rule-based + LM
只判断对错，不能修正
需要linguist rules

A Sentence Judgment System for Grammatical Error Detection

Lung-Hao Lee [1,2], Liang-Chih Yu[3,4], Kuei-Ching Lee[1,2],
Yuen-Hsien Tseng[1], Li-Ping Chang[5], Hsin-Hsi Chen[2]
[1]Information Technology Center, National Taiwan Normal University
[2]Dept. of Computer Science and Information Engineering, National Taiwan University
[3]Dept. of Information Management, Yuen Ze University
[4]Innovation Center for Big Data and Digital Convergence, Yuen Ze University
[5]Mandarin Training Center, National Taiwan Normal University
lcyu@saturn.yzu.edu.tw, {lhlee, johnlee, lchang,
samtseng}@ntnu.edu.tw, hhchen@ntu.edu.tw

Abstract

This study develops a sentence judgment system using both rule-based and n-gram statistical methods to detect grammatical errors in Chinese sentences. The rule-based method provides 142 rules developed by linguistic experts to identify potential rule violations in input sentences. The n-gram statistical method relies on the n-gram scores of both correct and incorrect training sentences to determine the correctness of the input sentences, providing learners with improved understanding of linguistic rules and n-gram frequencies.

1 Introduction

China's growing global influence has prompted a surge of interest in learning Chinese as a foreign language (CFL), and this trend is expected to continue. This has driven an increase in demand for automated IT-based tools designed to assist CFL learners in mastering the language, including so-called MOOCs (Massive Open Online Courses) which allows huge numbers of learners to simultaneously access instructional opportunities and resources. This, in turn, has driven demand for automatic proofreading techniques to help instructors review and respond to the large volume of assignments and tests submitted by enrolled learners.

However, whereas many computer-assisted learning tools have been developed for use by students of English as a Foreign Language (EFL), support for CFL learners is relatively sparse, especially in terms of tools designed to automatically detect and correct Chinese grammatical errors. For example, while Microsoft Word has integrated robust English spelling and grammar checking functions for years, such tools for Chinese are still quite primitive. In contrast to the plethora of research related to EFL learning, relatively few studies have focused on grammar checking for CFL learners. Wu et al. (2010) proposed relative position and parse template language models to detect Chinese errors written by US learner. Yu and Chen (2012) proposed a classifier to detect word-ordering errors in Chinese sentences from the HSK dynamic composition corpus. Chang et al. (2012) proposed a penalized probabilistic First-Order Inductive Learning (pFOIL) algorithm for error diagnosis. In summary, although there are many approaches and tools to help EFL learners, the research problem described above for CFL learning is still under-explored. In addition, no common platform is available to compare different approaches and to promote the study of this important issue.

This study develops a sentence judgment system using both rule-based and *n*-gram statistical methods to detect grammatical errors in sentences written by CFL learners. Learners can input Chinese sentences into the proposed system to check for possible grammatical errors. The rule-based method uses a set of rules developed by linguistic experts to identify potential rule violations in input sentences.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, pages 67–70, Dublin, Ireland, August 23-29 2014.

The *n*-gram statistical method relies on the *n*-gram scores of both correct and incorrect training sentences to determine the correctness of the input sentences. The system helps learners develop an improved understanding of both linguistic rules and *n*-gram frequencies. In addition, the proposed system can also be incorporated into online CFL MOOC platforms to help assess and/or score the numbers of assignments and tests.

2 A Sentence Judgement System

Figure 1 shows the user interface of the sentence judgment system, which can be accessed at http://sjf.itc.ntnu.edu.tw/demo/. Learners can submit single or multiple sentences through the textbox shown in the upper part of Fig. 1. Each input sentence is pre-processed for word segmentation and part-of-speech tagging, and then passed to both the rule-based and *n*-gram statistical methods for grammatical error detection. Finally, an input sentence will be marked as incorrect ( ) if both methods detect grammatical errors. Otherwise, it will be marked as correct ( ). In addition to the decision information, the explanation of the matched rules and *n*-gram frequencies are also presented for reference, as shown in the bottom part of Fig. 1. For instance, the following sentence is marked as incorrect:

我 從 這裡 走 往 北
(I from here go towads north.)

The rule-based method shows a rule violation is detected and explains that a preposition (e.g., "往" (towards)) cannot be used after a verb (e.g., "走" (go)). The *n*-gram frequencies also shows that the the frequency of the bigram "走 往" (go towards) is relative low. The following subsections describe in detail the pre-processing, rule-based method, and *n*-gram statistical method.

2.1 Pre-processing

Chinese is written without word boundaries. As a result, prior to the implementation of most Natural Language Processing (NLP) tasks, texts must undergo automatic word segmentation. Automatic Chinese word segmenters are generally trained by an input lexicon and probability models. However, it usually suffers from the unknown word (i.e., the out-of-vocabulary, or OOV) problem. In this study, a corpus-based learning method is used to merge unknown words to tackle the OOV problem (Chen and Ma, 2002). This is followed by a reliable and cost-effective POS-tagging method to label the segmented words with parts-of-speech (Tsai and Chen, 2004). For example, take the Chinese sentence "歐巴馬是美國總統" (Obama is the president of the USA). It was segmented and tagged in the form of "POS:Word" sequence shown as follows: Nb:歐巴馬 SHI:是 Nc:美國 Na:總統. Among these words, the translation of a foreign proper name "歐巴馬" (Obama) is not likely to be included in a lexicon and therefore is extracted by the unknown word detection mechanism. In this case, the special POS tag 'SHI' is a tag to represent the be-verb "是".

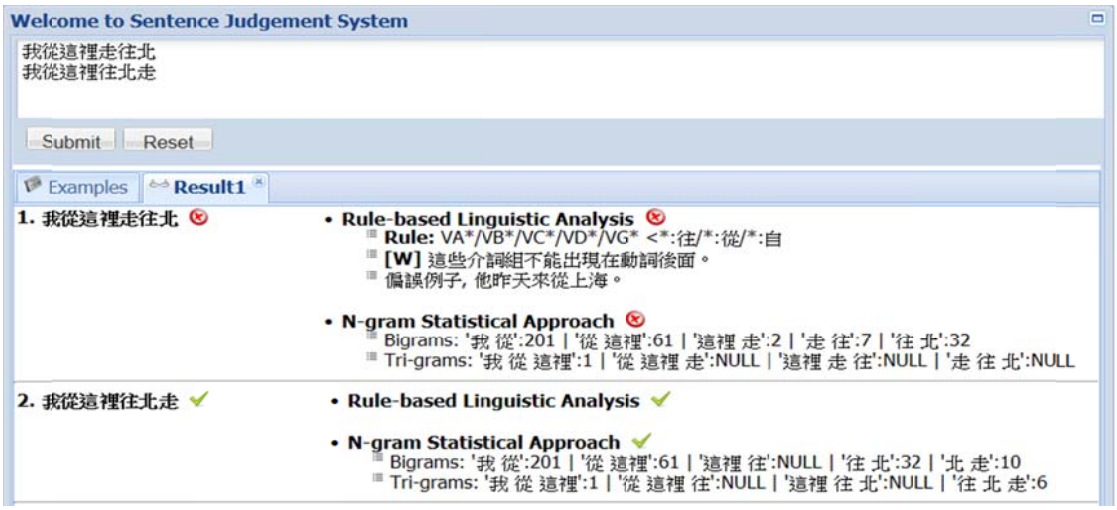

Figure 1. Screenshot of the sentence judgement system.

2.2 Rule-based Linguistic Analysis

Several symbols are used to represent the syntactic rules to facilitate the detection of errors embedded in Chinese sentences written by CFL learners: (1) "*" is a wild card, with "Nh*" denoting all subordinate tags of "Nh", *e.g.*, "Nhaa," "Nhab," "Nhac," "Nhb," and "Nhc". (2) "-" means an exclusion from the previous representation, with "N*-Nab-Nbc" indicating that the corresponding word should be any noun (N*) excluding countable entity nouns (Nab) and surnames (Nbc). (3) "/" means an alternative (i.e., "or"), where the expression "一些/這些/那些" (some/these/those) indicates that one of these three words satisfies the rule. (4) The rule mx{W1 W2} denotes the mutual exclusivity of the two words W1 and W2. (5) "<" denotes the follow-by condition, where the expression "Nhb < Nep" means the POS-tag "Nep" follows the tag "Nhb" that can exist several words ahead of the "Nep".

Using such rule symbols, we manually constructed syntactic rules to cover errors that frequently occur in sentences written by CFL learners. We adopted the "Analysis of 900 Common Erroneous Samples of Chinese Sentences" (Cheng, 1997) as the development set to handcraft the linguistic rules with syntactic information. If an input sentence satisfies any syntactic rule, the system will report the input as suspected of containing grammatical errors, creating a useful tool for autonomous CFL learners.

2.3 N-gram Statistical Analysis

Language modeling approaches to grammatical error detection are usually based on a score (log probability) output by an *n*-gram model trained on a large corpus. A sentence with grammatical errors usually has a low *n*-gram score. However, choosing an appropriate threshold to determine whether a sentence is correct is still a nontrivial task. Therefore, this study proposes the use of *n*-gram scores of correct and incorrect sentences to build the respective correct and incorrect statistical models for grammatical error detection. That is, a given sentence is denoted as incorrect (i.e., having grammatical errors) if its probability score output by the statistical model of incorrect sentences (i.e., the incorrect model) is greater than that of correct sentences (i.e., the correct model).

To build the incorrect and correct statistical models, a total of 19,080 sentences with grammatical errors were extracted from the HSK dynamic composition corpus. These sentences were then manually corrected. An *n*-gram (*n*= 2 and 3) language model was then built from the Sinica corpus released by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) using the SRILM toolkit (Stolcke, 2002). The trained language model was used to assign an *n*–gram score for each correct and incorrect sentence, which were then used to build the respective correct and incorrect models based on a normal probability density function (Manning and Schütze, 1999). Both models can then be used to evaluate each test sentence by transforming its *n*-gram score into a probability score to determine whether the sentence is correct or not.

3 Performance Evaluation

The test set included 880 sentences with grammatical errors generated by CSL learners in the NCKU Chinese Language Center, and the corresponding 880 manually corrected sentences. For the rule-based approach, a total of 142 rules were developed to identify incorrect sentences. For the n-gram statistical approach, both bi-gram and tri-gram language models were used for the correct and incorrect statistical models. In addition to precision, recall, and F1, the false positive rate (FPR) was defined as the number of correct sentences incorrectly identified as incorrect sentences divided by the total number of correct sentences in the test set.

Table 1 shows the comparative results of the rule-based and *n*-gram statistical approaches to grammatical error detection. The results show that the rule-based approach achieved high precision, low recall and low FPR. Conversely, the *n*-gram-based approach yielded low precision, high recall and high FPR. In addition, the tri-gram model outperformed the bi-gram model for all metrics. Given the different results yielded by the rule-based and *n*-gram statistical approaches, we present different combinations of these two methods for comparison. The "OR" combination means that a given sentence is identified as incorrect by only one of the methods, while the "AND" combination means that a given sentence is identified as incorrect by both methods. The results show that the "OR" combination yielded better recall than the individual methods, and the "AND" combination yielded better precision and FPR than the individual methods. Thus, the choice of methods may depend on application requirements or preferences

Method	Precision	Recall	F1	False Positive Rate
Rule	0.857	0.224	0.356	0.038
2-gram	0.555	0.751	0.638	0.603
3-gram	0.585	0.838	0.689	0.595
Rule OR 2-gram	0.500	1.000	0.667	1.000
Rule OR 3-gram	0.502	1.000	0.668	0.993
Rule AND 2-gram	0.924	0.083	0.153	0.007
Rule AND 3-gram	0.924	0.083	0.153	0.007

Table 1. Comparative results of the rule-based and *n*-gram statistical approaches.

Many learner corpora exist for EFL for use in machine learning, including the International Corpus of Learner English (ICLE) and Cambridge Learner Corpus (CLC). But collecting a representative sample of authentic errors from CFL learners poses a challenge. In addition, English and Chinese grammars are markedly different. In contrast to syntax-oriented English language, Chinese is discourse-oriented, with meaning often expressed in several clauses to make a complete sentence. These characteristics make syntactic parsing difficult, due to long dependency between words in a clause or across clauses in a sentence. These difficulties constrain system performance.

4 Conclusions

This study presents a sentence judgment system developed using both rule-based and *n*-gram statistical methods to detect grammatical errors in sentences written by CFL learners. The system not only alerts learners to potential grammatical errors in their input sentences, but also helps them learn about linguistic rules and *n*-gram frequencies. The major contributions of this work include: (a) demonstratingg the feasibility of detecting grammatical errors in sentences written by CFL learners, (b) developing a system to facilitate autonomous learning among CFL learners and (c) collecting real grammatical errors from CFL learners for the construction of a Chinese learner corpus.

Acknowledgments

This research was partially supported by Ministry of Science and Technology, Taiwan under the grant NSC102-2221-E-155-029-MY3, NSC 102-2221-E-002-103-MY3, and the "Aim for the Top University Project" sponsored by the Ministry of Education, Taiwan.

Reference

Andreas Stolcke. 2002. SRILM − An extensible language modeling toolkit. *Proceedings of ICSLP'02*, pages 901-904.

Chi-Hsin Yu and Hsin-Hsi Chen. 2012. Detecting word ordering errors in Chinese sentences for learning Chinese as a foreign language. *Proceedings of COLING'12*, pages 3003-3018.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA.

Chung-Hsien Wu, Chao-Hung Liu, Matthew Harris and Liang-Chih Yu. 2010. Sentence correction incorporating relative position and parse template language model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1170-1181.

Keh-Jiann Chen and Wei-Yun Ma. 2002. Unknown word extraction for Chinese documents. *Proceedings of COLING'02*, pages 169-175.

M. Cheng. 1997. Analysis of 900 Common Erroneous Samples of Chinese Sentences - for Chinese Learners from English Speaking Countries (in Chinese). Beijing, CN: Sinolingua.

Ru-Ying Chang, Chung-Hsien Wu, and Philips K. Prasetyo. 2012. Error diagnosis of Chinese sentences using inductive learning algorithm and decomposition-based testing mechanism. *ACM Transactions on Asian Language Information Processing*, 11(1):Article 3.

Yu-Fang Tsai and Keh-Jiann Chen. 2004. Reliable and cost-effective pos-tagging. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(1):83-96.