Chinese Word Spelling Correction Based on Rule Induction

Jui-Feng Yeh, Yun-Yun Lu, Chen-Hsien Lee, Yu-Hsiang Yu, Yong-Ting Chen
Department of Computer Science and Information Engineering, National Chiayi University
No.300 Syuefu Rd., Chiayi City 60004, Taiwan (R.O.C.)
{ralph, s1020447, s1020443, s1002967, s1003008}@mail.ncyu.edu.tw

Abstract

The importance of learning Chinese is increasing in the latest decades. However, the learning of Chinese is not easy for foreigners as a second language learning. Sometimes they write some text or document, but there always have many error words. So, how to detect the error word in document is becoming more then more important. This issue is very extensive, not only can help foreigners to learning Chinese but also can detect the error word. This paper had proposed method can divide five sections of structure: First sections are input sentence; second sections are parsing and word segmentation; third sections are fine the wrong word; forth sections are remove duplicate; fifth sections are final output. In this paper we use language model to detect Chinese spelling. It is had four part, E-Hownet, CKIP, similar pronunciation and shape dictionary, use the preset word to compare the word correction which in database. We use the bi-gram to promote our performance.

1 Introduction

Learning Chinese is very important in this era, because the Chinese is main market customers. Since the trend of the times, there have many of foreigners beginning to learn Chinese. But Chinese is not easy to learn, because sometime the same word has many pronouns, or same pronounce has different word, even the much the words have similar glyph. Chinese unlike English, there have thousands of words in Chinese, different combinations have different meaning. Although pronounce the same, but there will be different words with different meaning, sometimes there were having some misunderstanding because using the wrong word. So how to learn Chinese is very import research. This topic is extensive, not only for foreigners to learn Chinese, but also can help to detect the wrong word in the document.

In recent years, there has a lot of paper to research about Chinese learning. Chinese learning in today not only face to face teaching, but also can learn in a mobile system. There have many type smartphone applications about learning Chinese, sometimes there also has another country's language. Michael B. Syson et al. (2012) propose a system ABKD which is learning the game in multimodal, this system has two languages for learning, one is Chinese the other is Japanese. This system is learning about the Chinese Hanzi and Japanese Kanji by the game. Vincent Tam et al. (2012) use iOS-Base devices to propose an e-learning software, this device is extendible and ubiquitous, this paper proposes different learning type like it can learn the characters in correct stroke sequences of Chinese, it also has some mini-game to help learning Chinese. This author also proposes another paper is main on writing Chinese, and not only focus on iOS-base, but also for other smart phone (ex: android). Xiangyu Qiu et al. (2012) propose a method about learning Chinese font style and transferring, it's based on strokes and structure, they propose a new glyph decryption method, it divides the Chinese characters two parts, one is the stable side call structure, the other side is mutable called style. Mei-Jen Audrey Shih et al. (2011) propose an online system to learn Chinese, online learning system is convenience for user, it is assembled to abound environment and had a broad content search opportunity, this paper is focused on how to learn Chinese language effectively in an online learning environment. Lee Jo Kim et al. (2011) propose a tool which supports Chinese language teaching and learning system based on ICT-Base, this tool can help peer assisted

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 139–145,
Wuhan, China, 20-21 October 2014

learning environment. Lung-Hsiang et al. (2012) propose a mobile assisted system about learning vocabulary, they use the Mobile-Assisted language learning (MALL), they present two case studies in the Mobile-Assisted Language Learning, this system is main on two languages, one is English, the other is Chinese, specially it is not learning the word, it is learning about the "idioms" and learning how to construct sentences. Shang-Jen Chuanget al. (2011) propose a new recognition of Traditional Chinese handwriting by neural networks, their recognition of Traditional Chinese handwriting by PNN and SVM, their database is 20 people's Traditional Chinese handwriting, and use different quantization methods for everyone. Yingfei Wu (2011) proposes a learning system of "Chinese calligraphy" on mobile systems, Calligraphy is good for learning Traditional Chinese font, because it needs step by step to write the Chinese word, but calligraphy is not easy, even a word usually has many different font styles, the calligraphy need ink and paper, so they propose a new mobile system which can easy to learning Calligraphy without use paper and ink. David Tawei Ku et al. (2012) proposes the Chinese learning in situated learning, it is trend a ubiquitous learning environment, and the feature focuses on real life learning situation, and problem solving practice, this learning system divides two parts, one is integrating situated learning strategy and the other is context awareness technology. Yanwei Wang et al. (2011) proposes a discriminative learning method of MQDF (Modified quadratic discriminant function), MQDF is based on sample importance weights, this method is investigated and compared other discriminative learning methods about MQDF. DA-Han Wang (2012) propose a handwriting recognition system, this system commonly combines character classification confidence scores, they propose two regularized classes-dependent confidence transformation (CT) methods. Yunxue Shao (2011) propose a similar handwritten Chinese characters method base on multiple instance learning, they solved the problem by Asaboost framework, the method is found week classifiers to select some self-adapting critical regions. Lung-Hsiang Wong (2010) propose a Mobile-Assisted language learning (MALL), their have two case studies, and focus on "creative learner outputs", student in two studies language by one-to-one mobile devices, and capture the picture of the real life.Shih-hung Wu et al. (2013) propose a paper for Chinese Spelling Check task which at SIGHAN bake-off 2013, in this paper, thay describe all detail of the task for Chinese spelling check, include the task descripition, data preparation, performance metrics, and evaluation results.

This paper proposes five steps to find the wrong words in a document: First is input the sentence; the second uses the CKIP to word segment; the third is finding the wrong word. In the third step, the main method in this paper divides the words in document for three parts, The first is the single word, sometimes the single words mean there does not have any match word before or after this single word, in other words, is there maybe had word error, so we compose the word which before or after this single word, this word most be the single word too. After composing two of single words, it can generate a new word than regarded as a suspicious error word. The second is about idioms, most of the idioms are composed of four words, so we take the four words to pronounce and glyph to compare with the E-Hownet. The other words (ex: two words, three words), we also compare with the E-Hownet, if it can find the same word in E-Hownet, it means this word is correct, use it as a suspicious error word. Forth is remove duplicate, this step is remove the duplicate wrong word. Finally opput the file.

2 Method

In this section, our proposed method is to check out the foreigners will get the word wrong and then correct for the right word. The sentences written by people learning Chinese as a foreign language (CFL) may contain a variety of grammatical errors, such as word choice, missing words, and so on. It focuses on spelling errors in this bake-off. We will introduce the framework of the proposed system and method, which is divided into two parts: training phase and test phase that will describe in section 2.1 and section 2.2.

2.1 Training phase

As shown in figure 1, training phase is to construct the dictionary which is used in test phase, there are including the similar pronunciation & shape dictionary and training data dictionary. E-Hownet is used to find the wrong word and correct the wrong word, it also can use to construct the rule induction. And

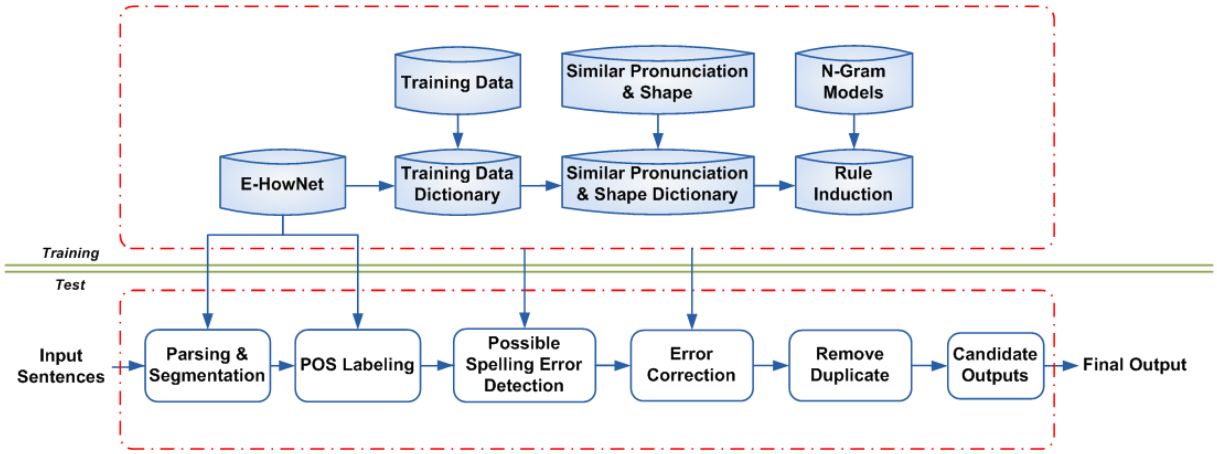

Figure 1: the framework of the proposed system.

n-gram models is also used to be the ranking score construct the rule induction. Finally, the candidate outputs are generated according to our rule induction. We will describe more detail in the follows.

First, we go to pre-process the data from the bake-off organizer. Step 1, we removed unnecessary portions of each sentence in the input file, such as PID number. The results will feed into the tool which is the CKIP Autotag, then it will do word segmentation and part-of-speech tagging based on E-Hownet. The corresponding part-of-speech (POS) of each word is obtained in the sentences. Each word has a part of speech at the end of a word in parentheses. Step 2, we are going to remove unessential blank spaces and parentheses. This step allows us to be more convenient for the implementation of our program. These processes are also used in the test phase.

Next, we introduce our rule induction in the following.

- Let A_i $(i = 1{\sim}n)$ mean incorrect word, $A_{ak}(k = 1{\sim}m,\ m \leq 4)$ mean k-th incorrect word, $Sim(A_{ak})$ mean the similar word with A_{ak}.
- Let $\mathrm{E-HN}(A_{a1}, A_{a2}, \ldots\ldots A_{am})$ mean that A_{a1} to A_{am} can combine into a word which can be find in E-Hownet, $\mathrm{LOC}(B_i)$ mean location of the word.
- $a_j = LOC\left(E - HN\left(A_{ap}, Sim\left(A_{aq}\right)\right)\right)$, $b_r = LOC(E - HN\left(Sim(A_{ap}), Sim\left(A_{aq}\right)\right))$, when $p = 1, \mathrm{q} = 2{\sim}\mathrm{m}$ or $q = 1, \mathrm{p} = 2{\sim}\mathrm{m}$. a_j $(\mathrm{j} = 1{\sim}\mathrm{m})$ mean that $A_{ap}, Sim(A_{aq})$ combine into a word which can find in E-Hownet, b_r $(r = 1{\sim}n)$ mean that $\mathrm{Sim}(A_{ap)}, Sim(A_{aq})$ combine into a word which can find in E-Hownet.
- $Min((a_1, a_2, \ldots, a_m), (b_1, b_2, \ldots, b_n)$ mean that output the minimum, this indicates the position of the front in the E-Hownet which is the more correct word.

2.2 Test phase

In the previous section, the rule induction is built in training phase. We will describe the test phase of the framework in this section. The word segmentation and part of speech (POS) labeling are the same as training phase. Then, we begin the processes with the third step, we have to detect the wrong word. There are some proposed method to find the wrong word in the following.

- In the previous step, we have the word segmentation, we choose the words more than two characters, then compared the words with E-Hownet or training data dictionary. If there is not the same words in E-Hownet or training data dictionary, we determine it as incorrect words.
- For the judge idioms, we choose all word with four characters, and compared the word with E-Hownet and the similar pronunciation & shape dictionary. If there is not the same words in the training data dictionary, E-Hownet or similar pronunciation & shape dictionary, we determine it as incorrect words.
- To the judge the sentences written by CFL, we focus on "的 (De)", "地 (De)", "得 (De)". Behind the "的 (De)" must connect the verb, behind the "地 (De)"

must be a noun. Further, behind the "得 (De)" must be a verb, fornt the "得 (De)" can be an adverb, Nv or Nh. If the characters do not comply with the POS of the above, we determine it as incorrect words.

- Finally, we strengthen the judgment of single character. Behind or found the single character is the same as single characters, we combine the character to the word which contain two characters. And we determine it as incorrect words.

According the above, we begin the processes which is comparing the wrong words with similar pronunciation & shape dictionary, that is in order to find the similar words, then if the similar words can be found in E-Hownet or training data dictionary, we saved the incorrect words in a text file named wrong, and saved the similar words in a text file named correct, this focuses on two characters of the word in the case of one character wrong. The proposed method also aim the two characters of the word in the case of all characters wrong, eg., 勞刀 (嘮叨). The processes as is same as above, but the incorrect words saved in a text file named double_wrong, and the similar saved in a text file named double_correct. The fifth step, we are going to remove duplicates. First, If the words in the text named wrong can be found in the text named double_wrong, we will remove the words in wrong. Second, if identify the words appear more than twice, we will remove the unnecessary words. It is helping us to reduce the process time. We will output the result in the final step. The processes will find the words and find the corresponding sentence, then save the position and correct word in the file named output. Finally, according the PID to sort the sentence and output to the specified format. For example, input: "(pid= A2-1051-1) 後天是小明的生日，我要開一個無會。", output: "A2-1051-1, 15, 舞", If the input contains no spelling errors, the system should return "pid, 0".

3 Experiments

According to the Chinese spelling check task in SIGHAN, this paper is dedicated to the detection and correction of errors in sentences. The evaluate is divided into two parts: Subtask 1 is detection level that is to find out the location of incorrect spelling characters in the sentences, then the subtask 2 is correction level, which is to find out the location of spelling error in subtask 1 and then correct the error. In section 3.1, we will describe the data sets, performance metrics, then we will show our evalution in section 3.2.

3.1 Data sets

```
<ESSAY title="少子化現象">
<TEXT>
<PASSAGE id="C1-1792-1">在日本行成「少子化」現象的可能原因有一些。其中一個是「晚婚化」。</PASSAGE>
</TEXT>
<MISTAKE id="C1-1792-1" location="4">
<WRONG>行成</WRONG>
<CORRECTION>形成</CORRECTION>
</MISTAKE>
</ESSAY>
```

Figure 2: an example of the training data.

In this bake-off, the evaluation is an open test. Participants can employ any linguistic and computational resources to develop the spelling checker, and provide passages of CFL's essays from the NTNU learner corpus for training purpose. The corpus was released in SGML format which is shown in figure 2. Moreover, there are at least 1000 different degrees of difficulty of testing passages for testing. In this paper, we use C++ to develop our proposed method.

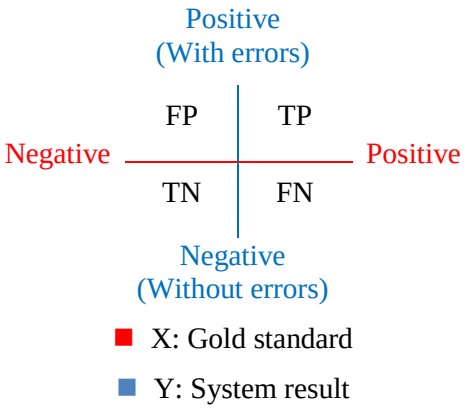

Figure 3: A quadrant map of performance metrics.

The judging correctness are divided into two parts: detection level and correction level. The following are showing some performance metrics and quadrant map shown in figure 3 that is measured in both levels of indicators:

- **TP**: System determines the character for errors related to the actual error, and the judgments the system is correct.
- **FP**: System determines the character for errors is not related to the actual error, and the judgments of the system is incorrect.
- **FN**: System determines the character for errors is related to the actual error, and the judgments of the system is incorrect.
- **TN**: System determines the character for errors is not related to the actual error, and the judgments of the system is correct.

The following is the performance metrics in this

- $\textbf{False Positive Rate} = \frac{FP}{(FP+TN)}$
- $\textbf{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$
- $\textbf{Precision} = \frac{TP}{(TP+FP)}$
- $\textbf{Recall} = \frac{TP}{(TP+FN)}$
- $\textbf{F1} - \textbf{Score} = \frac{2 \times Precision \times Recall}{(Precision+Recall)}$

3.2 Evaluation

Figure 4 is our data of evaluation, which the largest difference between the first and the second. The proposed method is only aimed to the training data in run1, then we make changes for run2 in the data which is provided by run1.

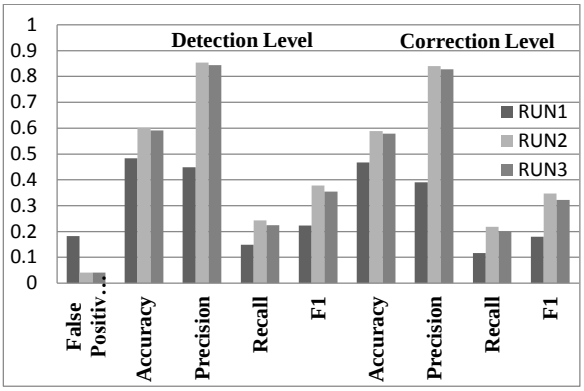

Figure 4: Performance evaluation.

According to the table 1, our false positive rate is the third in this bake-off, which means that our proposed method is feasible, but there is room for improvement. There are two parts of performance evaluation: detection level and correction level which is shown in table 2 and table 3.In the accuracy and precision, we can see that our proposed method can be the top three, but our method in recall is relatively weaker than another. This performance evaluation shows that our method is viable, but our method may be overly strict cause our relatively low

Table 1: Top five of the false positive rate.

Participating teams	False Positive Rate
NCYU*	0.0414
NCTU&NTUT	0.0377
SUDA	0.032
KUAS	0.0452
NTHU	0.0829

Table 2: Top four of performance evaluation in Detection Level.

Participating teams	Accuracy	Precision	Recall	F1
NCYU*	0.6008	0.8543	0.2429	0.3783
KUAS	0.7194	0.9146	0.484	0.633
CAS	0.6149	0.7148	0.3823	0.4982
SJTU	0.5471	0.5856	0.322	0.4156

Table 3: Top four of performance evaluation in Correction Level.

Participating teams	Accuracy	Precision	Recall	F1
NCYU*	0.5885	0.8406	0.2185	0.3468
KUAS	0.7081	0.9108	0.4614	0.6125
CAS	0.5829	0.676	0.3183	0.4328
SJTU	0.5377	0.5709	0.3032	0.3961

4 Conclusions

This study proposes a method for Chinese text detect spelling error. The method in our study is focus on word classify to easy detect Chinese spelling error. The word is classifying three class, single word, idioms and other words (two words, three words et.)The experimental result shows the performance it good, and we also apply this method in "SIGHAN 8 Chinese spelling check task", and the final result pretty good. In the feature, we hope can raise the performance and find the other word classifies. More word class can helpful to find the Chinese spelling error. After the Chinese spelling error, we will start to study the relationship between grammar and spelling errors, because in this paper we only care about the word pronouns and glyph, but in recent years some spelling error has been regularization, it most to understanding the context then detect it is right or wrong, so the issue about the relationship between grammar and spelling errors is need to study, if we can fine the relationship then the Chinese spelling detect correct rate must can raise higher.

Acknowledgments

This work is supported in part by the National Science Council, Taiwan, R.O.C., under the project grant numbers NSC 102-2221-E-415-006-MY3.

Reference

Qiu, X., Jia, W., and Li, H. 2012. A Font Style Learning and Transferring Method Based on Strokes and Structure of Chinese Characters. In Computer Science and Service System (CSSS), pp. 1836-1839.

Syson, M. B., Estuar, M. R. E., and See, K. T. 2012. ABKD: Multimodal Mobile Language Game for Collaborative Learning of Chinese Hanzi and Japanese Kanji Characters. In Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 03 pp. 311-315.

Tam, V., and Cheung, R. L. 2012. An Extendible and Ubiquitious E-learning Software for Foreigners to Learn Chinese on iOS-Based Devices. InAdvanced Learning Technologies (ICALT), 2012 IEEE 12th International Conference on pp. 46-48.

Tam, V., and Huang, C. 2011. An Extendible Software for Learning to Write Chinese Characters in Correct Stroke Sequences on Smartphones. InAdvanced Learning Technologies (ICALT), 2011 11th IEEE International Conference on pp. 118-119.

Li, K. H., Cheng, T. F., Lou, S. J., and Tsai, H. Y. 2012. Application of Game-based Learning (GBL) on Chinese language learning in elementary school. In Digital Game and Intelligent Toy Enhanced Learning (DIGITEL), 2012 IEEE Fourth International Conference on pp. 226-230.

Ku, D. T., and Chang, C. C. 2012. Development of context awareness learning system for elementary Chinese language learning. In Proceedings of the 2012 Sixth International Conference on Genetic and Evolutionary Computing on pp. 538-541.

Tam, V., and Luo, N. 2012. Exploring Chinese through learning objects and interactive interface on mobile devices. In Teaching, Assessment and Learning for Engineering (TALE), 2012 IEEE International Conference on pp. H3C-7.

Shih, M. J., and Yang, J. C. 2011. How to Learn Chinese through Online Tools? From the Perspective of Informal Learning to Culture Immersion. InAdvanced Learning Technologies (ICALT), 2011 11th IEEE International Conference on pp. 305-306.

Kim, L. J., Lim, S. H., and Ying, L. T. 2011. ICT-based peer assisted learning environment: Using online feedback tools for Chinese Language writing tasks. In Electrical and Control Engineering (ICECE), 2011 International Conference on pp. 6612-6614.

Wong, L. H., and Looi, C. K. (2010, April). Mobile-assisted vocabulary learning in real-life setting for primary school students: Two case studies. In Wireless, Mobile and Ubiquitous Technologies in Education (WMUTE), 2010 6th IEEE International Conference on pp. 88-95.

Chuang, S. J., Zeng, S. R., and Chou, Y. L. 2011. Neural Networks for the Recognition of Traditional Chinese Handwriting. In Computational Science and Engineering (CSE), 2011 IEEE 14th International Conference on pp. 645-648

Wu, Y., Yuan, Z., Zhou, D., and Cai, Y. 2013. Research of virtual Chinese calligraphic learning. In Multimedia and Expo (ICME), 2013 IEEE International Conference on pp. 1-5.

Ku, D. T., and Chang, C. C. 2012. Development of context awareness learning system for elementary Chinese language learning. In Proceedings of the 2012 Sixth International Conference on Genetic and Evolutionary Computing on pp. 538-541.

Zhao, Z., and Ma, X. 2012. Prediction of Prosodic Word Boundaries in Chinese TTS Based on Maximum Entropy Markov Model and Transformation Based Learning. In Computational Intelligence and Security (CIS), 2012 Eighth International Conference on pp. 258-261.

Lin, C. C., and Tsai, R. H. 2012. A Generative Data Augmentation Model for Enhancing Chinese Dialect Pronunciation Prediction. Audio, Speech, and Language Processing, Transactions on, 20(4), 1109-1117.

Wang, Y., Ding, X., and Liu, C. 2011. MQDF discriminative learning based offline handwritten Chinese character recognition. In Document Analysis and Recognition (ICDAR), 2011 International Conference on pp. 1100-1104.

Shao, Y., Wang, C., Xiao, B., Zhang, R., and Zhang, Y. 2011. Multiple instance learning based method for similar handwritten Chinese characters discrimination. In Document Analysis and Recognition (ICDAR), 2011 International Conference on pp. 1002-1006.

Wu, S. H., Liu, C. L., & Lee, L. H (2013). Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013. Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing (SIGHAN'13), Nagoya, Japan, 14 October, 2013, pp. 35-42

Academia Sinica CKIP. http://ckipsvr.iis.sinica.edu.tw/

Academia Sinica E-Hownet. http://ehownet.iis.sinica.edu.tw/