Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

Detecting Grammatical Error in Chinese Sentence for Foreign

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: Each language has its own unique grammar, Chinese is same as the other languages similarly. But each language has different type even there does not have any relations. So the foreigners learn the language not only need to learning the word pronouns and glyph, but also need to learn the grammar. This issue is very extensive, not only can help foreigners to learn Chinese, but also can detect the error grammar. This paper had proposed method can divide five sections of the structure: First sections are input sentence; second sections are parsed and word segmentation; third sections find the missing, wrong word; fourth sections find the redundant wrong word; fifth sections are final output. This paper has two parts, the first is how to detect the grammar error, and the second is how should we know the Chinese grammar error is what type. Finally, we can get the type of the grammar, and we can know how to correct.

Keywords: CFL, Chinese word correction, grammatical error, rule induction.

1. Introduction

Learning Chinese is more famous than before, not only there has more and more Chinese business, but also there has many tourist attractions csuse the foreigners love to go there to travel. So, there have many of foreigners beginning to learn Chinese. But Chinese is not easy to learn like the other languages, it is not only having many pronunciations and glyph of the word, but also have many grammar of the sentence. The Chinese although has subject, verbs and object too, but there has a combination of fixed text, if you do not to comply with these rules, the meaning of the sentence itself will be different.So how to learn Chinese grammar is very import research. This topic is extensive, not only for foreigners to learn Chinese, but also can help to detect the wrong grammar in the document.

In recent years, there has a lot of paper to research about Chinese learning grammar. Most of paper about learning Chinese paper not only talk about the sentence correct rate, but also talk about the Grammar correct rate. Ying Jiang (2012) proposes an arithmetic called “Rnture-Sentence” to segment the Chinese sentences, it can segment the sentence more completely, and solve the problem about the complicated Chinese grammar, proofreading method, their method to cross sentences is based on LanguageTool, their paper also presents to some method of new rule which can accomplish complicated Chinese grammar proofreading. These authors also propose another Chinese grammar, proofreading (2013), the presents an indexing method of a corpus base of the Chinese grammar, they can evaluate the rule of the accuracy and the frequency, each rule, they adopt an iterative approach to improve it, make sure its better performance in the real word, it also introduces the important role of the corpus of their method. Mei-Jen Audrey Shih et al. (2011) propose a Chinese online learning system, this online system is convenience and it is assembled to abound environment and had a broad content search opportunity, this paper is mained on how to learn Chinese language effectively in an online learning environment. Lee Jo Kim et al. (2011) propose a Chinese language teaching and learning system based on ICT-Base tool, this tool can help peer assisted learning environment. Ying Jiang et al. (2012) they provide a new method to deal with grammar error. They arrange a new rule for their new grammar system, this system has two combined characters with intention, then their grammar system can contain as well as the spelling error system, so their system can have perfect respect of precision and practicality. David Tawei Ku et al. (2012) proposes a situated learning for

Chinese learning, it is a trend that ubiquitous learning environment, and the feature main on real life learning situation, and problem solving practice, this learning system has 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. Lung-Hsiang Wong (2010) propose a Mobile-Assisted Language Learning (MALL), there have two case studies, and mained on "creative learner outputs", student in two studies language by one-to-one mobile devices, and capture the picture of the real life. Hui Yang et al. (2010) proposes a continuous prior polarity algorithm, their method reflects subtle changes of sentiment in contrast, its previous studies which expressed sentiment polarity discretely, they also proposed a method based on Chinese dependency grammar which can assess modified polarity, they can accurately identify subjective words and its modified according different Chinese dependency grammar, then predict the sentence by aggregate. Peng Li et al. (2012) proposes “A Hierarchy-based Constraint Dependency Grammar Parsing for Chinese”, they mentioned the Constraint Dependency Grammar (CDG) is a famous formalism which about the grammatical rules, and they have successfully adopted in Chinese, they propose to develop a three schema which is based on a study of constraining in the corpus. Haiping Zhu et al. (2011) propose a analyze Chinese sentence with semantic dependency method, the correlation between words and phrases can calculation of the similarity from Hownet, between the sentence and sentence’s similarity can analyze by formula, their method can be adopted to analyze the correct topic and to categorize.

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 find the missing wrong word; forth sections are find the redundant wrong word; fifth sections are final output.In the third section, we classify the four of the grammar rule, the part of speech(POS) can classify four type(Shi, Neu, D, DA), and there have regular POS behind the four type of the POS.After find, we put the worng grammar part in the dictionary file which only for these error, the dictionary file name is Miss . In the forth section, find the redundant wrong word, there has a special rule,behind the POS of DE must a DE, if did not, it will write in the dictionary file which name is “Redundant”. finally, we will use these dictionary files to detect the Chinese Grammar, if the error is bellowing “Missing” diIionary file, the error is bellowing “Redundant”dictionary file, if not miss error or redundant error, that means it it correct, output to correct the file.

2. Method

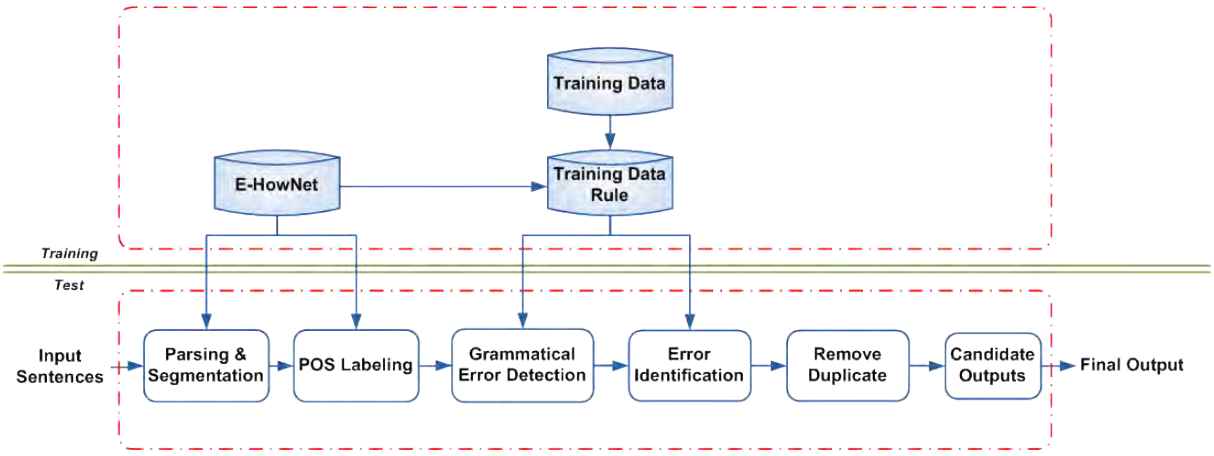

Figure 1. The framework of the proposed system.

In this section, we will introduce the framework of the proposed system and method. Our proposed method is aimed to detect and identify the sentence for learning Chinese as a foreign language (CFL). The sentences written by CFL may contain a variety of grammatical errors, such as word choice, missing words, word disorder and so on. It focuses on grammatical errors in this task. And the

framework 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

The training phase shown in figure 1, we have some data that are contain some sentences can be trained to find some useful information. First, we do the pre-process to the data from the task organizer, that will be input file and we removed unneeded portions of each sentence in this file, such as SID number, then the treated results will be further inputted into the tool which is CKIP AutoTag, that is to do the word segmentation and part-of-speech (POS) tagging based on E-Hownet. The corresponding part-of-speech of each word is obtained in the sentences, which is given a part of speech at the end of a word in parentheses. Second, we are going to remove unessential blank spaces and parentheses, that will be more convenient in the following file operations. In the test phase, we are also adopted in this way. Then, we want to find some rules with training data which can be used in test phase. We construct the training data rule from the results of process which have part-of-speech. Finally, the candidate outputs are generated according to our training data rule.

2.2 Test phase

In the previous section, the training data rule 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 & fourth step, we have to detect and identify the wrong word. The following is focused on finding the Missing type of the wrong words.

- Behind the word with POS of “Shi” is not connected the word with POS of “Verb”.
- Behind the word with POS of “Neu” is not connected the word with POS of “De”.
- Behind the word with POS of “D” is not connected the word with POS of “Neqa”.
- Behind the word with POS of “Da” is not connected the word with POS of “Neu”.

According to above, the Missing type of the incorrect words will save in the file named missing. And the following is focused on finding the Redundant type of the wrong words.

- Behind the word with POS of “De” is not connected the word with POS of “De”.
- Behind the word with POS of “P” is not connected the word with POS of “P”.
- Behind the word with POS of “Cbb” is not connected the word with POS of “D”.
- Behind the word with POS of “Vh” is not connected the word with POS of “D”.
- Behind the word with POS of “De” is not connected the word with POS of “Neqa”.
- Behind the word with POS of “D” is not connected the word with POS of “D”.

According to above, the Missing type of the incorrect words will save in the file named missing. We will remove repeated SID. It is helping us to reduce the process time. We will output the result in the final step. The processes will run the test data, if the word If the word exists in the missing file, we will output the sentence with Missing. The Redundant type of the incorrect word is same as Missing. Others are identified as correct. For example, input: “(sid=C1-1876-2) 對社會國家不同的影響”, output: “C1-1876-2, Missing”, If the input contains no errors, the system should return “C1-1876-2, correct”.

3. Experiments

According to the grammatical error diagnosis for learning Chinese as a foreign language in NLP-TEA-1, this paper is dedicated to the detection and identification of errors in sentences. The evaluate is divided into two parts: Subtask 1 is detection level that is to check out the sentence which

is incorrect or correct, then the subtask 2 is identification level, which is to identify the error type in sentences, i.e., Redundant, Missing, Disorder, and Selection. In section 3.1, we will describe the data sets, performance metrics, then we will show our evaluation in section 3.2.

3.1 Data sets

```
<ESSAY title="張愛文的一天">
<TEXT>
<SENTENCE id="A2-0101-1">下了課就去看他在學校對面的公車站</SENTENCE>
</TEXT>
<MISTAKE id="A2-0101-1">
<TYPE>Disorder</TYPE>
<CORRECTION>下了課就去在學校對面的公車站看他</CORRECTION>
</MISTAKE>
</ESSAY>
```

Figure 2. An example of the training data.

In this task, the evaluation is an open test. Participants can employ any linguistic and computational resources to develop the error diagnosis, and provide data 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.

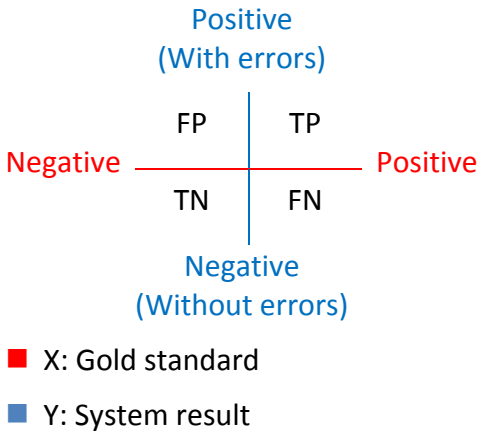

Figure 3. A quadrant map of performance metrics.

The judging correctness are divided into two parts: detection level and identification 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 of performance metrics are according to the quadrant map.

- **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

According to the table 1, our false positive rate is the best in this task, which means that our proposed method is feasible, but our proposed method just focuses on identifying two error type, . There are two parts of performance evaluation: detection level and identification level which is shown in table 2 and table 3. In the identification, we can see that accuracy is the best. Then, accuracy and precision are also comparable to others, but our method in recall is relatively weaker than another. This performance evaluation shows that our method is viable, but our method is still much room for improvement.

Table 1: Participating teams of the false positive rate.

Participating teams	False Positive Rate
NCYU*	0.1189
TMU	0.1691
UDS	0.2491
KUAS&NTNU	0.2686
CIRU	0.496
NTOU	1

Table 2. Participating teams of performance evaluation in Detection Level.

Participating teams	Accuracy	Precision	Recall	F1
NCYU*	0.4983	0.4927	0.1154	0.187
TMU	0.5171	0.5399	0.232	0.3245
UDS	0.4914	0.4945	0.7749	0.6037
KUAS&NTNU	0.5006	0.5003	0.9051	0.6444
CIRU	0.6446	0.6128	0.7851	0.6884
NTOU	0.5	0.5	1	0.6667

Table 3. Participating teams of performance evaluation in Identification Level.

Participating teams	Accuracy	Precision	Recall	F1
NCYU*	0.4594	0.2409	0.0377	0.0652
TMU	0.4554	0.3545	0.1086	0.1662
UDS	0.2337	0.2467	0.2594	0.2529
KUAS&NTNU	0.2149	0.2696	0.3337	0.2983
CIRU	0.4589	0.4548	0.4137	0.4333
NTOU	0.2074	0.2932	0.4149	0.3436

4. Conclusions

This study proposes a method for Chinese text detect grammar error. The method in our study is focus on word classify to easy detect Chinese grammar error. The grammar error is classifying four type, the verbs was not add behind POS of Shi, the De was not add behind the POS of Neu, the Neqa was not add behind POS of D, and the Neu was not add behind POS of Da. The experimental result shows the

performance it good, and we also apply this method in “grammatical error diagnosis for learning Chinese as a foreign language”, and the final result pretty good. In the feature, we hope can raise the performance and find the more grammar type . More grammar type can helpful to find the Chinese grammar error. After the Chinese grammar 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 grammar detect correct rate must can raise higher.

Acknowledgements

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.

References

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) on 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), pp. 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.

Li, P., Liao, L., and Li, X. 2012 . A hierarchy-based constraint dependency grammar parsing for Chinese. In Audio, Language and Image Processing (ICALIP), 2012 International Conference on pp. 328-332.

Jiang, Y., Wang, T., Lin, T., Wang, F., Cheng, W., Liu, X., ... and Zhang, W. 2012,. A rule based Chinese spelling and grammar detection system utility. In System Science and Engineering (ICSSE), 2012 International Conference on pp. 437-440.

Jiang, Y., Zhou, Z., Wan, L., Li, M., Zhao, W., Jing, M., and Liu, X. 2012. Cross sentence oriented complicated Chinese grammar proofreading method and practice. In Information Management, Innovation Management and Industrial Engineering (ICIII), 2012 International Conference on Vol. 3, pp. 254-258.

Jiang, Y., Lin, Z., Wang, J., Dai, M., Zhen, L., Li, N., ... and Meng, Y. 2013. Corpus Based Chinese Grammar Error Detection Rules Evaluation Method and System. In Proceedings of the 2013 Third International Conference on Intelligent System Design and Engineering Applications on pp. 496-499.

Zhu, H., Yang, Y., Chen, Y., and Ma, Q. 2011, September. Chinese sentence correlation analyzing based on semantic dependency method. In Electronics, Communications and Control (ICECC), 2011 International Conference on pp. 1932-1935.

Ma, L. L., and Wu, J. 2012, June. On-line handwritten Chinese character recognition based on inter-radical stochastic context-free grammar. In Neural Networks (IJCNN), The 2012 International Joint Conference on pp. 1-5.

Wu, C. H., Liu, C. H., Harris, M., & Yu, L. C. (2010). Sentence correction incorporating relative position and parse template language models. Audio, Speech, and Language Processing, IEEE Transactions on, 18(6), 1170-1181.

Chang, R. Y., Wu, C. H., & Prasetyo, P. K. (2012). Error Diagnosis of Chinese Sentences Using Inductive Learning Algorithm and Decomposition-Based Testing Mechanism. ACM Transactions on Asian Language Information Processing (TALIP), 11(1), 3.

Yu, C. H., & Chen, H. H. (2012). Detecting Word Ordering Errors in Chinese Sentences for Learning Chinese as a Foreign Language. In COLING on pp. 3003-3018.

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

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