Automated essay scoring

Automated essay scoring (AES) is the use of specialized computer programs to assign grades to essays written in an educational setting. It is a form of educational assessment and an application of natural language processing. Its objective is to classify a large set of textual entities into a small number of discrete categories, corresponding to the possible grades, for example, the numbers 1 to 6. Therefore, it can be considered a problem of statistical classification.

Several factors have contributed to a growing interest in AES. Among them are cost, accountability, standards, and technology. Rising education costs have led to pressure to hold the educational system accountable for results by imposing standards. The advance of information technology promises to measure educational achievement at reduced cost. The use of AES for high-stakes testing in education has generated significant backlash, with opponents pointing to research that computers cannot yet grade writing accurately and arguing that their use for such purposes promotes teaching writing in reductive ways (i.e. teaching to the test).

Most historical summaries of AES trace the origins of the field to the work of Ellis Batten Page.[1] In 1966, he argued[2] for the possibility of scoring essays by computer, and in 1968 he published[3] his successful work with a program called Project Essay Grade (PEG). Using the technology of that time, computerized essay scoring would not have been cost-effective,[4] so Page abated his efforts for about two decades. Eventually, Page sold PEG to Measurement Incorporated By 1990, desktop computers had become so powerful and so widespread that AES was a practical possibility. As early as 1982, a UNIX program called Writer's Workbench was able to offer punctuation, spelling and grammar advice.[5] In collaboration with several companies (notably Educational Testing Service), Page updated PEG and ran some successful trials in the early 1990s.[6] Peter Foltz and Thomas Landauer developed a system using a scoring engine called the Intelligent Essay Assessor (IEA). IEA was first used to score essays in 1997 for their undergraduate courses.[7] It is now a product from Pearson Educational Technologies and used for scoring within a number of commercial Contents History products and state and national exams. IntelliMetric is Vantage Learning's AES engine. Its development began in 1996.[8] It was first used commercially to score essays in 1998.[9] Educational Testing Service offers "e-rater", an automated essay scoring program. It was first used commercially in February 1999.[10] Jill Burstein was the team leader in its development. ETS's Criterion Online Writing Evaluation Service uses the e-rater engine to provide both scores and targeted feedback. Lawrence Rudner has done some work with Bayesian scoring, and developed a system called BETSY (Bayesian Essay Test Scoring sYstem).[11] Some of his results have been published in print or online, but no commercial system incorporates BETSY as yet. Under the leadership of Howard Mitzel and Sue Lottridge, Pacific Metrics developed a constructed response automated scoring engine, CRASE. Currently utilized by several state departments of education and in a U.S. Department of Education-funded Enhanced Assessment Grant, Pacific Metrics' technology has been used in large-scale formative and summative assessment environments since 2007. Measurement Inc. acquired the rights to PEG in 2002 and has continued to develop it.[12] In 2012, the Hewlett Foundation sponsored a competition on Kaggle called the Automated Student Assessment Prize (ASAP).[13] 201 challenge participants attempted to predict, using AES, the scores that human raters would give to thousands of essays written to eight different prompts. The intent was to demonstrate

that AES can be as reliable as human raters, or more so. The competition also hosted a separate demonstration among nine AES vendors on a subset of the ASAP data. Although the investigators reported that the automated essay scoring was as reliable as human scoring,[14] this claim was not substantiated by any statistical tests because some of the vendors required that no such tests be performed as a precondition for their participation.[15] Moreover, the claim that the Hewlett Study demonstrated that AES can be as reliable as human raters has since been strongly contested,[16][17] including by Randy E. Bennett, the Norman O. Frederiksen Chair in Assessment Innovation at the Educational Testing Service. [18] Some of the major criticisms of the study have been that five of the eight datasets consisted of paragraphs rather than essays, four of the eight data sets were graded by human readers for content only rather than for writing ability, and that rather than measuring human readers and the AES machines against the "true score", the average of the two readers' scores, the study employed an artificial construct, the "resolved score", which in four datasets consisted of the higher of the two human scores if there was a disagreement. This last practice, in particular, gave the machines an unfair advantage by allowing them to round up for these datasets.[16]

According to a rencent survey[19] , modern AES systems try to score different dimenions of an essay's quality in order to provide feedback to users. These dimensions include the following items: Grammaticality: following grammar rules Usage: using of prepositions, word usage Mechanics: following rules for spelling, punctuation, capitalization Sytle: world choice, sentense structure variety Relevance: how relevant of the content to the prompt Organization: how well the essay is structured Development: development of ideas with examples Coherison: appropriate use of transition phrases Coherence: appropriate transitions between ideas Different dimensions of essay quality Thesis Clarity: clarity of the thesis Persuasivenes: convincingness of the major argument

From the beginning, the basic procedure for AES has been to start with a training set of essays that have been carefully hand-scored.[20] The program evaluates surface features of the text of each essay, such as the total number of words, the number of subordinate clauses, or the ratio of uppercase to lowercase letters— quantities that can be measured without any human insight. It then constructs a mathematical model that relates these quantities to the scores that the essays received. The same model is then applied to calculate scores of new essays. Recently, one such mathematical model was created by Isaac Persing and Vincent Ng.[21] which not only evaluates essays on the above features, but also on their argument strength. It evaluates various features of the essay, such as the agreement level of the author and reasons for the same, adherence to the prompt's topic, locations of argument components (major claim, claim, premise), errors in the arguments, cohesion in the arguments among various other features. In contrast to the other models mentioned above, this model is closer in duplicating human insight while grading essays. The various AES programs differ in what specific surface features they measure, how many essays are required in the training set, and most significantly in the mathematical modeling technique. Early attempts used linear regression. Modern systems may use linear regression or other machine learning techniques often in combination with other statistical techniques such as latent semantic analysis[22] and Bayesian inference. [11]

Any method of assessment must be judged on validity, fairness, and reliability. [23] An instrument is valid if it actually measures the trait that it purports to measure. It is fair if it does not, in

effect, penalize or privilege any one class of people. It is reliable if its outcome is repeatable, even when irrelevant external factors are altered. Before computers entered the picture, high-stakes essays were typically given scores by two trained human raters. If the scores differed by more than one point, a more experienced third rater would settle the disagreement. In this system, there is an easy way to measure reliability: by inter-rater agreement. If raters do not consistently agree within one point, their training may be at fault. If a rater consistently disagrees with how other raters look at the same essays, that rater probably needs extra training. Various statistics have been proposed to measure inter-rater agreement. Among them are percent agreement, Scott's π, Cohen's κ, Krippendorf's α, Pearson's correlation coefficient r, Spearman's rank correlation coefficient ρ, and Lin's concordance correlation coefficient. Percent agreement is a simple statistic applicable to grading scales with scores from 1 to n, where usually $4 \leq n \leq 6$. It is reported as three figures, each a percent of the total number of essays scored: exact agreement (the two raters gave the essay the same score), adjacent agreement (the raters differed by at most one point; this includes exact agreement), and extreme disagreement (the raters differed by more than two points). Expert human graders were found to achieve exact agreement on 53% to 81% of all essays, and adjacent agreement on 97% to 100%.[24] Procedure Criteria for success Inter-rater agreement can now be applied to measuring the computer's performance. A set of essays is given to two human raters and an AES program. If the computer-assigned scores agree with one of the human raters as well as the raters agree with each other, the AES program is considered reliable. Alternatively, each essay is given a "true score" by taking the average of the two human raters' scores, and the two humans and the computer are compared on the basis of their agreement with the true score. Some researchers have reported that their AES systems can, in fact, do better than a human. Page made this claim for PEG in 1994.[6] Scott Elliot said in 2003 that IntelliMetric typically outperformed human scorers.[8] AES machines, however, appear to be less reliable than human readers for any kind of complex writing test.[25] In current practice, high-stakes assessments such as the GMAT are always scored by at least one human. AES is used in place of a second rater. A human rater resolves any disagreements of more than one point.[26]

AES has been criticized on various grounds. Yang et al. mention "the over-reliance on surface features of responses, the insensitivity to the content of responses and to creativity, and the vulnerability to new types of cheating and test-taking strategies."[26] Several critics are concerned that students' motivation will be diminished if they know that no human will read their writing.[27] Among the most telling critiques are reports of intentionally gibberish essays being given high scores.[28] On 12 March 2013, HumanReaders.Org launched an online petition, "Professionals Against Machine Scoring of Student Essays in High-Stakes Assessment". Within weeks, the petition gained thousands of signatures, including Noam Chomsky, [29] and was cited in a number of newspapers, including The New York Times, [30] and on a number of education and technology blogs.[31] The petition describes the use of AES for high-stakes testing as "trivial", "reductive", "inaccurate", "undiagnostic", "unfair" and "secretive".[32] In a detailed summary of research on AES, the petition site notes, "RESEARCH FINDINGS SHOW THAT no one—students, parents, teachers, employers, administrators, legislators—can rely on machine scoring of essays ... AND THAT machine scoring does not measure, and therefore does not promote, authentic acts of writing."[33] The petition specifically addresses the use of AES for high-stakes testing and says nothing about other possible uses. Most resources for automated essay scoring are proprietary.

eRater – published by Educational Testing Service Intellimetric – by Vantage Learning Project Essay Grade[34] – by Measurement, Inc. PaperRater Criticism HumanReaders.Org Petition Software