Detecting English Grammatical Errors based on Machine Translation

張竟 Jim Chang

吳鑑城 Jiancheng Wu

張俊盛 Jason S. Chang

國立清華大學資訊工程系

Department of Computer Science

National Tsing Hua University

{jim.chang.nthu@gmail.com; wujc86@gmail.com; jason.jschang@gmail.com}

Keywords: grammatical error correction, serial errors, machine translation, n-grams

Many people are learning English as a second or foreign language, and there are estimated 375 million English as a Second Language (ESL) and 750 million English as a Foreign Language (EFL) learners around the world according to Graddol (2006). Evidently, automatic grammar checkers are much needed to help learners improve their writing. However, typical English proofreading tools do not target specifically the most common errors made by second language learners. The grammar checkers available in popular word processors have been developed with a focus on native speaker errors such as subject-verb agreement and pronoun reference. Therefore, these word processors (e.g., Microsoft Word) often offer little or no help with common errors causing problems for English learners.

Grammatical Error Detection (GED) for language learners has been an area of active research. GED involves pinpointing some words in a given sentence as grammatically erroneous and possibly offering correction. Common errors in learners' writing include missing, unnecessary, and misuse of articles, prepositions, noun number, and verb form. Recently, the state-of-the-art research on GED has been surveyed by Leacock et al. (2010). In our work, we address serial errors in English learners' writing related to the proposition and verb form, an aspect that has not been dealt with in most GED research. We also consider the issues of broadening the training data for better coverage, and coping with data sparseness when unseen events happen.

Researchers have looked at grammatical errors related to the most common prepositions (e.g., De Felice and Pulman, 2007; Gamon 2010). In the research area of detecting verb form errors, methods based on template related to parse tree, maximum entropy with lexical and POS features have been proposed (e.g., Lee and Seneff, 2006; Izumi et al. 2003).

The Longman Dictionary of Common Errors, second edition (LDOCE) is the result of analyzing errors encoded in the Longman Learners' Corpus. The LDOCE shows that

grammatical errors in learners' writing are mostly isolated, but there are certainly a lot of consecutive errors (e.g., the unnecessary preposition "*of*" immediately followed by a wrong verb form "*thinking*" in "*These machines are destroying our ability of thinking* [*to think*]."). We refer to two or more errors appearing consecutively as *serial errors*. Previous work on grammar checkers either focuses on handling one common type of errors exclusively, or independently. However, if an error is not isolated, it becomes difficult to correct the error when another related error is in the immediate context. In other words, when serial errors occur in a sentence, a grammar checker needs to correct the first error in the presence of the second error (or vise versa), making correction hard to solve. These errors could be corrected more effectively, if the corrector recognized them as serial errors and attempt to correct the serial errors at once.

Consider an error sentence "*I have difficulty to understand English.*" The correct sentence for this should be "*I have difficulty in understanding English*." It is hard to correct these two errors one by one, since the errors are dependent on each other. Intuitively, by identifying "*difficulty to understand*" as containing serial errors and correcting it to "*difficulty in understanding*," we can handle this kind of problem more effectively.

We present a new method for correcting serial grammatical errors in a given sentence in learners' writing. In our approach, a statistical machine translation model based on trigram containing a word followed by preposition and verb, or infinitive in web-scale ngrams data is generated to attempt to translate the input into a grammatical sentence. The method involves automatically learning two translation models based on Web-scale n-gram (Brants and Franz, 2006). The first model translates trigrams containing serial preposition-verb errors into correct ones. The second model is a back-off model, used in the case where the trigram is not found in the training data.

Input: *I have difficulty to understand English.* .	
Phrase table of translation model:	Back-off translation model:
difficulty of understanding ||| difficulty in understanding ||| 1.00	difficulty of VERB+ing ||| difficulty in VERB+ing ||| 0.80
difficulty to understand ||| difficulty in understanding ||| 1.00	difficulty to VERB ||| difficulty in VERB+ing ||| 1.00
difficulty with understanding ||| difficulty in understanding ||| 1.00	difficulty with VERB+ing ||| difficulty in VERB+ing ||| 1.00
difficulty in understand ||| difficulty in understanding ||| 1.00	difficulty in VERB ||| difficulty in VERB+ing ||| 1.00
difficulty for understanding ||| difficulty in understanding ||| 1.00	difficulty for VERB+ing ||| difficulty in VERB+ing ||| 1.00
difficulty about understanding ||| difficulty in understanding ||| 1.00	difficulty about VERB+ing ||| difficulty in VERB+ing ||| 1.00
Output: *I have difficulty in understanding English.*	

Figure 1. Example system translates the sentence "I have difficulty to understand English."

At run-time, our system will generate multiple possible trigram by changing word's preposition and verb form in the original trigram. Example trigrams generated for "difficulty to understand" are shown in Figure 1. The system will then rank all these generated sentences and use the highest ranked sentence as suggestion.

To conclude, we have introduced a new method for correcting serial errors in a given sentence in learners' writing. In our approach, a statistical machine translation model is generated to attempt to translate the given sentence into a grammatical sentence. The method involves automatically learning two translation models based on Web-scale n-gram. Evaluation on a set of sentences in a learner corpus shows that the proposed method corrects serial errors reasonably well with a precision of 0.68, recall 0.33 and F-score 0.45. Our methodology exploits the state of the arts in machine translation to develop a system that can effectively deal with serial errors or many error types at the same time.

References

Thorsten Brants and Alex Franz. The Google Web 1T 5-gram corpus version 1.1. LDC2006T13, 2006.

Rachele De Felice and Stephen G. Pulman. Automatic detection of preposition errors in learner writing. *CALICO Journal*, 26(3):512–528, 2009.

Michael Gamon. Using mostly native data to correct errors in learners' writing. In *Proceedings of the Eleventh Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Los Angeles, 2010.

David Graddol, 2006. English next: Why global English may mean the end of 'English as a Foreign Language.' UK: British Council.

Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi, and Hitoshi Isahara. Automatic error detection in the Japanese learners' English spoken data. In *Companion Volume to the Pro- ceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 145–148, 2003.

Leacock Claudia et al. 2010. Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies,* 3(1) 1–134.

John Lee and Stephanie Seneff. 2006. Automatic grammar correction for second-language learners. In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech)*, pages 1978–1981.