1 History and Similar Efforts

Automated document classification has been a highly productive area of study in the fields of text mining and machine learning. Initial attempts at document classification, and automated classification in general, were in the form of expert systems. Expert systems consist of human-compiled rules which are applied to a particular instance, such as a document, and make decisions about that instance based on occurrences of predefined features [Clifford et al. 1983]. Such a system relies entirely on the knowledge of the human expert who compiles the rules. The goal of such a system is to automate the classification process, but not necessarily to classify any more accurately than might an expert human. While such a system can theoretically approach the accuracy of the human who compiled the rules, it is incapable of exceeding that accuracy, and it comes at a high cost in terms of development time. Early attempts to solve these problems, in particular the latter, explored the automated creation of rule sets. Apté et al. [1994] was an early study into the effectiveness of computer-generated rule sets. Because of the enormous time savings afforded by such a system, and because of the potential of such systems to exceed the accuracy of humans, classification systems using computer-generated rule sets have become very common in commercial products (consider, for instance, spam email filtering [Cormack 2008]) and continue to be a very active area of research. The reader who desires a more detailed account of the history of document classification and a survey of its current (as of 2004) state is referred to Berry [2004].

One of the goals of the current study is to explore how a text classification system can be used to develop a software tool which provides suggestions to language learners on how they can improve their written language. In light of that, it is worth exploring similar existing tools. Grammar checkers are one such tool. A great deal of work has been done in the field of automatic grammar checking and, in many ways, this a very mature field of computer science. English grammar checkers have been available in commercial word processors for nearly two decades [Vernon 2000], and there has been much progress,

recently, on grammar checkers targeted towards language learners. One such effort is Microsoft Research's ESL Assistant. One element of this system, described in Gamon [2010], focuses on identifying common learner errors in article and preposition usage. The system uses maximum entropy classifiers in a novel approach to determine whether a particular location in a text should have an article or preposition and, if so, which specific article or preposition. The system looks at word boundaries within the text. For each boundary, it gathers features from the six words to the right and left of the boundary. It then applies the first of two classifiers to this feature set. The *presence classifier* determines the probability that the boundary is an appropriate location for an article or preposition. Then, if a location with a high probability is identified, a second classifier, the *choice classifier*, is applied to this set of features to identify the most likely appropriate articles or prepositions for use in that location. The system then compares the results of the classifiers with any articles or prepositions actually used in that location, and makes suggestions to the user. One benefit of this approach is that the classifiers need only be trained on native texts, which are more plentiful than nonnative texts. The study also explains how meta-classifiers can be used to improve upon this approach. For this, a second error detection system is used as well, based on language models. Language models are systems that assign a probability to a sequence of words, indicating how likely it is for that sequence of words to occur in a natural language sample. Gamon then uses both of these error detection systems as part of a meta-classifier trained on a corpus of learner texts in which all preposition and article errors have been marked. He finds that the meta-classifier provides considerably better accuracy than either system used alone.

Lee and Seneff [2006] explore a generative approach to grammar correction for language learners. In this system, one sentence or utterance is processed at a time. A number of permutations on the input are produced by modifying articles and preposition, inflecting nouns and verbs, and modifying auxiliary verbs. A language model is then used to choose a small subset of these permutations as candidates for correction. The study used both hu-

man evaluators and automatic evaluators to determine whether the system was producing output more appropriate than what the learner had inputted.

Wagner et al. [2007] compare two automated grammatical error detection systems. Though the approaches they describe are not specific to learner English, the nature of the systems do make them well suited for such. The first approach they describe uses a *precision grammar*. A precision grammar is a set of grammatical rules (see Ch. **??**) designed to parse only strictly grammatical language. This is in contrast to the grammars used in general purpose parsers, which tend to be designed to accommodate common errors. In this approach, the precision grammar is used to identify errors (i.e. unparseable passages). Though this study does not focus on providing corrections, the authors note that it is possible to include special "mal-rules" to identify specific types of malformed syntax. The second approach Wagner et al. [2007] explore uses part of speech n-grams (i.e. considering only the parts of speech of sequences of words) with language models to identify grammatical errors.

The studies highlighted here are but a small sample of the work the has been done in the areas of automated grammatical error detection and correction, but they are representative of the two most common approaches used: a shallow approach, using language models and n-grams to identify errors; and a deep approach, using parsers. The technologies used to generate corrections are more varied, and the reader interested in a broader review is referred to Lee [2009].

As far as this author can tell, no research has been done towards a system based on distinguishing native English and well-formed nonnative English, nor directly towards a system that offers grammatical suggestions to improve already grammatical learner English. It is likely that a system using language models could be adapted towards these ends, though this is not the approach taken here. Most of the work done on providing automatic grammar evaluation has been focused on the routine grammatical errors made by native speakers, or on the errors typical of novice and intermediate English learners. There ap-

pears to be little in the way of automated tools for advanced learners who wish to bring their writing skills to near-native levels of proficiency.

References

APTÉ, C., DAMERAU, F., AND WEISS, S. M. 1994. Automated learning of decision rules for text categorization. *ACM Trans. Inf. Syst. 12,* 3, 233–251.

BERRY, M. 2004. *Survey of text mining: clustering, classification, and retrieval*. Number v. 1. Springer.

CLIFFORD, J., JARKE, M., AND VASSILIOU, Y. 1983. A Short Introduction to Expert Systems. *SSRN eLibrary*.

CORMACK, G. V. 2008. Email spam filtering: A systematic review. *Found. Trends Inf. Retr. 1,* 4, 335–455.

GAMON, M. 2010. Using mostly native data to correct errors in learners' writing: A meta-classifier approach. In *2010 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

LEE, J. AND SENEFF, S. 2006. Automatic grammar correction for second-language learners. In *INTERSPEECH-2006*. 1978–1981.

LEE, J. S. Y. 2009. Automatic correction of grammatical erros in non-native english text. Ph.D. thesis, Massachusetts Institute of Technology.

VERNON, A. 2000. Computerized grammar checkers 2000: capabilities, limitations, and pedagogical possibilities. *Computers and Composition 17,* 3, 329 – 349.

WAGNER, J., FOSTER, J., AND VAN GENABITH, J. 2007. A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors.

In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language processing and Computational Natural Language Learning*. Association for Computational Linguistics, 112–121.