Automated Essay Scoring: A Cross-Disciplinary Perspective

Mark D. Shermis and Jill C. Burstein (editors)
(Florida International University and ETS Technologies, Inc.)

Mahwah, NJ: Lawrence Erlbaum
Associates, 2003, xvi+238 pp;
hardbound, ISBN 0-8058-3973-9, $59.95
($24.50 prepaid)

Reviewed by
Lawrence M. Rudner
Graduate Management Admission Council

The vision of having effective algorithms score student essays should be appealing to the teacher, test publisher, and research scientist. Teachers would be freed of the burden of reading and hand-scoring maybe hundreds of student papers and consequently would be more likely to assign written questions and probe for deeper understanding. Test publishers would be able to score essays for less cost and conceivably provide higher-quality assigned grades. Research scientists, especially readers of this journal, would find this to be a fascinating area, one that merges research from multiple disciplines and having many avenues yearning for exploration.

Shermis and Burstein's is the first book dedicated to the topic of automated essay scoring (AES). As such, it is destined to be the seminal work in the area. The book is composed of thirteen chapters, each written by a different set of authors. Five of the chapters provide descriptions of five different approaches and form the heart of the book. There are also four chapters on psychometric issues, two on innovations, a formal introduction, and two introductory chapters. The authors are all authorities. The five approaches are described by their developers or major advocates.

As with other forms of artificial intelligence, the task of AES is to accomplish a human goal. This does not mean that the goal needs to be accomplished using the same techniques as humans use. In the case of AES, humans typically read a passage and look for certain prespecified key concepts defined in a scoring rubric. Readers call upon their content knowledge, literary experience, and language skills in evaluating the passage.

The computer cannot possibly score an essay the same way. Rather, AES seeks to use the computer's special capabilities. The computer can count surface features, examine individual words and phrases, look at word order, stem, identify stop words, parse each sentence, examine sentence-to-sentence relatedness, weigh different features, identify arguments, and compare each new essay to hundreds of prescored essays. The question is whether the results are adequate.

If the goal is to approximate human scores, then the answer is yes for all the approaches. Timothy Keith provides an extremely informative chapter on the predictive validity of several AES programs. The programs tend to yield impressively high correlations with the scores of human raters—generally between .70 and .90 and often between .80 and .85. Further, the correlations of AES with human raters cannot be distinguished from the correlations among human raters. As Ellis Page points out in a chapter describing Project Essay Grade (PEG), AES in a sense "passes the Turing test"—an outside observer cannot tell the difference between the machine and a human. Another way to look at the accuracy of AES is to examine the percentage of

agreement between AES and human scores. In practice, AES scores are considered to be comparable to human ratings if the two are within one point of each other: "adjacent accuracy." Several chapters report adjacent accuracies of 90–99%. By this criterion, it is fairly easy for an AES system to be adequate. Adjacency covers much of a scoring scale—half of a six-point scale and three-fourths of a four-point scale. Further, most scores are typically in the middle of the score range, again increasing the likelihood of obtaining a near-perfect adjacent accuracy.

As stated earlier, the heart of the book is the five chapters devoted to different methods. Three of the methods have been described in the professional literature and appear to be fairly mature—PEG, which is described by Ellis Page; e-Rater, which is described by Jill Burstein; and Intelligent Essay Assessor (IEA), which is described by Thomas Landauer, Darrell Laham, and Peter Foltz. Two additional approaches are presented. Leah Larkey and Bruce Croft present the details of a Bayesian approach based on the well-developed text classification literature. Scott Elliot provides a summary of studies conducted using Intellimetric. The five approaches are all quite different.

PEG and the Bayesian approach are the simplest. Using a large collection of surface features such as average sentence length, frequency of certain transitional words, number of semicolons, and word rarity, PEG yields extremely impressive AES–human rater correlations. These surface features appear to be effective proxies for the intrinsic variables that humans look for. The Bayesian approaches examine the probabilities of each token (typically a word or a stemmed word) being used in essays in each score group. Larkey and Croft present a wonderful analysis of a variety of approaches.

On the other end of the spectrum, IEA and e-Rater have a much deeper linguistic base. IEA examines content, style, and mechanics, with content expressed as independent measures of semantic quality and the amount of such content. E-Rater examines discourse structure, syntactic structure, and vocabulary usage. While the underlying mathematics is different, the two approaches share an underlying philosophy of relying on natural language processing rather than mechanical features.

If a reader is looking for an understanding of the approaches and potential of AES, this is the book to read. All the current approaches are presented in one volume. The authors do an excellent job of describing the philosophy and history of their approaches. Of particular interest are ideas for providing diagnostic and evaluative feedback that are sprinkled throughout the book. The chapters are, however, quite independent and in the wrong order. I suggest starting with the introduction, then moving to the descriptions of the approaches, psychometric issues in AES, and innovations in AES. The first two chapters, which are probably intended to provide a general framework and background, can be skipped without any loss.

Lawrence Rudner is the chief statistician with the Graduate Management Admission Council. He is the author of the Bayesian Essay Test Scoring sYstem (BETSY), which is available for research use at http://edres.org/betsy/. His research interests include AES, computer adaptive testing, and decision theory. Rudner's address is 1600 Tysons Boulevard, Suite 1400, McLean, VA 22102; e-mail: LRudner@gmac.com.