Automatically Identifying and Correcting Errors in Learner Writing using a Word Cluster Approach

Laurence Anthony

Center for English Language Education in Science and Engineering

School of Science and Engineering

Waseda University

3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

Tel/Fax : 03-5286-3845

Email : anthony@antlab.sci.waseda.ac.jp WWW : http://www.antlab.sci.waseda.ac.jp/

To date, there has been a great deal of work on the analysis and tagging of errors in learner writing. There has also been many attempts to build systems that can automatically identify and correct learner errors at the character, word, or part-of-speech (POS) level. Although such systems can provide valuable information on spelling and obvious grammatical mistakes, they are often less successful when dealing with sentence-level errors. For example, there are few systems that can identify, and more importantly correct, common errors made by foreign language learners, such as "What do you play sports?" and "I work a part-time job." A notably exception is the EDEN (Error Detection for English) system, recently developed by Saiga et al. (2003), that uses a corpus of learner generated data manually tagged for errors. This corpus is used as training data for a supervised learning system, in which target sentence errors are classified into pre-determined error classes using a maximum entropy approach. Indeed, the EDEN system has shown promising results when the target data corresponds closely with the training data. On the other hand, it performs less will on target data from different domains. Indeed, all systems that adopt a supervised learning approach based on a tagged learner corpus are likely to suffer from similar domain restrictions.

In this paper, I propose an alternative method for automatically identifying and correcting learner errors that operates at the sentence level, but does not require a pre-defined set of error classes, or a set of training data tagged for errors. Instead of attempting to locate all possible errors produced by learners, error identification is restricted to the task of finding differences (errors) between a target sentence and candidate sentences from a corpus of template "correct" sentences. First, the system searches for the most similar template sentence in the corpus using a standard distance measure. Next, the system locates word clusters in the target sentence that

match those in the template sentence, also detecting if there are any redundant or absent clusters. Following this, the clusters are ranked by length, and then the system presents the user with a minimum set of instructions (corrections) for re-arranging, deleting or adding word-clusters to the target sentence so that it eventually matches the template sentence. In this way, the system is able to instruct the user to move the word "sports" to the position after "What" to correct the first example above, and replace "work" with "do" in the second example automatically.

Although the proposed system is not applicable in contexts where a set of predetermined, "correct" template sentences are unavailable, such as in essay writing, it can be applied successfully to many popular CALL systems, such as online quizzes and restricted question/answer exercises. Currently, the system is able to deal with a wide range of learner errors, including word ordering, redundancy, repetition, word choice and so on. By performing a POS analysis of the template and test sentences, it is also possible to give increasingly detailed corrections, allowing the system to 'guide' the learner through the error correction process, using instructions such as "move the last word of the sentence to the beginning, use a verb for the second word, and change the tense of the third word". Finally, the system can serve as a practical aid for teachers assessing learner writing by hand, and will become increasingly effective if corrected learner sentences are dynamically added to the template sentence corpus.