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

KNGED: a Tool for Grammatical Error Diagnosis of Chinese Sentences

Tao-Hsing CHANG[a*], Yao-Ting SUNG[b], Jia-Fei HONG[c] & Jen-I CHANG[b]

[a]*Department of Computer Science and Information Engineering,*
National Kaohsiung University of Applied Sciences, Taiwan
[b]*Department of Educational Psychology and Counseling, National Taiwan Normal University, Taiwan*
[c]*Department of Applied Chinese Language and Culture, National Taiwan Normal University, Taiwan*
*changth@gm.kuas.edu.tw

Abstract: The main purpose of this paper is to propose a method that can automatically detect whether there are any grammatical errors as well as identify their error types. The framework of this method is based on a rule base to identify common grammatical errors. This rule base contains manually constructed rules and rules that are automatically machine generated. This paper further proposes algorithms which can apply these rules to determine whether a sentence is incorrect as well as what types of errors it belongs to. Experimental results show that the F1-measure of the proposed method is 0.64 and 0.30 on detection and identification, respectively.

Keywords: KNGED, Chinese grammar, grammatical error, CFL, automatic diagnosis, rule-based method.

1. Introduction

Automatically detecting grammatically incorrect sentences is fundamental and important for numerous NLP studies and related applications. For instance, in language teaching, automatic detection of grammatically incorrect sentences produced by learners can help a teacher teach grammar more effectively. However, the detection of grammatically incorrect sentences in Chinese is challenging. The main reason is that it is difficult to detect sentence boundaries in Chinese. In English, the period provides a clear signal of the end of a sentence, allowing the segment of text between two periods to be taken as a sentence and analyzed grammatically. However, two periods between Chinese sentences represent a complete semantic expression. An excessively long sentence may contain several commas as delimiters. Moreover, a sentence segment formed by a comma may comprise the complete sentence, a clause and even a phrase. This phenomenon makes the detection of sentence boundaries in Chinese difficult.

The above difficulty makes the method of detecting grammatical errors in Chinese sentences via using a parser to completely deconstruct a parsing tree infeasible. To detect errors in an English sentence, a parsing tree constructed using a parser can provide criteria for sentence judgment. However, to deal with grammatical errors generated by learners of Chinese as a second language, the parsing tree method may not have been thoroughly examined. This is because the main cause of common errors by second language learners is the language transfer phenomenon in language learning. For example, Korean students often write the following incorrect sentence.

我 來 台灣 四年 工作 了

The error pattern here is that the time noun ‘四年’ appears before the verb ‘工作’. This common error occurs because Korean is a subject-object-verb (SOV) language. An important characteristic of an SOV language is that all other elements such as nouns, adverbs, and numbers come before the verb. However, Chinese is a subject-verb-object (SVO) language, so the correct sentence would be '我 來 台灣 四年 工作 了 (I came to work in Taiwan for four years)' , where the verb ‘工作(work) ’ must appear before the time noun‘四年 (four years)’.

Consequently, by generalizing common errors made by learners of Chinese as a second language, it is possible to further analyze which specific syntactic structures those common errors belong to. These specific syntactic structures are considered error detection rules. Moreover, the error detection rules possess a syntactic error pattern and its corresponding syntactically correct pattern. Sufficient patterns and rules enable the generation of a rule base. Sentences can be compared using the rule base to identify grammatical errors. If a sentence contains multiple segments that conform to error detection rules, this segment is most likely the syntactic structure of that error. Therefore, sufficient error detection rules are collected, grammatical errors can be identified by comparing the rule base. The method of error detection rule collection can be obtained through analysis of sentences in a learner error corpus. The bigger the corpus, the more error detection rules can be generated and the more grammatical errors detected.

This study primarily proposes a method capable of automatically detecting and identifying grammatical errors. The framework of this method is based on a rule base to identify common grammatical errors. This rule base contains manually constructed rules and rules that are automatically machine generated. This study further proposes algorithms that can apply error detection rules to determine whether a sentence is incorrect and what types of errors it belongs to and classify any errors.

This paper is organized as follows. Section 2 reviews some related research and illustrates the impact of these studies on the research motivations. Section 3 then lists the corpus used in this paper and illustrates a learner corpus that is designed to automatically detect grammatically incorrect sentences produced by Chinese as second language writers. Next, section 4 introduces the method to manually construct error detection rules and the method for the program to automatically generate error detection rules. Subsequently, section 5 describes the algorithm for automatically identifying incorrect sentences. Section 6 demonstrates the performance of the proposed approach, while Section 7 draws conclusions.

2. Related Works

Recently, Chinese learning has become a growing trend, making Chinese one of the most popular foreign languages globally, besides English. To learners, learning a new language frequently involves grammar difficulties, and grammatically incorrect sentences are a common error. Previous research on second language acquisition indicated that effective provision of corrective feedback can contribute to the development of grammatical competence in second language learners (Fathman and Whalley, 1990; Ashwell, 2000; Ferris and Robers, 2001; Chandler, 2003). Currently, in the field of natural language processing, the development of tools and technologies to automatically detect grammatical errors is an important research trend.

On the one hand, regarding common error types made by learners of English as a second language] and the development of the related automatic detection research, Donahue (2001) used the error taxonomy of native English learners proposed by Connors and Lunsfor (1998) to analyze two hundred writing tests taken by learners of English as a second language. The most common error types committed by learners of English as a second language were found to differ from those of native English learners. The three most common error types of learners of English as a second language were incorrect usage of commas, incorrect word usage, and missing words. However, this corpus was insufficient to understand common error types committed by most learners of English as a second language. Cambridge University Press collaborated with the University of Cambridge to create the Cambridge Learner Corpus (CLC), which tags approximately 16 million words. Among these words, the three most common error types are incorrect word selection, preposition errors, and determiner errors (Nicholls, 2003).

During the past ten years, natural language processing specialists have designed automated grammatical error detection techniques and tools focused on common error types in the corpus. Examples include preposition error detection by Eeg-Olofsson and Knuttson (2003), Tetreault and Chodorow (2008), DeFelice and Pulman (2009) and Tetreault and Chodorow (2009), article and preposition error detection by Gamon et el. (2009) and Dale and Kilgarriff (2011), determiner and proposition error detection by Dale *et al.* (2012), and determiner, article, and proposition error detection by Ng *et al.* (2013).

On the other hand, regarding common error types in learners of Chinese as a second language, Wang (2011) observed that the most common grammatical error types among Chinese learners whose mother tongue is English are missing language components, incorrect word order, and incorrect sentence structure. Additionally, analysis of the HSK corpus of 35,884 erroneous sentences has demonstrated that the three most common error types are incorrect word order, missing adverb components, and missing predicate components (Cheng *et al.* 2014). With the development of related automatic detection research, Cheng *et al.* (2014) and Yu and Chen (2012) designed word order error detection technology focused on the Chinese sentences in the HSK Dynamic Composition Corpus. In developing a sentence grammatical error detection system, Lee *et al.* (2014) further used the HSK Dynamic Composition Corpus and additional manually constructed rules of common Chinese sentence errors.

The above literature indicates that, in English learning, there exists widespread use of learning assistance tools developed from natural language processing technology. These tools can automatically detect and correct the grammatical errors of learners. This is valuable as a means to help learners learn correct grammar and improve their compositional skills (Chodorow *et al.*, 2012; Leacock, Chodorow, Gamon, &Tetreault, 2010). However, little research has examined automatic detection of Chinese grammatical errors. This study proposes the integration of rule-based and machine learning methods to identify reliable rules from the corpora to detect the grammatical errors of learners of Chinese as a second language.

3. Corpora

This study seeks to obtain reliable rules to detect grammatical errors committed by learners of Chinese as a second language. The three corpora used in this study include (1) the dry run data provided by the convention; (2) the formal run data provided by the convention; (3) Chinese Written Corpus developed herein. The following focuses on introducing Chinese Written Corpus.

This study has continually developed a Chinese Written Corpus primarily comprising a single topic at different levels. This corpus was developed using Chinese composition scoring guidelines based on the ACTFL (2012) language proficiency criteria. The research samples are compositions written by foreign students who have learned enough Chinese to have basic competence. Samples were collected from September 2010 to June 2013. The source of the corpus is foreign Chinese learners studying at the National Taiwan Normal University Mandarin Training Center and 11 other Taiwanese Chinese educational institutions. The corpus currently includes foreign learners representing 37 different mother languages. During composition collection, complete information was collected on each composition. This information included the title of the composition, the Chinese and English names, nationality, and mother tongue of the learner, and the Chinese education institution in Taiwan. This information was saved as text and image documents. Currently, the texts of this corpus deal with two topics, and there are 1,147 compositions in total, comprising approximately 750 thousand words.

Following the creation of the corpus, each composition text was assessed by two experts or personnel trained in evaluation. To ensure reliability, the texts were cross-evaluated using the Chinese Composition Scoring Standard. This standard assigns compositions to different rankings of Distinguished, Superior, Advanced, Intermediate, and Novice. The Advanced, Intermediate, and Novice categories are each divided into three subcategories, including High, Medium, and Low, amounting for a total of 11 categories. These 11 categories account for Chinese users of all levels, from learners unable to construct a full sentence to native level writers. This study manually scored the compositions in each complete topic based on the above scoring standards and procedures. The composition scores were collected, and learner and corpus information were inputted into an error tagging system developed herein for compositions by learners of Chinese as a second language.

This error tagging system compiles learner and corpus information, and also includes word segmentation, part-of-speech tagging, and error tagging functions. In the main error tagging system, methods and standards for the analysis of learner language errors can be roughly divided into “linguistic form taxonomy” and “surface structure taxonomy”. Linguistic form taxonomy classifies error types – word class, sentence, and specific sentence errors – using language components as a structure. Meanwhile, surface structure taxonomy classifies error types using their structure. That is, it compares the correct and incorrect forms. Typical surface structures comprise four categories: omission, addition, selection, and disorder (Dulay, Burt & Krashen, 1982; James, 1998). The function of error tagging in this study integrates the two taxonomies, first classifying errors based on the surface structure, then carefully analyzing them based on the language form.

4. Rule Generation and Extraction

Based on the above three data types used in this study, this section explains the optimal method of generating error detection rules to identify ungrammatical sentences.

4.1 Manually Constructed Rules

The study uses five steps to generate manually constructed rules. First, based on the training data provided in this shared task, this study handcrafted syntactic patterns of grammatically incorrect sentences and corrected sentences. Second, to ensure the reliability of manually constructed rules for detecting incorrect sentences, this study also devised a program in which the Chinese Written Corpus developed in this study is embedded. Thirdly, on program completion, we enter syntactic patterns of grammatically incorrect sentences into the interface, and the program can then show the number of sentences contained in the Chinese Written Corpus. Moreover, those sentences conform to syntactic patterns of grammatically incorrect sentences.

Meanwhile, this study entered syntactic patterns of corrected sentences into the program, and then recorded the number of sentences contained in the Chinese Written Corpus, as well as those sentences that conform to the syntactic patterns of corrected sentences. Finally, this study retains the number of syntactic patterns of corrected sentences such that it exceeds that of incorrect sentences. These rules are considered the reliable error detection rules for identifying grammatically incorrect sentences in formal run data. This study contains 840 manually constructed rules, which contain 90 rules for identifying sentences with Missing words, 73 for identifying sentences with Redundant words, 51 for identifying sentences with Selection words, and 626 for identifying sentences with Word disorder.

4.2 Machine Generated Rules

The advantage of manually constructed rules is that complex rules can be detected with high accuracy. However, using manually constructed rules to identify grammatical errors suffers from a disadvantage. Specifically, the number of manually constructed rules is limited, and errors may exist. This study thus employs a program to retrieve syntactic rules of ungrammatical sentences from the learner corpus.

Unlike manually constructed rules, the rules generated by the program are fixed in length. For example, the learner corpus contains the following sentence.

這些	地方	是	在	巴西
Neqa	Na	SHI	P	Nc

In this sentence, each part of speech is labeled. This sentence in the learner corpus is tagged as the Redundancy error, and ‘SHI’ is a redundant word. We hypothesize that every word in this sentence can be collocated with its beginning and end, and their parts-of-speech to generate rules. Therefore, we combine “是” and its part-of-speech “SHI” with the first and last parts of the word “是” and their associated parts of speech, which yields 32 possible Redundant rules, as shown in Fig. 1.

In Figure 1, the symbol "+" represents two adjacent words or parts-of-speech, while the symbol ">" indicates that both the front and the back of a word or its associated part-of-speech should not be adjacent to that symbol. For a rule *pr* included in these 32 possible rules, if it meets the following criteria, it will be recognized as an error detection rule:

$$positive(pr) > p \quad \text{and} \quad r > \underline{\;k,}$$
$$r = positive(pr)\ /\ negative(pr)$$

where *positive(pr)* indicates the number of *pr* that occurred in the corpus with erroneous sentences; and *negative(pr)* indicates the number of *pr* that occurred in the corpus with correct sentence. In this study, the value of *positive(pr)* divided by *negative(pr)* is denoted as the *r*-value. The *r*-value of rules used by the grammatical error diagnosis algorithm described in Section 5.

Parameters *p* and *k* are thresholds obtained via experiment. Larger *p* is associated with more occurrence of rule *pr* in the incorrect sentences. That is, the rule of *pr* does not appear randomly. Meanwhile, larger *k* represents the possibility of a high degree of precision when using *pr* to identify a sentence as erroneous. Take 32 rules in Fig. 1 for example; if *p* and *k* are set to 2, then just 11 rules with borders in Fig. 1 are collected in the rule base for detection. This study uses the above method to automatically generate 13,890 Redundant rules and 2,497 Missing rules.

(1) 這些>是+在
(2) 這些>是+P
(3) 這些>是>巴西
(4) 這些>是>Nc
(5) 地方+是+在
(6) 地方+是+P
(7) 地方+是>巴西
(8) 地方+是>Nc
(9) 這些>SHI+在
(10) 這些>SHI+P
(11) 這些>SHI>巴西
(12) 這些>SHI>Nc
(13) 地方+SHI+在
(14) 地方+SHI+P
(15) 地方+SHI>巴西
(16) 地方+SHI>Nc
(17) Neqa>是+在
(18) Neqa>是+P
(19) Neqa>是>巴西
(20) Neqa>是>Nc
(21) Na+是+在
(22) Na+是+P
(23) Na+是>巴西
(24) Na+是>Nc
(25) Neqa>SHI+在
(26) Neqa>SHI+P
(27) Neqa>SHI>巴西
(28) Neqa>SHI>Nc
(29) Na+SHI+在
(30) Na+SHI+P
(31) Na+SHI>巴西
(32) Na+SHI>Nc

Figure 1. Examples of Rules Generated by Machine.

5. Grammatical Error Diagnosis Algorithm

For each sentence, the following steps are performed to determine whether it is incorrect.

Step 1. Check for rules that conform to the error detection rule of Word selection. If such rules exist, the sentence is considered to contain Word selection error and so the error identification is concluded].

Step 2. Check whether rules exist that conforms to the error detection rule of Word disorder. If so, the sentence is considered to contain Word order error and so the error identification is concluded.

Step 3. Check for rules that conform to the error detection rule for Redundant and Missing words.

Step 3.1. If the rule only conforms to one of the error detection rules related to redundant or missing words, then it is considered a sentence that contains that type of error and so the error identification is concluded.

Step 3.2. If the rule simultaneously conforms to more than one error detection rule of redundant or missing words, then among the rules that conforms to both types of error, that with the highest *r* value is selected.

Step 3.3. It is assumed that among the Missing word rules, the highest value of *r* is *mr*, and among the Redundant word rules, the highest value of *r* is *rr*. If the *r* value of *rr* exceeds *y* times that of *mr*, the sentence is considered to suffer from Redundant word error; otherwise, it is considered to suffer from Missing word error. The identification is concluded following sentence judgment.

Step 4. If the sentence is not recognized as erroneous via the last three steps, then it is considered correct.

Because different types of error detection rules exert different effects, based on analysis of error detection rules from the dry run corpus, their effectiveness reveals that the Selection has higher accuracy than other types of rule. Consequently, when a sentence is identified as containing segments of the rule of Selection, it is recognized that the sentence contains that type of error. Similarly, although the accuracy of the Word disorder rule is lower than that of the Selection rule, it is far higher than that of the Redundant word and Missing word rules. Therefore, when a sentence is identified as containing the Word disorder rule, it is first recognized that the sentence contains that type of error.

Compared to the Missing word rule, the redundant word rule can more easily obtain a higher *r* value. Thus, if the *r*-value of the Redundant word rule must exceed the missing word rule by y times, then the result of the detection of the rule of Redundant word can be reliable; otherwise, the sentence should be recognized as containing a Missing word error. The next section illustrates the value of each parameter used in the proposed method.

6. Experimental Results

In the NLPTEA 2014 CFL shared task, three parameters are established and combined with three runs to evaluate the effectiveness of the proposed method. In Run 1, the p-value is 3, the k-value is 2, and the y-value is 50. In Run 2, the p-value is 10, the k-value is 1000 and the y-value is 50. In Run 3, the p-value is 3, the k-value is 2 and the y-value is 1. Table 1 lists the experimental results.

Table 1: An example of a table for the ICCE proceedings.

Submission		Run1	Run2	Run3
False Positive Rate		0.9040	0.2686	0.9040
Detection Level	Accuracy	0.5006	0.5217	0.5006
	Precision	0.5003	0.5374	0.5003
	Recall	0.9051	0.3120	0.9051
	F1	0.6444	0.3948	0.6444
Identification Level	Accuracy	0.2149	0.4109	0.2074
	Precision	0.2696	0.2516	0.2607
	Recall	0.3337	0.0903	0.3189
	F1	0.2983	0.1329	0.2869

7. Discussion

We have made a few discoveries regarding the process of this experiment and the results obtained. First, manually constructed rules are more complicated than machine-generated rules. However, the accuracy of manually constructed rules does not necessarily exceed that of machine generated rules. Fairly reliable error detection rules can be obtained by establishing parameters based on automatically generated rules. Second, many automatically generated rules are not listed in manually constructed rules. This means the method of using machines to identify error detection rules is feasible. Considering these two perspectives, if the program has an enhanced ability to search for rules, then it is feasible to fully automatically identify grammatical errors made by Chinese as second language learners.

Several aspects of our proposed method can be further improved. First, rules in this study are primarily based on Chinese written error corpus. However, the corpus currently remains in the expansion phase. The increasingly rich content of the corpus can enhance the system performance. Second, only Redundant word and Missing word errors can be automatically generated by the current program. Also, the error detection rules contains only three terms. If more types of rules that are automatically generated by the program can be added in the program and the program can identify more complex rules, the system performance will be further improved.

Acknowledgements

This research is partially supported by the 'Aim for the Top University Project' and 'Center of Learning Technology for Chinese' of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, R.O.C. and the "International Research-Intensive Center of Excellence Program" of NTNU and Ministry of Science and Technology, Taiwan, R.O.C. under Grant no. NSC 103-2911-I-003-301

References

ACTFL Proficiency Guidelines 2012 - Writing. (2012). Retrieved August 25, 2014, from http://actflproficiencyguidelines2012.org/writing

Ashwell, T. (2000). Patterns of teacher response to student writing in a multi-draft composition classroom: Is content feedback followed by form feedback the best method? Journal of Second Language Writing, 9, 227–57.

Cheng, S. M., Yu, C. H., & Chen, H. H. (2014). Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Learners. *Proceedings of COLING 2014*, (pp. 279-289), Dublin, Ireland.

Chang, T. H., Sung, Y. T., & Lee, Y. T. (2012). A Chinese word segmentation and POS tagging system for readability research. *Proceedings of SCiP 2012*, Minneapolis, MN.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. Journal of Second Language Writing, 12, 267–96.

Chodorow, M., Dickinson, M., Israel, R., & Tetreault, J. R. (2012). Problems in Evaluating Grammatical Error Detection Systems. *Proceedings of COLING 2012* (pp. 611-628), Mumbai, India.

Connors, R. J., & Lunsford, A. A. (1988). Frequency of formal errors in current college writing, or ma and pa kettle do research. *College Composition and Communication, 39*(4), 395-409.

Dale, R., Anisimoff, I., & Narroway, G. (2012). HOO 2012: A report on the preposition and determiner errorcorrection shared task. *Proceedings of the 7th Workshop on Building Educational Applications Using NLP* (pp. 54-62), Montréal, Canada.

Dale, R., & Kilgarriff, A. (2011). Helping our own: The HOO 2011 Pilot Shared Task. *Proceedings of the 13th European Workshop on Natural Language Generation*, Nancy, France.

De Felice, R., & Pulman, S. 2009. Automatic detection of preposition errors in learner writing. *CALICO Journal, 26*(3), 512-528.

Donahue, S. (2001). Formal errors: Mainstream and ESL students. Presented at the 2001 Conference of the Two-Year College Association (TYCA); cited by Leacock et al. 2010.

Dulay, H. C., Burt, M. K., & Krashen, S. D. (1982). Language Two. New York: Oxford University Press.

Eeg-Olofsson, J., & Knuttson, O. (2003). Automatic grammar checking for second language learners: the use of prepositions. *Proceedings of the 14th Nordic Conference in Computational Linguistics*, Reykjavik, Iceland.

Fathman, A. K., & Whalley, E. (1990). Teacher response to student writing: Focus on form versus content. In B. Kroll (Ed.), Second language writing: Research insights for the classroom (pp. 178-190). Cambridge, UK: Cambridge University Press.

Ferris, D. & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? Journal of Second Language Writing, 10, 161–184.

James, C. (1998). Errors in Language Learning and Use: Exploring Error Analysis. London: Addison Wesley Longman.

Lee, L. H., Yu, L. C., Lee, K. C., Tseng, Y. H., Chang, L. P., & Chen, H. H. (2014). A Sentence Judgment System for Grammatical Error Detection. *Proceedings of COLING 2014* (pp. 67-70), Dublin, Ireland.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated Grammatical Error Detection for Language Learners. Morgan and Claypool Publishers.

Gamon, M., Leacock, C., Brockett, C., Dolan,W. B., Gao, J. F., Belenko, D., & Klementiev, A. (2009). Using Statistical Techniques and Web Search to Correct ESL Errors. *CALICO Journal, 26*(3), 491-511.

Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., & Tetreault. J. (2013). The conll-2013 shared task on grammatical error correction. *Proceedings of the 17th Conference on Computational Natural Language Learning*.

Nicholls, D. (2003) The Cambridge learner corpus: error coding and analysis for lexicography and ELT. *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 572-581), Lancaster, UK..

Tetreault, J., & Chodorow, M. (2009). Examining the use of region web counts for ESL error detection. *Proceedings of the Web as CorpusWorkshop (WAC-5),* San Sebastian, Spain.

Tetreault, J., & Chodorow, M. (2008). The ups and downs of preposition error detection in ESL writing. *Proceedings of the 22nd International Conference on Computational Linguistics* (pp. 865-872), Manchester, UK.

Wang, Z. (2011). A Study on the Teaching of Unique Syntactic Pattern in Modern Chinese for Native English-Speaking Students. Master Thesis. Northeast Normal University.

Yu, C. H., & Chen, H. H. (2012). Detecting word ordering errors in Chinese sentences for learning Chinese as a foreign language. *Proceedings of COLING 2012* (pp. 3003-3018), Bombay, India.