both character similarity and language coherence. To this end, a series of experiments were carried out to select a proper edge weight function. Furthermore, the graph model is prone to turn less frequent words into more frequent words due to the nature of the LM, regardless of the correctness of words with lower frequency. To prevent these kinds of superfluous error corrections, we propose to only correct the most possible errors by setting suitable filters over candidates. Specifically, two types of error filters are designed and examined. The experiments were all conducted on the SIGHAN Bake-off 2013 dataset.

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

—*Correction precision*:

$$\mathcal{P} = \frac{\text{number of correctly corrected characters}}{\text{number of all corrected characters}} \tag{4}$$

—*Correction recall*:

$$\mathcal{R} = \frac{\text{number of correctly corrected characters}}{\text{number of wrong characters of gold data}} \tag{5}$$

—*F1 macro*:

$$\mathcal{F} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P} + \mathcal{R}}. \tag{6}$$

Edge weight function. Multiplication of character similarity and logarithmic conditional probability in the LM is first used as a weight function:

$$\omega^M = -\omega_s \log \omega_l, \tag{7}$$

where ω_s for different kinds of characters are shown in Table XI. The numbers are heuristically determined according to Yang et al. [2012]. The word length threshold is empirically set to $\tau = 2$ and $T = 5$.

Experiments show that with the multiplication function of Equation (7), the graph model gives moderate performance at $\mathcal{P} = 0.49$, $\mathcal{R} = 0.61$, and $\mathcal{F} = 0.55$ on Train13.

[6]http://www.sogou.com/labs/dl/w.html.
[7]http://code.google.com/p/opencc/.
[8]http://www.ictclas.org/ictclas_download.aspx.

Table XI. ω_s Used in ω^M and ω^L

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

Note: These numbers are heuristically set.

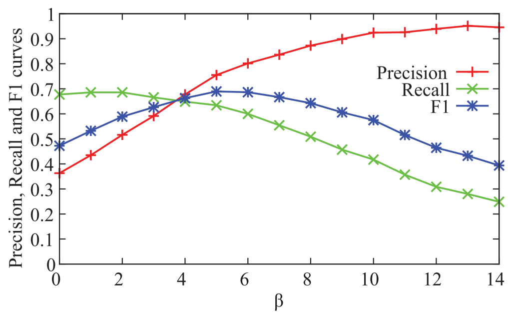


Fig. 4. $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}$ achieved by the graph model with different β on Train13.

A linear combination of character similarity and logarithmic conditional probability in the LM is then tried:

$$\omega^L = \omega_s - \beta \log P, \tag{8}$$

where ω_s for different kinds of characters are shown in Table XI.

We did experiments with Equation (8) and observed that with larger β, the spelling checker tends to perform more cautiously, which results in higher $\mathcal{P}$ but lower $\mathcal{R}$. The $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}$ on Train13 with different β are shown in Figure 4. As we can see, the highest F1 score ($\mathcal{F} = 0.68$) is achieved by setting $\beta = 5$, which is much better than the result ($\mathcal{F} = 0.55$) using Equation (7).

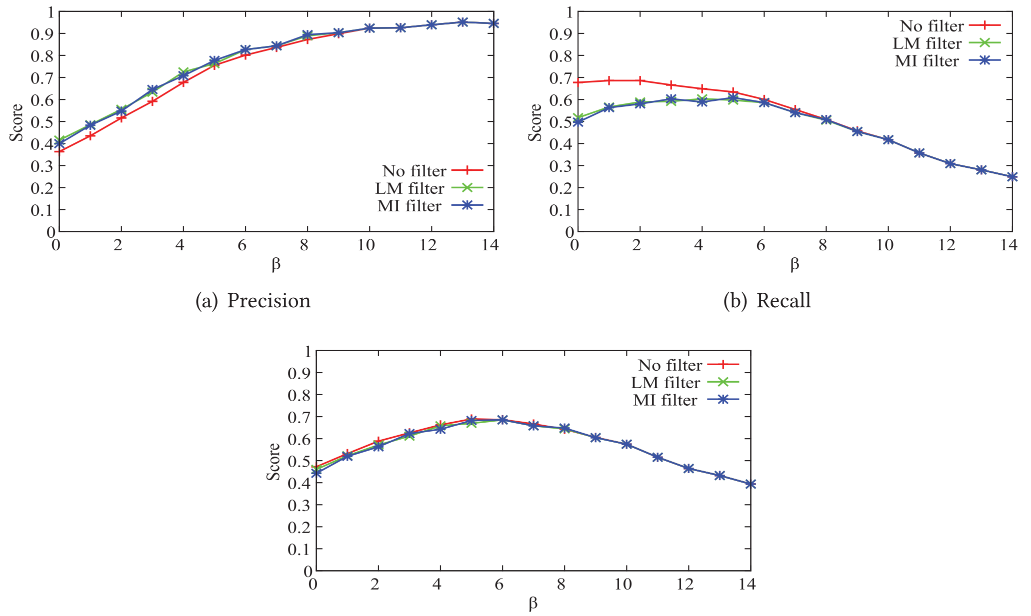
Filters. According to construction of the graph, our graph model tends to output a word sequence with higher sentence likelihood, which may turn less frequent yet correct words into more frequent words. In addition, there is at most one error in each sentence from the SIGHAN Bake-off 2013 dataset. This prior knowledge has been widely used to enhance model performance. (However, this cannot be exploited for the SIGHAN Bake-off 2014 dataset, in which more than one error might emerge in one sentence, e.g., continuous errors "建缸" (pinyin: *jian gang*).)

As the spelling checker might detect multiple errors, a filter according to PPL or MI is used to choose the most likely correction.

For the LM filter, sentence PPL is used as the metric. The correction is chosen according to the lowest PPL.

MI indicates the possibility of two characters being collocated together. For two adjacent characters c_1 and c_2, their MI score is

$$\mathrm{MI}(c_1, c_2) = \log \frac{P(c_1)P(c_2)}{P(c_1 c_2)}. \tag{9}$$



(a) Precision (b) Recall

(c) F1

Fig. 5. $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}$ by the graph model with filters on TRAIN13.

Table XII. Performance Using Single Models on TEST14

Model	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Graph	.4638	.2440	.3197
CRF	.6706	.0724	.1317
Rule	.4782	.1537	.2327

The correction is determined according to the highest MI gain Δ_{MI}:

$$\begin{aligned}\Delta_{\mathrm{MI}} = \max(&\mathrm{MI}(c_{i-1}, c_i') - \mathrm{MI}(c_{i-1}, c_i),\\ &\mathrm{MI}(c_i', c_{i+1}) - \mathrm{MI}(c_i, c_{i+1})).\end{aligned} \tag{10}$$

LM and MI filters slightly enhance the spelling checker. The results of applying two filters are shown in Figure 5. The MI filter is slightly better than the LM filter.

According to the empirical results of the proposed graph model on the SIGHAN Bake-off 2013 dataset, we decided to use Equation (8) as our edge weight function when constructing the graph, of which β is set to 5 for the MI filter. All later experiments follow this setting.

6.3. Performance Analysis

To reveal the individual effectiveness of each component in our hybrid model and how well they work with each other, we first tested each component separately. These results using single models are shown in Table XII. Note that all kinds of spelling errors are considered in this table. However, as each component is designed to deal with different spelling errors, it is desirable to investigate model performance according to the model's own aimed specific error types.

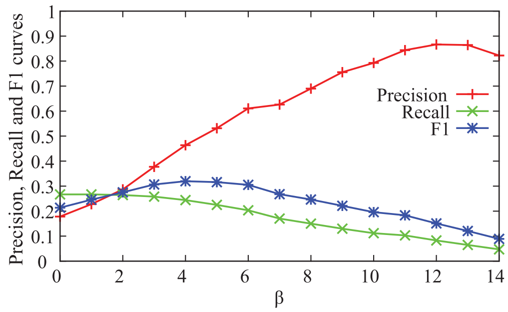


Fig. 6. Performance of graph model on TEST14.

Q12

Table XIII. Statistics for the Two CRFs on the Training Data

Training Set	Golden Label	Label	Number	Percentage (%)
的, 地, 得	的	的	8,070	96.86
		地	184	2.21
		得	78	9.36
	地	的	131	42.39
		地	171	55.34
		得	7	2.27
	得	的	78	19.21
		地	37	9.12
		得	291	71.67
在, 再	在	在	1,438	97.43
		再	38	2.57
	再	在	137	65.87
		再	71	34.13

Table XIV. Performance of the Two CRF Models on TEST14

Error Type	$\mathcal{P}$	$\mathcal{F}$	$\mathcal{F}$
的, 地, 得	.6622	.5765	.6164
在, 再	.6154	.5714	.5936

6.3.1. The Graph Model. We evaluated the graph model for continuous word errors, as it is specialized in our hybrid model to attack this type of error. Results on TEST14 with different β in ω^L are shown in Figure 6, in which the best F1 score ($\mathcal{F} = 0.32$) is achieved by setting $\beta = 4$.

6.3.2. The CRF Models. The training set for the CRF models is collected from TRAIN13, TRAIN14A, and TRAIN14B. All sentences containing the concerned words in these three datasets are used. Table XIII gives the statistics of the obtained training set. With first-order linear chain CRF, we trained two models and tested them on TEST14. The results are shown in Table XIV. Note that we only considered two specific confusions: “在, 再” and “的, 地, 得.”

6.3.3. The Rule-Based Model. The graph model cannot tackle specific word errors effectively (as discussed in Section 6.3.1). The CRF models only deal with two special types of errors. For other errors, we manually extract the rules in Section 5.2 from TRAIN13, TRAIN14A, and TRAIN14B. Results from the rule-based model on TEST14 are shown in Table XII.

Table XV. Official Results of Subtask 1 on SIGHAN Bake-off 2013

Submission	FAR	DA	DP	DR	DF1	ELA	ELP	ELR	ELF1
HLJU-Run2	.6529	.5290	.3849	.9533	.5484	.3390	.1292	.3200	.1841
KUAS & NTNU-Run1	.2257	.7890	.6099	.8233	.7007	.6940	.3753	.5067	.4312
NAIST-Run3	.2243	.7770	.5985	.7800	.6773	.6980	.3964	.5167	.4486
NCTU & NTUT-Run2	.8329	.4110	.3352	.9800	.4995	.2570	.1596	.4667	.2379
NCYU-Run3	.0929	.8250	.7451	.6333	.6847	.7480	.4431	.3767	.4072
NTHU-Run3	.0514	**.8610**	.8455	.6567	.7392	**.8200**	.6695	.5200	**.5854**
NTOU-Run1	.9800	.3140	.3043	**1.0000**	.4666	.1090	.0963	.3167	.1477
SinicaCKIP-Run3	.1629	.8420	.6919	.8533	**.7642**	.7710	.5000	**.6167**	.5523
SinicaIASL-Run2	.1857	.7540	.5873	.6167	.6016	.6860	.3714	.3900	.3805
SinicaSLMP & NTU-Run3	.1414	.8360	.7036	.7833	.7413	.7490	.4431	.4933	.4669
SJTU-Run3	**.0229**	.8440	**.9091**	.5333	.6722	.8090	**.7102**	.4167	.5252
YZU & NCKU-Run1	.0500	.7290	.6500	.2167	.3250	.7050	.4100	.1367	.2050

Note: SJTU-Run3 comes from our team.

6.4. Final Results

6.4.1. SIGHAN Bake-off 2013. We first report the final results on the SIGHAN Bake-off 2013 dataset output by our complete system. The 12 metrics used by the SIGHAN Bake-off 2013 shared task are as follows [Wu et al. 2013]:

—*False-alarm rate* (FAR): Number of sentences with false positive errors/number of testing sentences without errors
—*Detection accuracy* (DA): Number of sentences with correctly detected results/number of all testing sentences
—*Detection precision* (DP): Number of sentences with correctly detected results/number of sentences the evaluation system reports to have errors
—*Detection recall* (DR): Number of sentences with correctly detected errors/number of testing sentences with errors
—*Detection F1* (DF1): 2*DP*DR / (DP+DR)
—*Error location accuracy* (ELA): Number of sentences with correct location detection/number of all testing sentences
Q13 —*Error location precision* (ELP): Number of sentences with correct error locations/number of sentences that the evaluation system reports to have errors
—*Error location recall* (ELR): Number of sentences with correct error locations/number of testing sentences with errors
—*Error location F1* (ELF1): 2*ELP*ELR / (ELP+ELR)
—*Location accuracy* (LA): Number of sentences correctly detecting the error location/number of all testing sentences
—*Correction accuracy* (CA): Number of sentences correctly correcting the error/number of all testing sentences
—*Correction precision* (CP): Number of sentences correctly correcting the error/number of sentences that the system returns corrections.

The official results [Wu et al. 2013] are shown in Tables XV and XVI, in which SJTU-Run3 represents the proposed model. The best results of each metric are in bold. As shown in these two tables, our hybrid model on Test13Sub1 and Test13Sub2 achieves four first ranks out of 12 metrics.

6.4.2. SIGHAN Bake-off 2014. For SIGHAN Bake-off 2014, following conventions of this dataset, only $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}$ in Section 6.2 are utilized as metrics to illustrate model performance. As shown in Table XVII (results in the first block are those from SIGHAN

Table XVI. Official Results of Subtask 2 on SIGHAN Bake-off 2013

Submission	LA	CA	CP
HLJU-Run2	.3230	.2770	.3081
KUAS & NTNU-Run1	.4440	.3940	.5058
NAIST-Run2	.2610	.2540	.6530
NCTU & NTUT-Run1	.0700	.0650	.5118
NCYU-Run2	**.6630**	**.6250**	.7030
NTHU-Run2	.4420	.4310	.7020
SinicaCKIP-Run3	.5590	.5160	.6158
SinicaIASL-Run2	.4900	.4480	.4476
SinicaSLMP & NTU-Run1	.5070	.4670	.4670
SJTU-Run3	.3700	.3560	**.7050**
YZU & NCKU-Run1	.1170	.1090	.4658

Note: SJTU-Run3 comes from our team.

Table XVII. Results on TEST14

Model			$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
SIGHAN 2014	BIT [Liu et al. 2014]	Run1	.3206	.1582	.2119
		Run2	.365	.1883	.2484
	CAS [Xiong et al. 2014]	Run1	.676	.3183	**.4328**
		Run2	.6706	.3183	.4317
	NCTU&NTUT [Wang and Liao 2014]	Run1	.6	.0565	.1033
		Run2	.4592	.0847	.1431
	NCYU [Yeh et al. 2014]	Run1	.3899	.1168	.1797
		Run2	.8406	.2185	.3468
		Run3	.8281	.1996	.3217
	NJUPT [Gu et al. 2014]	Run1	.3191	.1827	.2323
		Run2	.1645	.1186	.1379
		Run3	.1416	.0923	.1117
	NTHU [Chiu et al. 2014]	Run1	.56	.1055	.1775
		Run2	.4406	.1186	.1869
		Run3	.2659	.1337	.1779
	NTOU [Chu and Lin 2014]	Run1	.3965	.1695	.2375
		Run2	.1143	.1281	.1208
	SCAU [Huang et al. 2014]	Run1	.4375	.1582	.2324
		Run2	.2083	.1695	.1869
		Run3	.2712	.1864	.221
	SUDA [Yu and Li 2014]	Run1	.3527	.1375	.1978
		Run2	.7119	.0791	.1424
Our system			.5550	.3914	**.4590**

Bake-off 2014 participants [Yu et al. 2014] [9]); our system obtained the highest correction F1 score among all methods. In other words, the proposed hybrid model outperforms previous state-of-the-art methods.

7. CONCLUSION

In this article, we present a hybrid model for CSC. The hybrid model includes a graph model and two independently trained models. To begin with, the graph model is utilized to solve the generic spelling check problem and the SSSP algorithm is adopted as the

[9] Researchers from KUAS, PKU, and SinicaCKIP also participated in SIGHAN Bake-off 2014. However, there is no technical report from them, and therefore their results are not presented here.

model implementation. By adjusting edge weight function, a trade-off could be made between precision and recall. Furthermore, two CRF models and a rule-based model are used to cover the shortage of the graph model for specific errors. The effectiveness of the proposed model is verified on the benchmark data released by the SIGHAN Bake-off shared tasks.

APPENDIX

In this appendix, we present the full list of Rule 3 as mentioned in Section 5.2.

Table XVIII. Rule 3: Corrections Related to the Current Word's Neighbors, w_{-1} and w_1

w_{-1}	w_0	w_1	Corrected w_0	w_{-1}	w_0	w_1	Corrected w_0
–	感	才	剛	–	性	苦	辛
–	新	福	幸	–	防	應	反
–	忒	別	特	–	每	鮮	海
–	學	行	舉	–		子	靴
–	慍	度	温	–	斥	賞	欣
–	有	善	友	–	酉	吧	酒
–	教	順	孝	–	總	於	終
–	覺	定	決	–	間	！	見
–	現	！	見	–	摟	！	嘍
–	滿	久	蠻	–	應	為	因
–	名	星	明	–	性	期	星
–	單	是	但	–		算	總
–	質	料	資	–	住	你	祝
–	幫	護	保	–	因	該	應
–	力	害	厲	–	令	的	冷
–	想	信	相	–	氣	車	汽
–	旅	行	旅	–	談	水	淡
–	器	著	哭	–	朱	母	祖
–	彩	光	採	–	知	到	道
–	渡	到	度	–	瞭	解	了
–	榮	養	營	–	消	夜	宵
–	癡	呆	癡	–	傍	邊	旁
–	陳	熟	成	–	面	甸	緬
知	到	–	道	電	音	–	影
一	式	–	試	周		–	末
那	裏	–	裡	化	裝	–	妝
否	者	–	則	怎	麽	–	麼
興	棼	–	奮	排	對	–	隊
時	擦	–	差	鍛	鍊	–	練
每	可	–	個	企	鳾	–	鵝
氣	分	–	氛	關	系	–	係
了	假	–	解	台	兆	–	北
安	青	–	靜	印	像	–	象
海	漢	–	灘	對	絕	–	決
申	僅	–	請	懷	年	–	念
情	女	–	侶	瘋	侹	–	狂
秘	決	–	訣	見	量	–	諒
裝	璜	–	潢	歌	首	–	手
禮	帽	–	貌	工	做	–	作
演	作	–	奏	演	坐	–	奏
體	別	–	貼	很	像	–	相
很	九	–	久	皮	戴	–	帶
無	倫	–	論	武	候	–	候
玄	莊	–	奘	導	圓	–	演
打	戰	–	仗	東	四	–	西
和	協	–	諧	生	熱	–	日
喜	藋	–	歡	因	位	–	為
照	像	–	相	–	慎	惜	珍

Table XIX. Rule 3 (continued): Corrections Related to the Current Word's Neighbors, w_{-1} and w_1

w_{-1}	w_0	w_1	Corrected w_0
–	附	合	符
–	讓	費	浪
–	位	了	為
–	讓	後	然
–	身	命	生
–	俄	¬(羅)	餓
–	以	, 從	已
–	生	體, 邊	身
–	像	¬(機, 片)	相
–	式	著, 穿, 一	試
–	以	種, 分, 個	一
–	重	我, 來, 小	從
–	方	子, 間, 東	房
–	情	你, 來, 您, 我	請
–	一	前, 後, 外, 內	以
–	懷	事, 掉, 人, 的, 了	壞
–	周	末, 一, 二, 三, 四, 五, 六, 日	週
–	情, 清	愛, 切, 戚	親
改	以	改	一
用	以	用	一
楊	承	琳	丞
來	所	:	說
年	月	飯	夜
感	安	節	恩
周	潔, 傑	倫	杰
表	員, 潭	–	演
時	後, 候	–	候
台	彎, 往	–	灣
當, 雖	讓	–	然
T, 体	血	–	恤
可, 所	一	–	以
洲, 國	筆	–	幣
雖, 當	熱	–	然
學, 人	身	–	生
作, 或, 再	著	–	者
¬(內, 肝, 腎)	臟	–	髒
名, 漢, 些, 千, 的	子	–	字
手, 書, 那, 哪, 房, 夢, 這	理	–	裡
¬(很)	多	很	都
¬(對)	立	面	裡
¬(必)	須	要	需
心	裡	, 研	理
心	理	¬(, 研)	裡
¬(一, 二, 這, 兩, 幾, 草, 壓)	根	¬(部, 本, 據, 源, 基, 治, 除)	跟

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers and editors for their invaluable comments and suggestions to improve this article.

REFERENCES

Farooq Ahmad and Grzegorz Kondrak. 2005. Learning a spelling error model from search query logs. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*. 955–962.

Deng Cai and Hai Zhao. 2016. Neural word segmentation learning for Chinese. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 409–420.

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 the Natural Language Processing Pacific Rim Symposium*. 278–283.

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 7th SIGHAN Workshop on Chinese Language Processing*. 79–83.

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

Hsunwen Chiu, Jiancheng Wu, and Jason S. Chang. 2013. Chinese spelling checker based on statistical machine translation. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*. 49–53.

Hsun-Wen Chiu, Jian-Cheng Wu, and Jason S. Chang. 2014. Chinese spell checking based on noisy channel model. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*. 202–209.

Wei-Cheng Chu and Chuan-Jie Lin. 2014. NTOU Chinese spelling check system in CLP Bake-off 2014. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*. 210–215.

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*. 123–133.

David Eppstein. 1998. Finding the *k* shortest paths. *SIAM Journal on Computing* 28, 2, 652–673.

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

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*. 358–366.

Lei Gu, Yong Wang, and Xitao Liang. 2014. Introduction to NJUPT Chinese spelling check systems in CLP-2014 Bakeoff. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*. 167–172.

Dongxu Han and Baobao Chang. 2013. A maximum entropy approach to Chinese spelling check. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*. 74–78.

Yu He and Guohong Fu. 2013. Description of HLJU Chinese spelling checker for SIGHAN Bakeoff 2013. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*. 84–87.

Yuming Hsieh, Minghong Bai, and Kehjiann Chen. 2013. Introduction to CKIP Chinese spelling check system for SIGHAN Bakeoff 2013 evaluation. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*. 59–63.

Changning Huang and Hai Zhao. 2007. Chinese word segmentation: A decade review. *Journal of Chinese Information Processing* 21, 3, 8–20.

Qiang Huang, Peijie Huang, Xinrui Zhang, Weijian Xie, Kaiduo Hong, Bingzhou Chen, and Lei Huang. 2014. Chinese spelling check system based on tri-gram model. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*. 173–178.

Zhongye Jia, Peilu Wang, and Hai Zhao. 2013. Graph model for Chinese spell checking. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*. 88–92.

Zhongye Jia and Hai Zhao. 2014. A joint graph model for Pinyin-to-Chinese conversion with typo correction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1512–1523.

Junhui Li, Guodong Zhou, Hai Zhao, Qiaoming Zhu, and Peide Qian. 2009. Improving nominal SRL in Chinese language with verbal SRL information and automatic predicate recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. 1280–1288.

Chuanjie Lin and Weicheng Chu. 2013. NTOU Chinese spelling check system in SIGHAN Bake-off 2013. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*. 102–107.

Chao-Lin Liu, Min-Hua Lai, Yi-Hsuan Chuang, and Chia-Ying Lee. 2010. Visually and phonologically similar characters in incorrect simplified Chinese words. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. 739–747.

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.

Min Liu, Ping Jian, and Heyan Huang. 2014. Introduction to BIT Chinese spelling correction system at CLP 2014 Bake-off. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*. 179–185.

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 7th SIGHAN Workshop on Chinese Language Processing*. 54–58.

Xuezhe Ma, Xiaotian Zhang, Hai Zhao, and Bao-Liang Lu. 2010. Dependency parser for Chinese constituent parsing. In *Proceedings of the Joint Conference on Chinese Language Processing*. 1–6.

Heming Shou and Hai Zhao. 2012. Hybrid rule-based algorithm for coreference resolution. In *Proceedings of the Joint Conference on EMNLP and CoNLL-Shared Task*. 118–121.

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*. 266–274.

Yih-Ru Wang and Yuan-Fu Liao. 2014. NCTU and NTUT's entry to CLP-2014 Chinese spelling check evaluation. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*. 216–219.

Shihhung Wu, Chaolin Liu, and Lunghao Lee. 2013. Chinese spelling check evaluation at SIGHAN Bake-off 2013. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*. 35–42.

Yang Xin, Hai Zhao, Yuzhu Wang, and Zhongye Jia. 2014. An improved graph model for Chinese spell checking. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing.* 157–166.

Jinhua Xiong, Qiao Zhang, Jianpeng Hou, Qianbo Wang, Yuanzhuo Wang, and Xueqi Cheng. 2014. Extended HMM and ranking models for Chinese spelling correction. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*. 133–138.

Shaohua Yang, Hai Zhao, Xiaolin Wang, and Baoliang Lu. 2012. Spell checking for Chinese. In *Proceedings of the International Conference on Language Resources and Evaluation*. 730–736.

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 7th SIGHAN Workshop on Chinese Language Processing*. 93–96.

Jui-Feng Yeh, Sheng-Feng Li, Mei-Rong Wu, Wen-Yi Chen, and Mao-Chuan Su. 2013. Chinese word spelling correction based on *N*-gram ranked inverted index list. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*. 43–48.

Jui-Feng Yeh, Yun-Yun Lu, Chen-Hsien Lee, Yu-Hsiang Yu, and Yong-Ting Chen. 2014. Chinese word spelling correction based on rule induction. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*. 139–145.

Junjie Yu and Zhenghua Li. 2014. Chinese spelling error detection and correction based on language model, pronunciation, and shape. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*. 220–223.

Liang-Chih Yu, Chao-Hong Liu, and Chung-Hsien Wu. 2013. Candidate scoring using Web-based measure for Chinese spelling error correction. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*. 108–112.

Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2014. Overview of SIGHAN 2014 Bake-off for Chinese spelling check. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*. 126–132.

Xiaotian Zhang, Chunyang Wu, and Hai Zhao. 2012. Chinese coreference resolution via ordered filtering. In *Proceedings of the Joint Conference on EMNLP and CoNLL—Shared Task*. 95–99.

Xiaotian Zhang and Hai Zhao. 2011. Unsupervised Chinese phrase parsing based on tree pattern mining. In *Proceedings of the 11th China National Conference on Computational Linguistics*.

Hai Zhao. 2009. Character-level dependencies in Chinese: Usefulness and learning. In *Proceedings of the 12th Conference of the European Chapter of the ACL*. 879–887.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006a. An improved Chinese word segmentation system with conditional random field. In *Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing*. 162–165.

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 Pacific Asia Conference on Language, Information and Computation*, Vol. 20. 87–94.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2010a. A unified character-based tagging framework for Chinese word segmentation. *ACM Transactions on Asian Language Information Processing* 9, 2, 5.

Hai Zhao and Chunyu Kit. 2007. Scaling conditional random field with application to Chinese word segmentation. In *Proceedings of the 3rd International Conference on Natural Computation*, Vol. 5. 95–99.

Hai Zhao and Chunyu Kit. 2008. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In *Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing*. 106–111.

Hai Zhao and Chunyu Kit. 2009. A simple and efficient model pruning method for conditional random fields. In *Proceedings of the International Conference on Computer Processing of Oriental Languages*. 145–155.

Hai Zhao, Yan Song, and Chunyu Kit. 2010b. How large a corpus do we need: Statistical method versus rule-based method. In *Proceedings of the 7th Conference on International Language Resources and Evaluation.* 1672–1677.

Hai Zhao, Masao Utiyama, Eiichiro Sumita, and Bao-Liang Lu. 2013. An empirical study on word segmentation for Chinese machine translation. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*. 248–263.

Received July 2016; revised November 2016; accepted January 2017

QUERIES

Q1: AU: Please review this article very carefully.
Q2: AU: Please provide full mailing and email addresses for all authors.
Q3: AU: Please review phrasing of sentence beginning with “Although there are various attempts,” paying particular attention to “which is also followed . . .” Please rephrase for clarity.
Q4: AU: Please review phrasing: “diverse errors spotting” and “useful Chinese language natures.”
Q5: AU: Please review single quotes used in list entries throughout this article.
Q6: AU: Please confirm “two common character confusions,” as the commas are separating four entries.
Q7: AU: Please rephrase: “To effectively handle . . .” is not a complete sentence.
Q8: AU: Please be sure that foonotes are shown sequentially in the article (considering that footnote 4 is shown in the Table note for Table III).
Q9: AU: Please review the random “4” on the left side of Table III.
Q10: AU: Please rephrase to make a complete sentence with “As the original dictionary . . .”
Q11: AU: Please review URL in footnote 8. Unable to connect.
Q12: AU: Please confirm closing up commas with thousands in Table XIII.
Q13: AU: Please note that RLP and RLR have been changed to ELP and ELR to match table column heads.