eGrader, a software application that automatically scores student essays: with a postscript on ethical complexities

Authors

Roxanne Byrne. Ph.D., Department of Mathematical and Statistical Sciences, University of Colorado Denver
Michael Tang, Ph.D., Science, Technology Studies, University of Colorado Denver
John Tranduc, (M.S. –pending), G.I.S. Program, University of Colorado Denver
Matthew Tang (M.S.—pending), G.I.S. Program, University of Colorado Denver

Abstract

Online and traditional teachers face several instructional challenges with regard to assessing student learning. This paper focuses on a software application that automatically scores student essay. The first part gives a brief overview of three commercial automated essay scoring systems. Then it describes the technical aspects of the machine grader developed by the authors, including an assessment of its performance. Although the statistical results were significant in finding a strong correlation between human and machine scorers and the other measures, follow-up non-quantitative evaluations led the researchers to discontinue using the eGrader. They concluded that while the eGrader's ability to measure objective evaluation criteria was successful, measuring subjective ideas proved to more complex and problematic.

Introduction

A survey conducted by authors of this paper found that essay assignments were perceived as among the more effective learning devices in higher education (Byrne and Tang, 2008). At the same time, the question arises as to how a human grader can score essays adequately when the number of essays to be graded is large and the time to evaluate them short (Hartley, et. al. 2006; Weseley and Addyson, 2007; Walvoord, et. al., 2008). One possible solution to the problem may be the adoption of an automated assessment tool for essays. Such a system, so it has been argued, could bring more consistency to the scoring of essays and at the same time promises cost and time savings.

The major impulse for the development, testing and use of automated scoring machines comes not from the groves of academe but from corporate testing enterprises such as the Education Testing Service (ETS). It is therefore of no surprise, that the one area in traditional academia where automated essay scoring services is making great in-roads is in the scoring of student essays by university admissions. In the placement area over 900 universities use machines to score written exams of over 5,000,000 students. In addition commercial test makers have entered actual classrooms by providing teachers with their software through foundation funding (Ericson, 2006, 3-4).

Commercial Services

What follows is a brief overview of three commercial automated essay scoring systems available today. For other summaries of these and other systems such as C-Rater, BETSY, Intelligent Essay Marking System, SEAR, Paperless School free text Marking Engine and Automark, see: Velanti (2003) and Shermis and Burnstein (2003).

1) Project Essay Grade (http://www.measinc.com/Default.aspx?Page= ETS.AutomatedEssayScoring)

Project Essay Grade, or PEG, is one of the earliest implementations of automated essay grading. It was developed by Page and others (Hearst, 2000; Page, 1996). According to the commercial website that sells the system, it is based on more than 40 years of research in computational linguistics, and its authors claim that PEG's scoring results have been validated in more independent studies than all other essay scoring solutions combined. For analysis of functions and design, see: (Valenti, Neri & Cucchiarelli, 2003;Shermis, 2003).

2) IntelliMetric® (http://www.vantagelearning.com/school/products/intellimetric/)

IntelliMetric is the result of Lawrence M. Rudner's (Rudner and Liang, 2002) early research on an automated essay grading system called the Bayesian Essay Scoring sYstem (BETSY). IntelligMetric uses Bayesian computer adaptive testing (Frick 1992; Madigan, Hunt, Levidow, and Donnell ,1995; and Rudner , 2001) to classify select "items" or essay features into a three or four point categorical scale. Its most profitable product is MyAccess, an automated writing tool to improve student writing and prepare them for the essay portions of exams, such as the Graduate Management Admission Test (GMAT®) for entrance into business schools with MBA programs.

3) Criterion's e-Rater® (http://www.ets.org, then Products)

E-Rater is a software engine, develop in the mid-90's is perhaps the most successful of the commercial automated writing evaluators and has been used since 1999 to score the essay portion of the GMAT (Burstein, 2003; Kukich,2000). E-Rater uses Microsoft's natural language parser and a companion software application called Critique to rate essays according to rates of errors as flagged by Microsoft's style and grammar checker. Critique takes into account statistics based on redundancy, length of essay, vocabulary and the number of required discourse elements such as thesis statement, main idea, or supporting idea.

Evaluation of commercial services

Most of the systems developed are aimed to grade essays both for style and content. Recent research, however, indicates that using content as a criterion for scoring may not be as essential as one would think as indicated by the success of PEG, which takes content as a minimal criterion. (Shermis, M.D., Shneyderman, A. and Attali, Y., 2005). We also found that for content analysis, Bayesian analysis, NLP, and LSA appear to be the most successful techniques used in automated essay grading. Thirdly, the main methods to measure a system's performance are experiments designed to find a correlation between the scores of human readers versus machine readers.

The eGrader

This machine essay scorer in contrast to the commercial services we surveyed: 1) operates on a client PC; 2) is cost effective; 3) needs little human training; and 4) does not require a huge data base and large computing power. The design approach is a-theoretical, empirical, and statistical (Anderson, 2008). Its development is influenced by three relatively unknown applications originated by S. R. Hawkins (1993) for natural language processing, by Alan Mole (1994) for machine translation and, by Barbara S. Glatt (1984) for plagiarism detection. Hawkins showed that a machine can be built that appears to understand textual meaning and do well on a Turing Test using Ogden and I.A. Richard's semantic theory. Mole built a successful translating machine for 33 foreign languages in the 1990's and discovered that that grammar, word order and other niceties of

language such as prepositions are unnecessary in understanding textual meaning. Glatt developed a plagiarism detection program based on the Cloze procedure that is used in foreign language and English as a second language teaching to test for reading comprehension. See: http://www.plagiarism.com/screening.htm

The eGrader (eG) uses a key word search function to download web pages that are then stored in a client computer directory to use as benchmark data to score student essays and other forms of writing. The Web documents are analyzed in turn by a semantic technique to provide a content analysis of the targeted writing. A second directory can be used to store specific content data such as relevant readings and sample student essays for similar analysis. Its core algorithm to analyze content may have similarities to Intellimetric's where: *meaning of* $word_1$ + *meaning of* $word_2$ + *meaning of* $word_3$ = *meaning of passage* (Ericsson, 2006, 29). The algorithm is based on a key word and concordance analyzer to measure similar concepts and usage between benchmark writing samples stored in the two directories.

For writing structure, like PEG, eGrader uses readability statistics based on Flesch Kincaid equations. These readability statistics include essay length, grade levels, and a proprietary algorithm that measures complexity of sentence structure based on a connective word counting device. The eGrader does not analyze grammar or mechanics and does not rely on traditional NPL or LSA theory or techniques to make its calculations. In addition, unlike other systems, it does not use vectors to associate words, concepts and documents to build a relational database. The eG is implemented on Netbeans IDE 6.1. It exploits Java Desktop Application template and Java Swing components for designing the user interface

In summary, the program uses the following rubrics to calculate its essay score:

1) key word and important concept comparison
2) concordance and similar usage comparison;
3) writing style and complexity measures; and
4) grade level, length and reading ease of writing.

The program's output then gives results in columns that can be sorted from highest to lowest so the user can assign grades scores according to an absolute or relative curve. This relative method of reporting simplifies the problem of weighting scores according to some pre-determined standard or writing level of students. A final column gives a composite score of the average of all the rubrics.

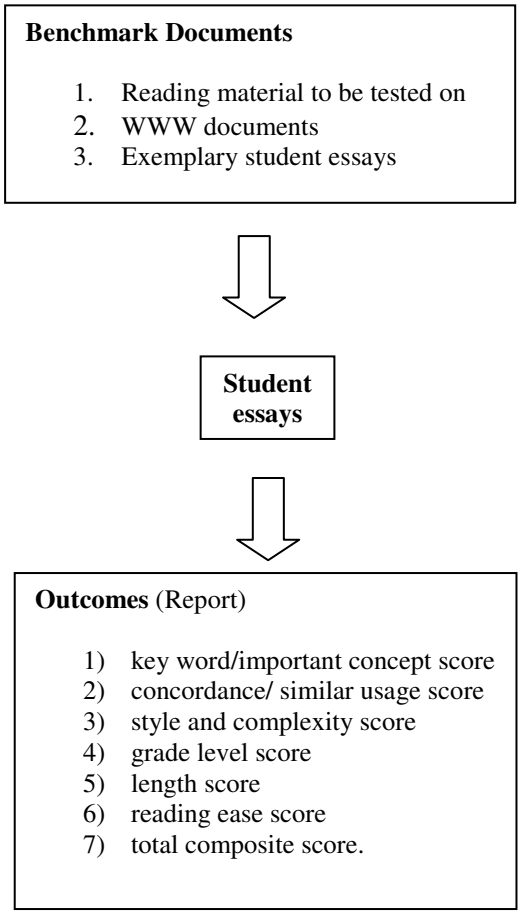

Fig 2: flow chart and structure of eG's input and output.

Performance

The eGrader's scores for 33 student essays were compared with the scores of the same papers given by 3 different human readers. The scores were then compared to determine their degree of correlation.

1. Human Reader 1 scores versus machine scores: $r = 85\%$
2. Human Reader 2 scores versus machine scores: $r = 75\%$
3. Human Reader 3 scores versus machine scores: $r = 74\%$

These results are comparable to other commercial systems. For example, in a similar experiment, researchers evaluated the IntelliMetric™ automated essay scoring system's performance by comparing human scorers versus machine scorers of essays from the Analytic Writing Assessment of GMAT. In two experiments, they found that Pearson *r* correlations of agreement between human raters and the IntelliMetric system averaged 83%. (My Access, 2008). According to ETS and Velanti et. al. (2003), over 750,000 GMAT essays have been scored with Criterion's eRater. By comparing human and e-Rater grades across 15 test questions, the empirical results range from 87% to 94%.

Conclusion with an ethical postscript

Even though the testing results were comparable to those claimed by the commercial testing services its developers decided not to continue eGrader's use in the classroom. While developing and testing the software several issues led to conclude that the LSA technique we used for eGrader could not detect meaning as per its classical definition of the word (Ericsson 2006). Moreover, while commercial firms claim that their algorithms "simulate human judgments and behavior . . . quite well," it is our judgment that this simulation undervalued the human effort of students. It is probably for this reason that several students expressed their concern of having a machine read their papers.

When using eGrader in a class (not part of the original experiments), an instructor informed students that their essays would be graded by a machine but students could ask for a second, human reading if they felt the grade was not a fair measurement of their work. Instructor qualms about the use of automated essay scoring machines emerged after 10 students asked for a rereading of their work. After doing so, the instructor changed three of the grades to A's by increasing their scores an average of 24%. During the grading of the papers, the instructor found a disturbing pattern. The machine algorithm could not detect ideas that were not contained in the readings or Web benchmark documents although the ideas expressed were germane to the essay question. For example, one student who received an average machine score wrote an essay that compared the required readings with ideas from another course she was taking. The machine content analyzer of course did not recognize the ideas that the student used from another class. Consequently, eGrader scored her essay low in content. Further discrepancies between the human reader and the machine reader suggest that machine readers could not detect other subtleties of writing such as irony, metaphor, puns, connotation and other rhetorical devices. For these and other reasons, the instructor decided not to use eGrader in further scoring of student essays. The machine reader appears to penalize those students we want to nurture, those who think and write in original or different ways. For us the subjective element, which was as important as the objective aspects of the essays, proved too complex to measure.

Works Cited

Anderson, Chris (June 23, 2008). The end of theory: the data deluge makes the scientific method obsolete Wired Magazine. 16.07

Attali, Yigal and Powers, Don (April, 2008) A Developmental Writing Scale, ETS Research Report**.** Princeton, NJ: ETS. 4, 1-59.

Bonwell, Charles C. (Fall 1996). Enhancing the lecture: revitalizing the traditional format. New Directions for Teaching and Learning, n67 p31-44

Christie, J. R. (1999). Automated essay marking-for both style and content. In M. Danson (Ed.), Proceedings of the Third Annual Computer Assisted Assessment Conference. Loughborough University, Loughborough, UK.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman R. A. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science. *41*(6), 391-407.

Ericsson, Patricia Freitag and Haswell, Richard (2006), eds. Machine Scoring of Student Essays: Truth and Consequences**,** Utah State UP, 2006

Glatt, Barbara S.; Haertel, Edward H. (Spr 1982) The use of the Cloze Testing Procedure for detecting plagiarism. Journal of Experimental Education, v50 n3 p127-36

Hartley, James; Trueman, Mark; Betts, Lucy; Brodie, Lauren (Oct 2006) What price presentation? The effects of typographic variables on essay grades. Assessment & Evaluation in Higher Education, v31 n5 p523-534

Hawkins, S. R. (1993). Ogden's Basic English as a lexical database for natural language processing. Submitted in partial fulfillment of the requirements for a degree of Master of Science in the Department of Computer Science, University of South Carolina.

Hearst, M. (2000). The debate on automated essay grading. IEEE Intelligent Systems, IEEE CS Press. 15(5), 22-37.

Kakkonen, Tuomo; Myller, Niko; Sutinen, Erkki; Timonen, Jari (2008). Comparison of dimension reduction methods for automated essay grading. Educational Technology & Society, v11 n3 p275-288 2008

Kincaid, J. P.; Fishburne, R. P., Jr.; Rogers, R. L.; and Chissom, B. S. (March, 1975); Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel, Research Branch Report 8-75, Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN

Landauer, T.K., Laham, D., & Derr, M. (2004). From paragraph to graph: Latent semantic analysis for information visualization. Proceedings of the National Academy of Science, 101, 5214-5219.

Mole, R.A. (1994). English to any language: conversational translator program. Entente$^{TM.}$ Boulder, CO.

MY Access!® Efficacy Report September 2007, Vantage Learning. 1-12.

Page, E.B. (1994). New computer grading of student prose, using modern concepts and software. Journal of Experimental Education, 62(2), 127-142

Rudner, L.M. & Liang, T. (2002). Automated essay scoring using Bayes' Theorem. The Journal of Technology, Learning and Assessment, 1(2), 3-21

Shermis, M. D. & Burstein, J. (2003). Automated Essay Scoring: A Cross Disciplinary Perspective. Mahwah, NJ: Lawrence Erlbaum Associates.

Shermis, M.D., Shneyderman, A. and Attali, Y. (March, 2005). How important is content in the ratings of essay assessments? Paper presented at the annual meeting of the National Council of Measurement in Education. Montreal.

Shores, Michael and Weseley, Allyson J (2007). When the A is for agreement: factors that affect educators' evaluations of student essays. Action in Teacher Education, 29(3), 4-11.

Valenti, Salvatore, Neri, Burstein, J., Leacock, C., & Swartz, R. (2001). Automated evaluation of essay and short answers. In M. Danson (Ed.). Proceedings of the Sixth International Computer Assisted Assessment Conference. Loughborough University, Loughborough, UK.

Valenti, Landauer, T.K., Laham, D., & Foltz, P.W. (2003). Automatic essay assessment. Assessment in Education: Principles, Policy & Practice, *10(3),* 295-308.

Walvoord, Mark E.; Hoefnagels, Marielle H.; Gaffin, Douglas D.; Chumchal, Matthew M.; Long, David A. (Mar-Apr 2008). An analysis of calibrated peer review (CPR) in a science lecture classroom. Journal of College Science Teaching, 37(4), 66-73

Warschauer, Mark & Paige Ware, Page (2006) Automated writing evaluation: defining the classroom research agenda Language Learning Research, 10 (2), 1-24