A SYSTEM TO DISTINGUISH NATIVE AND NONNATIVE WRITTEN ENGLISH

Computer Science

Missouri State University, May 2012

Master of Natural and Applied Science

Philip Magnus Robinsson White

ABSTRACT

English grammar correction tools intended for native speakers of English are common, and those directed towards English-learners are becoming increasingly common as well. However, there is no current software that provides grammatical feedback to advanced language-learners — individuals who are generating grammatically correct samples of the target language, yet whose output is still distinguishable from that produced by a native speaker. This study focuses on the development of such a system. In particular, this study shows how text can be mined for grammatical features which can then be used with machine-learning algorithms to classify the text as having been generated by a native or nonnative speaker. Using a number of native and nonnative corpora of written English as training and testing data, this study shows that such classification can be done with a high level of accuracy. Three separate experiments are presented, each dealing with a different method of extracting features from automatically-generated parse trees and dependency graphs. The first and simplest of these experiments explores using dependency relations as classifier features. The other two experiments delve more deeply into specifics of grammar, looking at verbal argument usage and verb forms, respectively. The accuracy of the resultant classifiers ranges from 70% up to 94%. The classifiers are then analyzed to determine the linguistic significance of the features used. This study considers only nonnative English generated by first-language Spanish speakers, but the methods shown are not in any way restricted to that particular class of nonnative English. This study also proposes and describes an interactive system that would take advantage of this classification process to provide the learner with detailed information on which aspects of his or her language mark him or her as a nonnative speaker.

KEYWORDS: language acquisition, English, Spanish, classification, parsing

This abstract is approved as to form and content

Lloyd A. Smith, Ph.D.
Chairperson, Advisory Committee
Missouri State University