Computational Linguistics and Chinese Language Processing
Vol. 20, No. 1, June 2015, pp. 49-64

Automatically Detecting Syntactic Errors in Sentences Written by Learners of Chinese as a Foreign Language

Tao-Hsing CHANG[*], Yao-Ting SUNG[+] and Jia-Fei HONG[#]

Abstract

This paper proposed a method that can automatically detect syntax errors in Chinese sentences. The algorithm for identifying syntax errors proposed in this study is known as KNGED, which uses a large database of rules to identify whether syntax errors exist in a sentence. The rules were generated either manually or automatically. This paper further proposed an algorithm for identifying the type of error that a sentence contained. Experimental results shown that the false positive rate and F1-measure of the proposed method for detecting syntax errors in Chinese sentences are 0.90 and 0.65.

Keywords: Syntactic Errors, Chinese Grammar, Chinese Written Corpus.

1. Introduction

The teaching of languages has always been an important area of research and a commercially viable market. An important topic of research is the means by which the linguistic abilities of learners can be enhanced efficiently. This is especially so for learners of foreign languages, who have to learn the target language within a limited time period while being in a non-immersive learning environment, unlike the ample time they had for learning their native language. Contrastive linguistics is a tool that can be used to improve the efficiency of learning a foreign language effectively. Since most learners would already have well-developed capabilities in their native language, pointing out and analyzing the differences between the native and foreign language can help learners to understand the differences between the two, thereby facilitating the conversion from the former to the latter.

[*] Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, Taiwan
E-mail: changth@gm.kuas.edu.tw
[+] Department of Educational Psychology and Counseling, National Taiwan Normal University, Taiwan
E-mail: sungtc@ntnu.edu.tw
[#] Department of Applied Chinese Language and Culture, National Taiwan Normal University, Taiwan
E-mail: jiafeihong@ntnu.edu.tw

However, simply understanding the differences between two languages does not mean that a person can make the conversion from one to the other effectively in real-life usage situations. In comparative linguistics, two phenomena often appear in the patterns of language usage. First, since the types and quantity of differences are substantial, learners may not necessarily notice each and every difference between the native and foreign languages when using the latter. Second, when learners are not familiar with the linguistic differences, they become susceptible to the phenomenon of language transfer.

An example is the use of suffixes that signal tenses of English verbs, which has no parallel in the grammar of Chinese verbs. Although learners of English are aware that they need to pay attention to the tenses of verbs, they often make the mistake of using the wrong tense. Learners must keep practicing to become familiar with the relevant linguistic differences. During the learning process, teachers must also point out the errors committed. Only then can learners internalize the differences and gain the ability to use the foreign language. Unfortunately, the labor costs of making these corrections are high. In the existing educational model where one teacher is often responsible for teaching many students, it is not possible for him/her to conduct intensive practices for all students, nor correct the errors of each individual student.

To overcome this issue, many studies have proposed the concept of "automatic detection of learners' errors during language usage." These methods mainly employ detection models that target word or syntax errors. Many useful methods have already been proposed for the automatic detection and correction of English syntax errors. Some of these rely on having an excellent grammar parser. If the parser is unable to deconstruct a sentence completely and convert it to a parsing tree, then some syntax errors in this sentence will fail to be detected and corrected. However, it is difficult to apply such a concept to the issue of identifying Chinese syntax errors for two main reasons. First, it is difficult to identify the limits of a single sentence. For English sentences, the contents between two periods can be treated as a syntactic structure and unit of analysis. For Chinese sentences, a segment that ends with a comma can be a sentence with a complete syntactic and semantic structure, just a clause, or even a phrase. Second, the Chinese language contains many more syntactical changes, making it difficult for learners to distinguish between correct and erroneous usage. Hence, using a grammar parser for learning Chinese is not as effective as using one for learning English. These reasons make the detection and correction of errors in Chinese sentences more difficult than in those of English.

We believe that the identification of patterns in syntax errors is a possible solution. Common syntax errors usually involve part of a sentence rather than its overall structure. This situation is particularly pronounced for syntax errors committed by learners of a second language, the root cause of which is the phenomenon of language transfer. The following is an

example of an error that is often committed by Korean students when writing Chinese sentences.

Erroneous: "他來台北一年讀書了" (He has been in Taipei a year for studying.)
Correct: "他來台北讀書一年了" (He has been studying in Taipei for a year.)

In Korean, a temporal noun is always placed before the verb. As a result, many continue to do so when writing Chinese sentences, thus committing errors. If this and other commonly made errors can be compiled and sorted into general categories, further analysis can be done to determine the identification rules for each category of errors. If part of a sentence contains a grammatical structure that may be flagged by an identification rule, then that structure is likely to be erroneous. When sufficient identification rules have been compiled, a comparison of written sentences with the rules base will highlight those with syntax errors. Statistical methods can also be used to analyze the large number of sentences contained in learners' corpora to identify frequently occurring grammatical structures. The larger the corpus, the bigger the number of identification rules that can be generated, which in turn help to detect more errors.

The main aim of this paper is to propose a method that can automatically detect syntax errors in Chinese sentences and then state the type of error that has been committed. In terms of framework, this method employs learners' writing corpora as the basis and two methods to generate rules for identifying syntax errors. In the first method, linguistic experts generate rules by examining corpora through a system; the second method uses formulas to establish rules automatically through the application of statistical methods to corpora. After establishing the rules, we applied them to determine whether a sentence was erroneous. For erroneous sentences, we further proposed an algorithm for identifying the type of error that the sentence contained.

The organization of the rest of this paper is as follows: an analysis of related studies and their impact on our research motivation is done in Section 2; the corpora used in this study are listed in Section 3, with detailed explanations of a learners' corpus that has been specially created to identify erroneous sentences written by those for whom Chinese is a second language; manually identified rules created by this study are also introduced in the section, together with the method of using formulas to automatically establish identification rules; the proposed algorithm for automatic identification of erroneous sentences is also explained in the section; the effectiveness of the proposed approach is illustrated in Section 4; and Section 5 is the conclusion.

2. Related Works

Syntax errors are usually classified as belonging to either the category of "language form" or "surface structure." The former uses the language subsystems as the framework by which to classify the type of error. Specifically, this refers to errors in parts of speech (POS), syntax and semantics. The latter uses the structural method to classify the type of error, that is, by comparing the erroneous and correct forms. Surface structure errors are generally divided into four types: omissions, erroneous additions, overpresentations and misorders (Dulay, Burt, & Krashen, 1982; James, 1998).

Many analytical studies have been done on errors made by learners. One of the most famous English learners' corpora is the Cambridge Learner Corpus (CLC), with as many as 16 million words having been tagged as erroneous. The three most common types of errors include wrong selection of words, wrong prepositions, and wrong qualifiers (Nicholls, 2003). After 200 learners for whom English is a second language had taken writing ability tests, Donahue (2001) analyzed their performance and compared his findings with the linguistic errors made by native English speakers as proposed by Connors and Lunsfor (1988). Donahue found that the most common types of errors made by non-native versus native English speakers were different. For the former, these included mistakes in the use of commas or words, as well as omission of words.

In recent years, common syntax errors made by learners for whom Chinese is a second language have become a popular research topic. Wang (2011) indicates that for Chinese language learners who are native English speakers, the most common syntax errors include the omission of language elements, wrong word order, and structural errors. Cheng, Yu & Chen (2014) used the corpus of the Chinese Proficiency Test (HSK), which comprised 35,884 erroneous sentences in total, to analyze the types of syntax errors. The study found that the most common problems involved wrong word order, as well as omission of adverbial elements and predicates.

With the development of natural language processing technologies over the past decade, various researches have been done and tools for the automatic detection of English syntax errors have been proposed. The most common types of errors detected by these studies involve prepositions (Eeg-Olofsson & Knuttson, 2003; Tetreault & Chodorow, 2008; Gamon *et al*., 2009; De Felice & Pulman, 2009; Dale, Anisimoff, & Narroway, 2012; Ng *et al*., 2013), articles (Gamon *et al.*, 2009; Dale & Kilgarriff, 2011; Ng *et al*., 2013), and qualifiers (Dale *et al*., 2012; Ng *et al*., 2013).

These tools automatically detect errors in the learners' usage of qualifiers, articles, and prepositions, and then correct learners' grammatical errors. By using these tools, foreign language learners in mastering the correct grammar and are useful for the improvement of

writing skills (Chodorow *et al*., 2012; Leacock *et al*., 2010). However, there have been very few studies on leaners' corpora for the automatic detection of Chinese grammatical errors. Cheng *et al*. (2014) and Yu & Chen (2012) had used the Chinese sentences included in the HSK corpus for dynamic composition to develop detection techniques for errors in word order. For the method proposed by Lee *et al*. (2014), other than the HSK corpus for dynamic composition, the study had also included manual rules for common Chinese erroneous sentences when developing their system for detecting various errors in sentence construction and grammar.

Three conclusions can be derived from the aforementioned literature review. First, most studies have classified the types of syntax errors in terms of grammar or form, for example, omission of prepositions and redundancy of articles. Second, for the identification of errors, automatic detection methods make use of either manually established rules or statistical models. The identification results of the rule-based method detects some error types well, but most error types are such that this method does not capture them (Lee *et al*., 2013). On the other hand, the statistical approach requires a considerably large learners' corpus to be effective. Third, there are very few learners' corpora for Chinese learners, and methods involving the use of statistical models to generate rules for identifying errors are even rarer.

3. Method

The algorithm for identifying syntax errors proposed in this study is known as KNGED, which uses a large database of rules to identify whether syntax errors exist in a sentence. The rules were generated either manually or automatically, the details of which will be elaborated upon in Subsections 3.2 and 3.3 respectively. Data sets of erroneous sentences had to be used during the rule-generating process. This study made use of two such data sets to generate identification rules for syntax errors: (i) dry run data (hereinafter referred to as TEA1-DRY) from the Shared Task on Grammatical Error Diagnosis for Learning Chinese as a Foreign Language (hereinafter referred to as NLPTEA1-CFL), which was organized by the 1st Workshop on Natural Language Processing Techniques for Educational Applications; and (ii) the Chinese Written Corpus (CWC) that we had developed, which will be described in detail in the next subsection.

3.1 Chinese Written Corpus

The CWC comprises 1,147 essays divided into two data subsets, with a total of approximately 750,000 words. Within each data set are essays on the same topic written by different authors who are expatriates learning Chinese in one of 11 Chinese language center of 11 universities in Taiwan. This group of authors had very diverse linguistic backgrounds; the total number of different native languages in it was 37. The texts were collected and compiled between

September 2010 and June 2013. Each essay was graded by two trained raters using the criteria from the *Chinese Composition Scoring Standard* developed by Hsiung *et al*. (2014). These criteria reference the classification structure of ACTFL (2012) and are prescribed for rating Chinese essays written by expatriates for whom Chinese is a second language. Specifically, writing abilities are rated as "distinguished," "superior," "advanced," "intermediate," or "novice." The latter three grades are in turn subdivided into "high," "medium," and "low," yielding 11 levels in total.

Each Chinese sentence of every essay in the CWC had undergone tagging for segmentation and POS based on WECAn system (Chang, Sung, & Lee, 2012), followed by the correction of errors by trained taggers. Forty-eight POS tags were used, including the 46 simplified tags for Chinese POS as defined in CKIP (1993), the verb nominalization tag *Nv*, and the unknown POS tag *b*. Each sentence had been checked by the taggers for syntax errors. If found, the position and type of error were tagged accordingly, together with the corrected sentence. The main types of errors included erroneous additions/errors of redundancy, omissions, incorrect word order, and erroneous word selection.

3.2 Automatic Machine-generated Rules

The assumptions for our proposed method were based on two pieces of observed information. First, some of the erroneous positions and terms within a sentence are related to the preceding or subsequent word or POS. Second, most errors will occur repeatedly if the corpus is sufficiently large. Hence, the proposed method first examines all the possible patterns for syntax errors that can be generated by an erroneous sentence. Next, each pattern is individually checked to see if it appears in any other sentences within the corpus. A pattern is treated as a candidate rule if it occurs more than once. The following sentence is an example:

這些 地方 是 在 日本 (These places are located in Japan)
Neqa Na SHI P Nc

The tags below the sentence are the POS of each word. In the corpus, the "是" (are) character in the sentence was marked as being an error of the redundant type. Based on the aforementioned assumptions, all 32 possible combinations based on the word "是," its POS tag "SHI," and the preceding or subsequent word or POS tag are listed in Figure 1.

The symbol "+" in the figure indicates that the preceding/subsequent word/POS tag is immediately adjacent to the erroneous position, while the symbol ">" indicates that the preceding/subsequent word/POS tag is not immediately adjacent to the erroneous position. Each combination is treated as a candidate identification rule. The corrected pattern

corresponding to the combination is denoted as correction rule. For instance, the correction rule for candidate rule "Na + SHI + P" is "Na + P".

The 32 candidate rules can be subjected to a further conditional test. A recurring pattern *r* is an identification rule if the following conditions are met:

$$\text{FreqInErr}(r) \geq p \text{ and } \text{Reliability}(r) \geq k, \text{ where } \text{Reliability}(r) = \text{FreqInCol}(re)/\text{FreqInCor}(r)$$

FreqInErr(*r*) represents the number of times that rule *r* applies to the erroneous sentences which are identified by rule *r*. FreqInCor(*x*) represents the number of corrected sentences in the corpus that complies with the rule *r*. *re* represents the correction rule for rule *r*. Parameters *p* and *k* are thresholds obtained during the experiment.

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

Figure 1. Examples of machine-generated candidate identification rules

If the value of *p* is large, it indicates that more erroneous sentences contain the possible rule and hence, it should be included in the database of identification rules. In other words, the possible rule *r* should not be a random product that appears after the combinations have been listed. If the value of *k* is large, it indicates that a smaller ratio of false alarms will be generated when the possible rule *r* is used to identify erroneous sentences. In other words, the accuracy rate of identification will be higher. Using the 32 rules in Figure 1 as an example, if *p* and *k* are both set at 2, only 11 of the rules will be included as identification rules for errors

(please refer to Figure 2). When an identification rule for errors is included in the rules database, its corresponding correction rule will also be included.

(9) 這些 > SHI + 在
(10) 這些 > SHI + P
(12) 這些 > SHI > Nc
(14) 地方 + SHI + P
(25) Neqa > SHI + 在
(26) Neqa > SHI + P
(28) Neqa > SHI > Nc
(29) Na + SHI + 在
(30) Na + SHI + P
(31) Na + SHI > 日本
(32) Na + SHI > Nc

Figure 2. Rules from Fig. 1 that are added to the rule base after screening

Theoretically, the length of a rule extracted using this method need not be restricted to one preceding/subsequent word/POS tag. However, since there are many erroneous sentences, the possible rules that can be generated will be too numerous, making the computation process too time consuming. Therefore, in terms of the format of the rule, this study only considered the immediately preceding/subsequent word/POS tag. Given this premise, the automatic machine-generated method only generated rules for two types of errors: redundancy and omission. Moreover, these rules were produced based on CWC.

In addition, we observed that many examples of the selection type of error involved the wrong use of a unit, for example, "一個公車" (a bus) instead of "一輛公車." So, we compiled all the units that are used with each noun from the Sinica corpus (Chen, Huang, Chang, & Hsu, 1996). Since each noun can be matched with more than one type of unit, all units that can be used were included in the database of units. If one of the patterns "Neu + Nf + Na" or "Neu + Nf > DE + Na" appears in a sentence, the words corresponding to the two POS—Nf and Na—will be treated as the unit and designated noun respectively. The pair formed by the unit and designated noun of this pattern is then sent to the database of units for checking. If the pair has not appeared previously, it means that an error of the selection type has been detected. The correct pair of unit and designated noun is then treated as the rule for correction.

3.3 Manually-generated Rules

All manually-generated rules are established by linguistic experts through the following four steps. First, the experts observed the erroneous sentences in TEA1-DRY and then listed the candidate rules for identifying and correcting syntax errors. Next, they used an inspection program to analyze whether each syntactic rule is correct. The program would indicate the number of sentences that satisfy the three separate conditions stipulated in the CWC: (i) the number of erroneous sentences that complied with a rule identifying wrong syntax; (ii) the number of corrected sentences that complied with the rule for correction; and (iii) the number of corrected sentences that complied with the rule for identification.

An effective rule for identifying and correcting grammatical errors must generate as many results as possible under the first and second conditions, but as few results as possible under the third condition. If more sentences satisfy the first condition, it means that the rule can identify more of the erroneous sentences. On the other hand, if more sentences comply with the third condition, it means that the rules for error identification will wrongly treat more of the correct sentences as being erroneous. Hence, the smaller the number of sentences identified under the third condition, the better are the results. If many sentences satisfy the second condition, it means that the rules for correction are common and correct forms of usage, thus their general presence in the corpus. Consequently, the likelihood of the rules for correction being effective will also be higher.

The format of the manually-generated identification and correction rules is similar to the machine-generated rules, although there is no restriction on the number of preceding/subsequent words/POS. Hence, the former has a higher accuracy rate for detection. However, non-limitation on the number of preceding/subsequent words/POS also resulted in rules with sequential errors. Eight hundred and forty manually-generated identification rules were used in this study, which could be broken down into the following types: 90 missing, 73 redundant, 51 selection, and 626 wrong order. Since the proposed method for automatic machine-generated rules could not generate rules with disorder errors, the number for this type of manually generated rules far exceeded the other types.

3.4 Detection of Erroneous Sentences and Algorithm for Detected Types of Errors

After setting up the rule base generated by machine and manually, each test sentence was compared with the rules to determine if it was erroneous and if so, the type of error and rules for correction. Since one sentence could be simultaneously identified by multiple rules, we designed an algorithm shown in Figure 3 to identify the most likely error.

```
KNGED (sentence S, integer y)
Begin
    maximum = 0;
    rule-pointer = null;
    Tag the segmentation and POS of the sentence using WECAn;

    for every identification rule r_i for the selection error type in the rule base
        if sentence S contains any structure that can be identified by r_i
       then tag the erroneous portion of sentence S and return the corrected sentence;

    for every identification rule r_i for the disorder error type in the rule base
        if sentence S contains any structure that can be identified by r_i
        then tag the erroneous portion of sentence S and return the corrected sentence;

    for every identification rule r_i for the redundant and missing error types in the rule base
     {  if sentence S contains any structure that can be identified by r_i
       then if r_i is the redundant error type
              then  {
                    if Reliability(r_i) > maximum
                    then
                          maximum = Reliability(r_i);
                          rule-pointer = r_i;
              }
              else  if (Reliability(r_i) * y) > maximum
                     then {
                          maximum = Reliability(r_i) * y;
                          rule-pointer = r_i;
                     }
    }

    Tag the erroneous portion of sentence S with the rule identified by the rule-pointer and
        return the corrected sentence;
   return sentence S is the correct sentence;
End.
```

Figure 3. Proposed KNGED algorithm for the detection and correction of syntax errors

The methods for generating identification rules for different types of errors vary, and so does their effectiveness. We applied the various types of identification rules to the TEA1-DRY data set and then analyzed their effectiveness. We found that the identification rules for the selection type of errors had a much higher degree of accuracy compared to the rules for the other types of errors. This is because the identification of errors in the use of units is completely based on the vocabulary, resulting in a relatively lower rate of error. Thus, under the proposed algorithm, once a sentence has been identified as having an error of the selection type, that type of error would be ascribed to the sentence first. On the other hand, results for the wrong order type of error all arise from manually generated rules, hence the relatively lower rate of accuracy. Nevertheless, they are still more accurate than the identification rules for the redundant and missing types of errors. Thus, when a sentence is identified as having the wrong order type of error but not that of selection, it should first be ascribed the former type of error.

The value for sentences that have not been identified under the selection and wrong order types of errors but have been identified under the missing and redundant types is calculated based on the reliability value for each rule as shown in Formula (1). Compared to the rule for the omission error it is easier for the rule for the redundant type to achieve a higher value in terms of reliability. Hence, if a sentence complies with an identification rule for the redundant type of error and another for the omission type, the reliability value of the former must be several times greater than that for the latter (*i.e.*, the *y* value of the algorithm). It is only in this situation that the identification results for the redundant type of errors are adopted. Otherwise, the sentence should be treated as the omission type of errors.

4. Experimental Results

The formal run data provided by NLPTEA1-CFL (Yu, Lee, & Chang, 2014) was used to evaluate the effectiveness of the proposed method. The data consist of 1,750 sentences. A half of these sentences have no grammatical errors while each of the remainder only contain one grammatical error. The number of sentences with error type redundant, missing, disorder, and selection is 279, 350, 120, and 126, respectively. Three indicators for evaluating the performance of our proposed method are defined as follows:

$$\text{Precision} = TP/(TP + FP)$$

$$\text{Recall} = TP/(TP + FN)$$

$$F1 = 2 * \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall})$$

where TP refers to the number of sentences for which the error type was correctly detected, FP

refers to the number of sentences with no errors that were nevertheless identified as erroneous, and FN refers to the number of sentences with errors that were not detected or detected but ascribed the incorrect error type. Since the assessment facets for recall and precision are different, the F1-measure was used as the overall indicator of assessment effectiveness. In NLPTEA1-CFL, the evaluation is divided into detection level and identification level. In detection level, the proposed method only grouped test sentences into two types: correct or incorrect. In identification level, the proposed method should clearly identifies test sentences to be one of four error types: Redundant, Missing, Disorder, and Selection.

The performance of KNGED based on the three assessment indicators is shown in Table 1. Since the performance of KNGED is affected by the parameter settings, Table 1 also shows the calculation results for KNGED's effectiveness under various parameters settings. When the parameter settings for KNGED-1 were p = 1, k = 2, y = 50, the number of rules generated for the redundant and omission types of errors was 53,834 and 3,781, respectively. When the parameter settings for KNGED-2 were p = 1, k ≈∞ (i.e. *FreqInCor(r)*=0), y = 50, the numbers of rules generated for the same two types of errors were 10,114 and 145. The parameter settings for KNGED-3 were p = 1, k = 2, y = 1. Because the parameter p and k of KNGED-3 were the same as of for KNGED-1, the numbers of rules generated for the same two types of errors were also 53,834 and 3,781 respectively.

Table 1. Comparison of results for different parameter settings for the previous experiment

Submission		KNGED-1	KNGED-2	KNGED-3
False positive rate		0.9040	0.2686	0.9040
Detection Level	Precision	0.5015	0.5164	0.5015
	Recall	0.9326	0.2880	0.9326
	F1	0.6523	0.3698	0.6523
Identification Level	Precision	0.2600	0.2555	0.2505
	Recall	0.3257	0.0926	0.3097
	F1	0.2892	0.1359	0.2770

In detection level, the F1-measure values of KNGED-1 and KNGED-3 were the highest and far exceeded the effectiveness of KNGED-2. The main reason is because the parameter settings of KNGED-2 resulted in only few rules in the rule base, causing the recall to decrease. It can thus be seen that the setting of parameter values have considerable impact on effectiveness. In addition, the performance of three parameter settings of KNGED do not perform well in identification level. The main reason is the inclusion of many invalid rules in

the rules database. It causes the accuracy to decrease.

A comparison between the effectiveness of manually-generated identification rules and machine-generated rules under KNGED-1 is shown in Table 2. In KNGED-1, the machine-generated rules do not contain the disorder type of errors, whereas the numerical variations between the various types of errors for manually-generated identification rules are large. Thus, we cannot deduce arbitrarily which method was better. However, it can be seen from Table 2 that it is insufficient to only employ manually-generated rules to identify grammatical errors. On the other hand, Table 2 also shows that the machine-generated rules of KNGED-1 are effective even all rules are simple bi-gram or tri-gram patterns.

Table 2. Comparison of effectiveness between manually-generated rules and machine-generated rules under KNGED-1

Rules		Manually-generated	Machine-generated
Detection Level	Precision	0.5217	0.5019
	Recall	0.3978	0.9399
	F1	0.4514	0.6543
Identification Level	Precision	0.1429	0.2697
	Recall	0.0608	0.3445
	F1	0.0853	0.3025

Since the information in NLPTEA1-CFL includes the language proficiency level for each sentence, we tested the effectiveness of KNGED-1 at detecting syntax mistakes by authors at different proficiency levels. The results are shown in Table 3. The language proficiency levels were in line with the grading standards of the Common European Framework of Reference for Languages (CEFR). The A1 and C2 grade represents the lowest and highest level of proficiency. It can be seen that the KNGED-1 for identifying erroneous sentences by writers with poor capabilities were more effective than that with good proficiency. This may be because for the writers with good proficiency, the erroneous structures that they make and the related causes are more complex, such that it was inadequate to use simple rules for identification.

Table 3. KNGED-1 identification results of erroneous sentences produced by writers of different CEFR linguistic proficiency levels

Level of CEFR		A2	B1	B2	C1
Detection Level	Precision	0.5104	0.5005	0.4971	0.5263
	Recall	0.9111	0.9342	0.9399	1.0000
	F1	0.6543	0.6518	0.6503	0.6897
Identification Level	Precision	0.2849	0.2683	0.2162	0.2500
	Recall	0.3481	0.3419	0.2623	0.3000
	F1	0.3133	0.3006	0.2370	0.2727

5. Conclusion and Future Work

We made several discoveries based on the processes and results of this experiment. First, although manually-generated rules are more complex than those generated automatically using formulas, their accuracy rates are not necessarily higher. Through manipulation of parameter settings, automatic generation can actually result in more reliable identification rules. Second, automatic generation leads to many rules that have not been manually proposed. This means that the use of machines to determine identification rules is a feasible method. Integrating these two points of view, if the effectiveness of search rules in programs can be significantly enhanced, then it is actually feasible to have a fully automatic system to identify syntax errors by writers for whom Chinese is a second language.

There are several areas in which the proposed method can be further improved. First, the contents of the CWC were the main basis for establishing the rules. Currently, this corpus is still at the expansion phase. As the contents become increasingly enriched, the effectiveness of the system should improve correspondingly. Second, for automatic machine-generated rules, only the immediately preceding/subsequent words/POS are currently considered for rules to identify the redundant and missing types of errors. If the effectiveness of screening the possible rules can be improved, more precise rules will be generated, thereby further enhancing the system's performance.

Third, the heuristic algorithm that we have proposed is unable to handle the issue of one sentence having multiple errors. In terms of practical application, it is very important to develop an algorithm that is able to identify sentences with multiple syntax errors. Fourth, many selection and word order types of syntax errors are related to context rather than syntactic hierarchy. The proposed method has already included the generation of identification rules for erroneous usage of units, which is context-related. Subsequently, further in-depth analysis can be made for other patterns of errors under this category. This will facilitate the

extraction of methods to generate identification rules for errors that are based on or related to context.

Acknowledgement

This work is supported in part by the Ministry of Science and Technology, Taiwan, R.O.C. under the Grants MOST 103-2511-S-151-001. It is also 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 MOST 104-2911-I-003-301.

References

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

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

Chen, K. J., Huang, C. R., Chang, L. P., & Hsu, H. L. (1996). Sinica corpus: Design methodology for balanced corpora. *Language*, 167-176.

Cheng, S. M., Yu, C. H., & Chen, H. H. (2014). Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Learners. In *Proceedings of COLING 201*4, 279-289, Dublin, Ireland.

Chodorow, M., Dickinson, M., Israel, R., & Tetreault, J. R. (2012). Problems in Evaluating Grammatical Error Detection Systems. In *Proceedings of COLING 201,* 611-628, Mumbai, India.

CKIP, (1993). *Analysis of Syntactic Categories for Chinese*. CKIP Tech. Report#93-05, Sinica, Taipei,

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. In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP,* 54-62, Montréal, Canada.

Dale, R., & Kilgarriff, A. (2011). Helping our own: The HOO 2011 Pilot Shared Task. In *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. In *Proceedings of the 14th Nordic Conference in Computational Linguistics*, Reykjavik, Iceland.

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.

Hsiung, Y. W., Lee, H. H. & Sung, Y. T. (2014). Examining the ACTFL writing assessment rating scale for L2 Chinese learners, *Journal of Chinese Language Teaching*, *11*(4), 105-133.

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

Lee, L. H., Chang, L. P., Lee, K. C., Tseng, Y. H.,&d Chen, H. H. (2013). Linguistic Rules Based Chinese Error Detection for Second Language Learning. In *Proceedings of ICCE 2013*, 27-29.

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. In *Proceedings of COLING 2014,* 67-70, Dublin, Ireland.

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

Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., & Tetreault. J. (2013). The conll-2013 shared task on grammatical error correction. In *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. In *Proceedings of the Corpus Linguistics 2003 Conference,* 572-581, Lancaster, UK.

Tetreault, J., & Chodorow, M. (2008). The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics*, 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. In *Proceedings of COLING 2012,* 3003-3018, Bombay, India.

Yu, L. C, Lee, L. H., & Chang, L. P. (2014). Overview of Grammatical Error Diagnosis for Learning Chinese as a Foreign Language. In *Proceedings of ICCE 2014,* 42-47, Nara, Japan.