Choy, D. et al. (Eds.) (2013). Work-in-Progress Poster Proceedings of the 21st International Conference on Computers in Education. Indonesia: Asia-Pacific Society for Computers in Education

Linguistic Rules Based Chinese Error Detection for Second Language Learning

Lung-Hao LEE[a,c], Li-Ping CHANG[b]
Kuei-Ching LEE[a,c], Yuen-Hsien TSENG[a*], and Hsin-Hsi CHEN[c]
[a] *Information Technology Center, National Taiwan Normal University, Taiwan*
[b] *Mandarin Training Center, National Taiwan Normal University, Taiwan*
[c] *Dept. of Computer Science and Information Engineering, National Taiwan University, Taiwan*
*samtseng@ntnu.edu.tw

Abstract: In this paper, we handcraft a set of linguistic rules with syntactic information to detect errors occurred in Chinese sentences written by SLL. Experimental results come the similar conclusions with well-known ALEK system used by ETS for English Learning. Our developed Chinese sentence error detection system will be helpful for Chinese self-learners.

Keywords: Computer-aided language learning, second language learning, computer education

10. Introduction

Second Language Learners (SLL) usually write ungrammatical sentences with various types of errors. SLL tends to make mistakes in writing Chinese sentences in their early stage of learning Chinese. For example, the learner may like to express: “這對夫妻很恩愛” (The couple is very affectionate to each other), where the “恩愛” (affectionate) was mistakenly written as another similar word “恩情” (kind) as observed in the learner’s corpora. Error detection systems that indicate different kinds of errors embedded in a given sentence are important and invaluable to SLL for self-learning.

Assessing LExical Knowledge (ALEK) system (Chodorow and Leacock, 2000) adopted statistical analysis to detect the errors of an English sentence. Using 20 target words from the Test of English as a Foreign Language (TOEFL), it performed with about 80% precision and 20% recall. Izumi et al. (2003) detected English grammatical and lexical errors made by Japanese learners. Recently, relative position and parse template language models were proposed to detect various types of Chinese errors written by US learners (Wu et al. 2010). Different from most of the previous studies, which have focused on corpus-based statistical methods, we attempt to develop a rule-based system to detect the common errors embedded in Chinese sentences written by SLL.

In this work, we manually construct a set of linguistic rules with syntactic information to detect erroneous sentences that were frequently written by the SLL. If a sentence satisfies at least one syntactic-rule, the developed system will regard the input sentence as erroneous and response with suggestions to indicate the possible errors.

11. Linguistic Rules Based Chinese Error Detection

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 to merge the unknown words as described in Chen and Ma (2002) is adopted to tackle the OOV problem. This is followed by a reliable and cost-effective POS-tagging method to label the segmented words with part-of-speeches similar to the approach proposed by Tsai and Chen (2004). Take the Chinese sentence “歐巴馬是美國總統” (Obama is the president of USA) for instance. 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 method. In this case, the special POS tag ‘SHI’ is a tag to represent the be-verb “是”. The complete set of part-of-speech tags is defined in the technical report by CKIP (2003).

To represent the syntactic rules for employing them easily to detect errors embedded in Chinese sentences written by SLL, several symbols are defined. Some of them are explained as follows: 1) The symbol “*” means a wild card. For example, the whole subordinate tags of “Nh”, *i.e.*, “Nhaa,” “Nhab,” “Nhac,” “Nhb,” and “Nhc”, can be denoted as “Nh*”. 2) The symbol “-” means an exclusion from the previous representation. Take this expression “N*-Nab-Nbc” as an example, it denotes that the corresponding word should be any noun (N*) excluding countable entity nouns (Nab) and surnames (Nbc). 3) The symbol “/” means alternative (the “or” situation). The expression “一些/這些/那些” (some/these/those) represents that one of these three words satisfies the rule. 4) The rule mx{W1 W2} denotes that the two words W1 and W2 should not co-exist (should be mutual exclusive). 5) The symbol “<” denotes the follow-by condition. For instance, this expression “Nhb < Nep” means the POS-tag “Nep” follows the tag “Nhb” that can exist several words ahead of the “Nep”.

With the rule symbols like the above, we manually construct syntactic rules to cover frequent errors occurred in Chinese sentences written by SLL. We adopt the “Analysis of 900 Common Erroneous Samples of Chinese Sentences” (Cheng, 1997) as the development set to handcraft the linguistic rules with syntactic information. Based on these samples compiled by Chinese teachers in Beijing, we constructed 60 syntactic rules to detect errors in the samples. Table 1 shows some rules accompanying with their example sentences. If an input sentence satisfies any syntactic rule, our developed system will report the input as an erroneous sentence. This can be helpful to SLL for self-learning of Chinese.

Table 1: Some developed syntactic rules and their detected erroneous sentences.

Rule	Dfa N*-Nb*-Nc*/A/VA*/VB*/VC*/VD*/VE*/VF*/V_12
Example	Nhaa:她 VK1:覺得 Nhab:自己 DE:的 Nab:丈夫 Dfa:很 Nhab:私人 (She feels that her husband is very private)
Notes	“私人” (private) is an improper word in this sentence. The correct word should be 自私 (selfish). So the correct sentence is “她覺得自己的丈夫很自私”.
Rule	mx{Dbab:可以 Dbab:能}
Example	VH11:舊 DE:的 Nab:雜誌 Dbab:可以 Dbab:能 VD:借 Td:嗎 (Can old magazines be borrowed?)
Notes	“能” (able) is a redundant word in this sentence. This word cannot be collocated with another word “可以” (can). The correct sentence is “舊的雜誌能借嗎”.
Rule	*:把/*:被 < Da*/Db*:-來-去/Dc/Dd
Example	Nab:自行車 P02:被 Nb:丁力 Dc:沒 VC:騎走 (The bicycle is not ridden by Dingli)
Notes	The word “沒” (not) is put in a wrong position. This sentence contains a word ordering error. The correct sentence is “自行車沒被丁力騎走”.

12. Experiments and Performance Evaluation

The test data comes from a set of real error sentences written by learners of Chinese as a second language at National Cheng Kung University in Taiwan (Wu et al., 2010). Each erroneous sentence (positive instance) is accompanied with a correct one (negative instance) in this test set. In total, there are 1,866 pairs of sentences collected in years around 2009.

Table 2 shows the confusion matrix of our approach. The results indicate that our linguistic rules for error detection achieved an accuracy of 58.47%=(418+1764)/(1866+1866), while maintaining a promising precision of 80.38%=418/(418+102), and a recall of 22.4%=418/(418+1448). The performance level is similar with that of the ALEK system used by Educational Testing Service (ETS) for erroneous English sentence detection (Chodorow and Leacock, 2000). In addition, maintaining low false-alarm rate (which is the ratio of correct sentences that are detected as erroneous ones) is important for a system to be practical. In the experiments, our approach achieved a false-alarm rate of 5.47% (among 1,866 correct sentences, 102 were detected as erroneous). This shows that our approach is feasible to detect errors while not causing much trouble to the users.

Table 2: Confusion matrix using our linguistic rule based detection.

Confusion Matrix		Gold Standard	
		Positive	Negative
Detected Results	Positive	418	102
	Negative	1448	1764

13. Conclusions and Future Work

This paper proposes a linguistic rule based Chinese error detection approach. The syntactic rules handcrafted based on a smaller development set achieve promising performance on a totally different and larger test set, while maintaining a favorably low false-alarm rate. The major contributions of this work include: 1) indicating the usefulness of common error samples manually analyzed/collected by Chinese teachers in previous work; 2) demonstrating the feasibility of linguistic rules handcrafted from these samples; and 3) developing a system to help self-learning of Chinese for SLL.

This work is our first exploration to automatically detect Chinese erroneous sentences. The research result can be extended to automatic essay evaluation, which is especially useful for Massive Open Online Courses (MOOC), because manually evaluating a large scale of Chinese writing homework and exams is a very challenging issue.

Acknowledgements

This research was partially supported by National Science Council (NSC), Taiwan, under the grant NSC102-2221-E-002-103-MY3, and the “Aim for the Top University Project” of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan. We are also grateful to the support of International Research-Intensive Center of Excellence Program of NTNU and NSC, Taiwan, under the grant NSC102-2911-I-003-301.

References

Chen, K.-J., & Ma, W.-Y. (2002). Unknown word extraction for Chinese documents. *Proceedings of COLING’02* (pp. 169-175). Taipei, Taiwan: ACL Press.

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

Chinese Knowledge Information Processing (CKIP) Group. (1993). Categorical analysis of Chinese. *ACLCLP Technical Report #* 93-05, Academia Sinica. Available online at: http://rocling.iis.sinica.edu.tw/CKIP/tr/9305_2013%20revision.pdf

Chodorow, M., & Leacock, C. (2000). An unsupervised method for detecting grammatical errors. *Proceedings of NAACL’00* (pp. 140-147). Seattle, Washington: ACL Press.

Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., & Isahara, H. (2003). Automatic error detection in the Japanese learner’s English spoken data. *Proceedings of ACL’03* (pp. 145-148). Sapporo, Japan: ACL Press.

Tsai, Y.-F., & Keh-Jiann Chen, K.-J. (2004). Reliable and cost-effective pos-tagging. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(1), 83-96.

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