Spelling checking using conditional random fields with feature induction for secondary language learners

Jui-Feng Yeh, Kai-Hsiang Yu and Chan-Kun Yeh

Department of Computer Science & Information Engineering, National Chia-Yi University, Chiayi, Taiwan, ROC

ABSTRACT

This paper presents a framework for Chinese spelling error detection and correction using conditional random fields (CRFs) with feature induction for secondary language learners. The trend of learning Chinese as second language is increasing recently. CRFs are adopted here as the model that models the global and local information to judge if the word is correct or not. Local features are usually considered to make the decision for intelligent systems. Herein, CRFs are one of the most used statistical approaches those which can adjust the corresponding weights for features to achieve near optimal results. This paper invested an automatic rule induction method to capture the hidden features for spellcheck in Chinese. Considering position information, the features are inducted by counting in the training corpus automatically. Therefore, the CRFs integrate the features and achieve approximated optimum by adjusting the weights corresponding to related features. From the experimental results, we can find the proposed method outperforms the traditional approaches and obtains improvement for finding the misspelled words and correcting them. It should be concluded, from what has been illustrated above that the proposed method is near to practice and useful for the learners who take Chinese as a secondary language.

ARTICLE HISTORY

Received 4 February 2016
Accepted 14 March 2016

KEYWORDS

Spelling check; conditional random fields; feature induction; secondary language

1. Introduction

Spelling checker is one of the most essential tools for keystroke especial for natural language processing. Recently, the trend of big data causes individuals to extract the intelligence from the data/open data from internet. Therefore, search engine based on keyword query is invested by research efforts in the latest decades. In spite of the keyword-based query search engine has achieved a good performance for users. However, semantic component usually is lacking in this style. As we have known, natural language is the most convenient interface as the human–machine interaction. However, it is not easy for individuals to input the natural language sentence especially for Asian languages such as Chinese. Furthermore, even in some editor tools such as Microsoft office, spelling error detection and correction have been considered as one of the most useful functions for English. However, the Chinese spelling checker is absent in nowadays applications.

This paper aims at developing the spelling checker for the second language learners of Chinese. Many foreigners are learning Chinese as secondary language in the recent decades due to internationalization. Additionally, they learn Chinese as one-fifth of population speaks Chinese on the planet and the population is increasing. That is to say, over one billion people speak and write Chinese in their daily life. However, it is hard for the native alphabetical speaker to learn Chinese as their secondary language. A Chinese character is composed of three main components; shape, syllable, and meaning. For illustration, we consider the versatility of Chinese character in these three aspects, also the number of characters is significantly larger than alphabetical languages such as English. Consequently, the first problem for learners of Chinese as a secondary language is the usage of Chinese character. That is to say, they must prevent the spelling errors when they learn and use Chinese. The number of characters in Chinese word is more than one thousand. Chinese words usually consist of two, three, and four characters. Since there is no space between words in Chinese, the word segmentation plays an essential role for Chinese natural language processing. Besides, different/similar pronounces and glyphs of the Chinese characters cause the solution for spellchecking is hard to achieve near to perfect. Different from English, there are more than ten thousand words in Chinese, different combinations have different meanings. Sometimes, there are some misunderstandings because of using the
wrong word as they have the same pronunciation. Even if one word is mistaken it might lead the sentence to have an entirely different meaning. So, we must carefully choose the word. Pronunciation and glyphs are instinctive ways to learn Chinese. If we can speak, read or write Chinese then we can quickly learn Chinese. And the most important part is what the Chinese word means. For different meanings and parts of speech (POS), the Chinese word is rich and hard to understand its usage for secondary language learners. In English, there are eight parts of speech (nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions, and interjections). In Chinese, there are ten categories in POS; they are nouns, adjectives, verbs, adverbs, pronouns, interjections, prepositions, conjunctions, auxiliary words, and quantifiers. For obtaining a detailed illustration of the POS-related features, we describe the foundational definition and their attributes about POS in Chinese herein. Nouns indicate the names of people or things, they can be further divided into four sub sorts, proper nouns, common nouns, abstract nouns, time nouns, and place nouns. The parts of speech, adjective shows the quality or forms of people or things, or the state of action or behavior. Verbs indicate behaviors, action or changes of people or things. They have several subsidiary categories: model verbs, tendency verb and deciding verb. Adverb is used in front of verbs or adjectives to show degree, extent, time or negation. Pronoun is used to replace nouns or numerals. Preposition introduces nouns, pronouns and other linguistic units to verbs or adjectives and shows the relationship between time, space, objects or methods. Conjunction connects words, phrases or sentences.

The remainder of this paper is organized as follows. Related works are illustrated in Section 2. Section 3 details the proposed conditional random fields (CRFs) with automatic rule induction for spelling error detection. For assessing the development of the proposed approach system, the experiments conducted are reported in Section 4, and concluding remarks and suggestions for future research are provided in Section 5.

2. Related works

In recent years, there are many researches about Chinese word spelling correction.[1–4] Herein, we divided the related works into three categories, attribute, method and applications, according to their main contributions. A joint graph model was used to pinyin-to-Chinese conversion with typo correction by.[5] Zampieri and de Amorim [6] combined phonetics and clustering algorithms for improving the word recovery. Many research efforts were invested for practical applications. Zhuang et al. used a statistical language model for spelling error checking for optical character recognition (OCR).[7] For OCR, multi-knowledge resources were used in.[8] Chen et al. adopted N-gram-based approach to detect and correct the typos in user's queries for web search engine.[9] For Internet chatting systems, a detection and correction scheme based on statistic and pinyin similarity was used in.[10] Bangqun designed and implemented the spelling check function of Online Public Access Catalogue (OPAC).[11]

For secondary language learning, the web site (http://www.pigai.org/) provided the necessary spelling error checking functionality in English. Hao et al. combined weighted finite state automata using incremental latent semantic analysis for automated Chinese essay scoring systems.[12] Rimrott and Heift aimed at spelling checker for second language learners, they classed the language influence into two categories: inter-lingual and intra-lingual. Besides, they also divided the factors resulting in spelling errors in linguistic subsystem into lexical, morphological, phonologic, and orthographic.[13] An error models was adopted here for improving the detection of improperly used Chinese characters in students’ essays. [14] For Chinese as foreign language, Wang and Wu [15] reported the observations about the conversation repair by learners in class. Chen et al. invested a probabilistic framework for Chinese spelling check.[16]

Jiang et al. proposed Chinese sentence error including two parts, spelling error and grammar error.[17] And there are two approaches to detecting errors: rule based and statistics based. They provide a complete and practical rule based combined grammar character with instantiation detecting grammar errors system which is written by XML. Xiong et al. use an HMM-based approach to segment sentences and generate candidates for sentences with spelling correction. They also use search engine results to help in decision-making for candidates.[18] Zheng et al. use sentiment object recognition to identify objects from the opinion sentence. Association mining and rules are established to deal with the reviews.[19] They change the rules. Lee et al. manually constructed a set of linguistic rules with syntactic information to detect erroneous sentences that were written by the second language learners.[20] If a sentence satisfies a syntactic rule, they will regard the input sentence as erroneous and respond with suggestions to indicate the possible errors. Tseng et al. proposed an approach to sieve out the candidate new terms from a large and long lasting stream of news.[21] It is based on the previously developed techniques for keyword extraction, hot topic detection, and term association analysis. Tseng et al. used a method based on the combination of trigger words, dictionary and rules to realize the personal attributes extraction is introduced.[22] They also built a basic framework including dictionary, trigger words, and rules that are relative to the task to extract personal specific
attributes. Xiang et al. integrated intent and sentiment in an online manner, which automatically identifies the two factors and gives the managers real-time feedbacks. [23] Considering the efficiency in search space, Yeh et al. proposed an approach based on inverted index list with a rescoring mechanism which is composed of detection and correction feature of those integrated in the maximum entropy framework. [24] Yeh and Yeh [25] used the rule induction for classifying the usage of ‘De’ in Chinese. Lin and Chu [26] studied Chinese spelling check using confusion sets and N-gram statistics. Additionally, Liu et al. considered the larger number of the homonymous or homomorphic characters in Chinese, they presented a hybrid ranking approach. [27] Hsieh et al. proposed a framework to correct Chinese spelling errors with word lattice decoding. [28]

In this paper, we proposed five steps to find the wrong words in a document. First, we input the testing sentence then we use the CKIP to segment the word, the third step is finding the wrong word. The third step is the main method in this paper. Three spelling error categories caused by phonetic similarity, semantic-similarity and graphemic similarity are illustrated in Table 1. And then we remove the duplicate wrong word. Finally output the error word file. Single words mean there is no match word before or after this single word, in other words, it might be an error word, so we compose the word which is before or after this single word. After composing two single words, it can generate a new word and regard it as a suspicious error word. Idioms usually are composed of four words, so we take the four words to pronounce and glyph to compare with the E-Hownet.

### 3. The proposed CRFs with feature induction

Since N-gram models can achieve an acceptable performance for Chinese spelling checker, the local information is useful and fully utilized. However, global information should be necessary to obtain a better performance. Additionally, for the Chinese as the secondary language learners, to extract the differences between their first/source language and secondary/target language is very essential. It is not capable of to alter the spelling errors that usually result from the influence of their mother tongue. Besides, the first difficulty they face is the number of character being too large for Chinese learners. Instead of 26 characters as in English, tens of thousands characters are used in Chinese especially for traditional Chinese. Furthermore, the complexity and the strokes of a Chinese character are very hard to learn for second language learners. Several observations in the last few aspects have shown how to capture the reasons that cause the Chinese spelling error and it is important for achieving an improvement. Herein, CRFs are used as the frameworks for a statistical model that is able to capture the global and local information and obtain a better performance. A feature induction algorithm is adopted here to describe the difference between first/source language and secondary/target language. This paper illustrates the CRFs and features induction in Sections 3.1 and 3.2 separately.

#### 3.1. CRFs for Chinese spelling checker

In natural language processing, word is the basic unit for semantic interpretation. Several methods are based on word-based approaches such as N-gram models and parser. Due to the limitations about corpus size, concept- and class-based approaches are used to capture word level information. However, spelling error checking is defined as a word-level or character-level event. Since the sentence first is segmented into a word sequence in Chinese language, the typos usually turn the original word containing spelling errors into some words with fewer characters. To sum up, for capturing the information either in word-level or character-level, we need a framework to integrate these features and obtain near to optimal performance. To deal with the spellcheck problems, we must integrate different features especially in Chinese language processing. This paper invested a CRF-based approach to integrate the different features. CRFs are a class of statistical modeling method often applied in automatic annotation by machine learning in natural language processing, where they are used for structured prediction considering not only local information but also global characteristic. Therefore, whereas a traditional ordinary method predicts a label for a single sample without regard to other samples, a CRF can take context into account. Herein, this paper proposed a CRF-based approach in natural language processing for predicting possible spelling errors and corrects them by classing sequences of labels for sequences of input characters in the user's input sentences. Similar to naive Bayes classifier, CRF can be further described as follows under the assumption of conditional independence:

$$P(y, x) = P(y) \prod_{k=1}^{K} P(x_k | y), \quad (1)$$

where $y$ means the output or annotation and $x$ denotes the observations. As we have known, the CRFs for natural language processing usually have the logistic regression

| Error type                  | Misspelled          | Corrections          |
|-----------------------------|---------------------|----------------------|
| Phonetic-similar error      | 粽藝天王            | 粽藝天王            |
| Semantic-similar error      | 我現在上課          | 我正在上課          |
| Graphemic-similar error     | 令天天氣情         | 今天天天氣情         |

Table 1. Examples of spelling error.
3.2. Feature induction by aligned corpus

Spelling error detection and correction is defined as the character selection problems. Word is the basic unit of semantic representation. Due to data scarcity, POS-based features are also considered here for gathering the related information. It is linked with something previously mentioned that building block is constructed by character and word level information as shown in Figure 2. The observations and candidate characters expended by the phonologically and visually similar character sets are included in character level features. Word and corresponding POS are also considered to form the word level features. Instead of general chain CRFs, the sequential model/feature is considered here for describing the character/word transitions. LEM-2 algorithm [29] is adopted here for obtaining useful features. The process flow of the feature induction is similar to that illustrated by Yeh and Yeh[25]. Compared to the algorithm used in [25], the main contribution of this forms. Therefore, the proposed CRFs adopted here can be expressed by the conditional probability of labels as:

\[
P(x|y) = \frac{1}{Z(x)} \exp \left\{ \theta_y + \sum_{j=1}^{K} \theta_{y,j} x_j \right\},
\]  

(2)

The normalizing factor \( Z(x) \) is defined by Equation (3).

\[
Z(x) = \sum_{y} \exp \left\{ \theta_y + \sum_{j=1}^{K} \theta_{y,j} x_j \right\},
\]  

(3)

By introducing the feature function the conditional probability and normalizing factor \( Z(x) \) are illustrated as the following equations:

\[
P(x|y) = \frac{1}{Z(x)} \exp \left\{ \sum_{k=1}^{K} \theta_{y,k} f_k(y, x) \right\},
\]  

(4)

\[
Z(x) = \sum_{y} \exp \left\{ \sum_{k=1}^{K} \theta_{y,k} f_k(y, x) \right\},
\]  

(5)

In this paper, observation means the characters in the original input text. Since the problem is defined as some typos that are embedded in the original sentence, the character candidates are expanded according to confusion sets including the phonologically and visually similar character sets. That is to say, the potential character candidates are selected to substitute for the spelling errors in the original sentence. Furthermore, the words are built using characters from bottom to top as shown in Figure 1.
4. Experimental results

Due to the use of experiments to verify the proposed approach is practical or not, a Chinese spelling checker has been developed here for performance evaluation. For evaluating the proposed approach, a CRF-based Chinese spelling checker with automatic feature induction is developed here. Due to the proposed approach is a model-based method, the procedure can be further distinguished into two parts: training phase and testing phase. For natural language processing, some tools are necessary especially in Chinese. Since no blank between Chinese words in sentence, the word segmentation is necessary for Chinese natural language processing. Herein, the tool AutoTag, developed by CKIP in Sinica, is used for word segment. After word segmentation, a Chinese sentence into word sequence with POS tagging.[30]

In training phase, it is to construct the character confusion dictionary and conditional random field models which are used in test phase. The character confusion includes the similar in pronunciation and shape. Besides, E-Hownet is also used as the knowledge base to find the misspelled characters and provide the possible candidates. That is to say, the confusion dictionary and E-Hownet is the main resources for word-building. The LEM-2 algorithm is used to induct the features from the training corpus. Finally, the parameter estimation of CRFs is done by iterative procedure.

There are two levels to experiments our system: detection level and correction level. Detection level is to find out the location of incorrect spelling characters in the sentences. Correction level is further designed to correct the possible errors.

4.1. Data preparation and evaluation metrics

There are two levels to experiments our system: detection level and correction level. Detection level is to find out the location of incorrect spelling characters in the sentences. Correction level is further designed to correct the possible errors.

The training and test corpus provided by SIGHAN 2015 Bake-off for Chinese Spelling Check are annotated in SGML format, an example is shown in Figure 3. However, the data are not rich enough to train the CRFs model. Therefore, the data-set gathered from internet is also integrated as part of the training data. The test corpus for SIGHAN 2015 Bake-off for Chinese Spelling Check is used here as the test set.[31]

Four matrices: accuracy, precision, recall and $F_1$ score are used here to measure the performance of the proposed approach. Precision is the fraction of retrieved documents that are relevant to the query. Recall is the fraction of the documents that are relevant to the query. The following is an essay from the corpus illustrated in SGML format.

```xml
<ESSAY title="寫給即將初次見面的筆友的一封信">
<TEXT>
<PASSAGE id="B1-0112-1">那一天我會穿牛仔褲和紅色的外套；頭會帶著藍色的帽子。如果你找不到我，可以打我的手機。</PASSAGE>
<PASSAGE id="B1-0112-2">我記得你說你想試試看越南菜有什麼味道，午餐我會帶你去吃。我也想試試看那一家的越南菜，網路站說很多人喜歡那一家的餐廳。</PASSAGE>
</TEXT>
</ESSAY>
```

Figure 3. An essay from the corpus illustrated in SGML format.
that are successfully retrieved. F1 score is a measure of a test’s accuracy. These metrics are all based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN). It considers both the precision and the recall of test to compute the score. In confusion matrix, TP is the ratio of the system that determines the character for errors related to the actual error, and the judgments the system is correct. FP means that system determines the character for errors is not related to the actual errors, and the judgments of the system is incorrect. FN denotes system determines the character of errors related to the actual error, and the judgment of the system is incorrect. Finally, TN is the ratio of the system determines the character of errors is not related to the actual error, and the judgment of the system is correct. The precision, recall, and F1 score are defined as the Equations (6)–(8).

\[
\text{Precision rate} = \frac{TP}{TP + FP}, \tag{6}
\]

\[
\text{Recall rate} = \frac{TP}{TP + FN}, \tag{7}
\]

\[
F1 \text{ score} = \frac{2 \times \text{Precision rate} \times \text{Recall rate}}{\text{Precision rate} + \text{Recall rate}}. \tag{8}
\]

4.2. Experimental results and discussion

Tables 2 and 3 show the performances of the proposed approach using the SigHan 2015 data-set and compared to trigrams.

According to the observation about the experimental results, the proposed approach outperforms the trigrams approach either in detection or correction phrases. The main contribution comes from the fact that CRFs can integrate different categories features to achieve near optimal results. Additionally, the aligned parallel corpus provides useful error patterns for secondary learners. By automatic rule induction algorithm LEM-2, we can find many features that are unseen by human tagging and labor consuming. Considering local information, tri-gram is usually able to achieve a good enough performance in natural language processing. By the observation of the results shown in Tables 2 and 3, we can find that the proposed approach outperforms tri-gram significantly in accuracy and precision measures. For recall rate, the performance about the proposed approach is also better than that about tri-gram. Deservedly, the F1 score of the proposed method is higher than tri-gram. These results lead to the conclusion that the proposed approach is operative and practicable for Chinese spelling error detection and correction.

5. Conclusions

In this paper, we present an automatic approach based on CRFs with feature induction for Chinese spelling error detection and correction. Instead of N-gram, the proposed CRFs are with global and local information to achieve better performance. Besides, the rule induction algorithm LEM-2 is adopted here for generating the feature set automatically. Since there are some error patterns resulting from the influence upon the learners’ native language, the LEM-2 algorithm is able to induct the useful features. Finally, the parameters of CRFs are estimated to achieve near optimal conditional. Compared to traditional N-gram approach, CRFs provide a framework for integrating the local and global information to achieve a better performance. Considering error patterns, LEM-2 algorithm is applied to induct the features of those that are used in CRFs and achieve significant improvement especially in accuracy and precision rates. To obtain the contribution, the feature sets concluding three categories, character, word and POS are adopted for describing the Chinese text. Furthermore, considering the error patterns, the word is divided into three classes, the single word, idioms and the other words. An automatic induction algorithm is used to find the helpful features. According to the experimental results, we find the proposed approach outperforms the traditional N-gram-based approach significantly.

Disclosure statement

No potential conflict of interest was reported by the authors.
References

[1] Huang CM, Wu MC, Chang CC. Error detection and correction based on Chinese phonemic alphabet in Chinese text. In: Vicenç Torra, Yasuo Narukawa, Yuji Yoshida, editors. Modeling decisions for artificial intelligence. Berlin Heidelberg: Springer; 2007. p. 463–476.

[2] Lei Z, Ming Z, Changning H, et al. Automatic detection and correction of typed errors in Chinese text. Appl. Linguist. 2001;1:19–26. Available from: http://www.fed.cuhk.edu.hk/en/lang/200100000001/0019.htm

[3] Liu B, Zheng J, Zhang H. Consistency check of segment using combination of rule and statistics. Comput. Eng. Design. 2008;29:1814–1816.

[4] Yang S, Zhao H, Wang X, et al. Spell checking for Chinese. LREC. 2012:730–736. Available from: http://www.lrecconf.org/proceedings/lrec2012/LREC2012_Proceeding.zip

[5] Jia Z, Zhao H. A joint graph model for pinyin-Chinese conversion with typo correction. Proceedings of ACL. 2014;1:1512–1523. Available from http://www.aclweb.org/anthology/P14-1142

[6] Zampieri M, de Amorim RC. Between sound and essays with error model. Int. J. Cont. Eng. Educ Life of improperly used Chinese characters in students’ checkers in CALL. Calico J. 2013;23:17–48.

[7] Huang CM, Wu MC, Chang CC. Error detection and correction based on Chinese phonemic alphabet in Chinese text. In: Vicenç Torra, Yasuo Narukawa, Yuji Yoshida, editors. Modeling decisions for artificial intelligence. Berlin Heidelberg: Springer; 2007. p. 463–476.

[8] Lei Z, Ming Z, Changning H, et al. Automatic detection and correction of typed errors in Chinese text. Appl. Linguist. 2001;1:19–26. Available from: http://www.fed.cuhk.edu.hk/en/lang/200100000001/0019.htm

[9] Liu B, Zheng J, Zhang H. Consistency check of segment using combination of rule and statistics. Comput. Eng. Design. 2008;29:1814–1816.

[10] Han B, Li Z. Detection and correction scheme of internet chat lingo based on statistic and pinyin similarity. In: Chunfeng Liu, Leizhen Wang, Aimin Yang, editors. Advances in natural language processing. Berlin Heidelberg: Springer; 2014. p. 438–449.

[11] Zhuang L, Bao T, Zhu X, et al. A Chinese OCR spelling check approach based on statistical language models. Systems, Man and Cybernetics, 2004 IEEE International Conference on IEEE. 2004;5:4727–4732.

[12] Chen Z, Li Y, Liu H, et al. Chinese spelling correlation in search engines based on N-gram model. J. China Acad. Electron. In-form. Technol. 2009;4:323–326.

[13] Han B, Li Z. Detection and correction scheme of internet chat lingo based on statistic and pinyin similarity. In: Chunfeng Liu, Leizhen Wang, Aimin Yang, editors. Information computing and applications. Springer: Berlin Heidelberg; 2012. p. 169–174.

[14] Banggun L. Design and implementation of the spelling check function of OPAC. Res. Library Sci. 2012;4:67–72. Available from: http://www.oriprobe.com/journals/tsgxjy/2012_4.html

[15] Wang J, Wu J. Conversation repair in the class with Chinese as foreign language. Theor Pract. Lang. Stud. 2015;5:1525–1533. Available from: http://www.academypublication.com/ojs/index.php/tpls/article/view/tpls050715251533/320

[16] Chen KY, Wang HM, Chen HH. A probabilistic framework for Chinese spelling check. ACM Trans. Asian Low-Resour. Lang. Inform. Process. 2015;14:15. Available from: http://dl.acm.org/citation.cfm?id=2826234

[17] Jiang Y, Wang T, Lin T, et al. A rule based Chinese spelling and grammar detection system utility. System Science and Engineering (ICSSSE), 2012 International Conference on IEEE. 2012:437–440. Available from: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6257223&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6257223

[18] Xiong J, Zhao Q, Hou J, et al. Extended HMM and ranking models for Chinese spelling correction. CLP. 2014;2013:133–138. Available from: www.aclweb.org/anthology/W14-6821

[19] Zheng M, Lei Z, Yue L. Identifying sentiment objects from Chinese sentences based on skip chain conditional random fields model. Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012 Sixth International Conference on IEEE. 2012, 390–394. Available from: http://www.gcccce2012.org/Conference%20Proceedings.pdf

[20] Lee LH, Chang LP, Lee KC, et al. Linguistic rules based Chinese error detection for second language learning. Proceedings of the 21st International Conference on Computers in Education (ICCE’13). 2013:27–29.

[21] Tseng YH, Lee LH, Yu LC. Traditional Chinese parsing evaluation at SIGHAN bake-offs 2012. Proceedings of the 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP’12). 2012;199–205. Available from: http://www.aclweb.org/anthology/W12-6335

[22] Tseng YH, Tseng CC, Hsin SC, et al. Detecting new vocabulary from news articles for Chinese teaching as a second language. Proceedings of Global Chinese Conference on Computers in Education (GCCCE’12). 2012, 551–552. Available from: http://www.gcccce2012.org/Conference%20Proceedings.pdf

[23] Xiong J, Zhang Y, Zhou X, et al. Problematic situation analysis and automatic recognition for Chinese online conversational system. CLP. 2014;2014:43–51. Available from: http://www.aclweb.org/anthology/W14-6808

[24] Yeh JF, Chen WY, Su MC. Chinese spelling checker based on an inverted index list with a rescoring mechanism. ACM Trans. Asian Low-Resour. Lang. Inform. Process. 2015;14:17. Available from: http://dl.acm.org/citation.cfm?id=2826235

[25] Yeh JF, Yeh CK. Automatic classification of the “De” word usage for Chinese as a foreign language. Comput. Linguist. Chin. Lang. Process. 2015;20:65–78.

[26] Lin CJ, Chu WC. A study on Chinese spelling check using confusion sets and N-gram statistics. Int. J. Comput. Linguist. Chin. Lang. Process. 2013;23:17–48.

[27] Liu X, Cheng F, Duh K, et al. A hybrid ranking approach to Chinese spelling check. ACM Trans. Asian Low-Resour. Lang. Inform. Process. 2015;14:16. Available from: http://dl.acm.org/citation.cfm?id=2822264
[28] Hsieh YM, Bai MH, Huang SL, et al. Correcting Chinese spelling errors with word lattice decoding. ACM Trans. Asian Low-Resour. Lang. Inform. Process. 2015;14:18. Available from: http://dl.acm.org/citation.cfm?id=2791389

[29] Grzymala-Busse JW. A new version of the rule induction system LERS. Fund. Inform. 1997;31:27–39.

[30] Academia Sinica CKIP. Available from: http://ckipsvr.iis.sinica.edu.tw/

[31] Yu LC, Lee LH, Tseng YH, et al. Overview of SIGHAN 2014 bake-off for Chinese spelling check. Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP’14); 2014:126–132. Available from: http://www.aclweb.org/anthology/W14-6820