An Unsupervised Automated Essay-Scoring System

Yen-Yu Chen, *Industrial Technology Research Institute*

Chien-Liang Liu and Chia-Hoang Lee, *National Chiao Tung University*

Tao-Hsing Chang, *National Kaohsiung University of Applied Sciences*

The proposed automated essay-scoring system uses an unsupervised-learning approach based on a voting algorithm. Experiments show that this approach works well compared to supervised-learning approaches.

Automated essay scoring (AES) is the ability of computer technology to evaluate and score written prose. Proposed in 1966, AES has since been used successfully on large-scale essay exams. The goal is not to replace human raters. In current large exams, each essay is scored by two or more human raters, and the final scores are averaged over these scores. For example, in the Graduate Record Examination (GRE) analytical writing section, two trained readers score each essay. If there is more than a one-point difference between the two readers' scores, then a third reader grades the essay, and the score for that essay will be the average of the two highest scores. In general, the whole essay-scoring process is time consuming and requires considerable manpower. Therefore, instead of having two people score the essays, each essay could be scored by AES and a human rater, with the final then determined by both. The combined approach would still require the AES system and the human rater to assign a score within one scale point of each other. Otherwise, a third human rater would resolve the discrepancy.

Companies such as Vantage Learning and ETS Technologies have published research results that demonstrate strong correlations and nonsignificant differences between AES and human scoring.[1] In essence, the human raters grade the essays according to some criteria. For example, the GRE analytical writing score is based on a strong focus on the topic, good evidence to support arguments, and proper use of grammar. If an essay includes all of these factors, it could earn a top score. Therefore, the aim of AES systems is to simulate a human rater's grading process, and a system is usable only if it can perform the grading as accurately as human raters.

In this article, we propose an unsupervised AES system that requires only a small number of essays within the same topic without any scoring information. (See the "Related Research in Automated Essay Scoring" sidebar for details on other approaches.) The scoring scheme is based on feature information and the similarities between essays. We use a voting algorithm based on the initial scores and similarities between essays to iteratively train the system to score the essays. Our experiments yield an adjacent agreement rate of approximately 94 percent and

1541-1672/10/$26.00 © 2010 IEEE
Published by the IEEE Computer Society

Related Research in Automated Essay Scoring

Automated Essay Scoring (AES) has been a real and viable alternative and complement to human scoring for many years. In 1996, Ellis Page designed the Project Essay Grader (PEG) computer grading program.[1] Page looked for the kind of textual features that computers could extract from the texts and then applied multiple linear regressions to determine an optimal combination of weighted features that best predicted the teachers' grades. The features Page identified as having predictive power included word length and the number of words, commas, prepositions, and uncommon words in the essay. Page called these features proxies for some intrinsic qualities of writing competence. He had to use indirect measures because of the computational difficulty of implementing more direct measures.[2]

Because it only uses indirect features, however, this type of system is vulnerable to cheating. Therefore, it is a significant research challenge to identify and extract more direct measures of writing quality. For example, later research used machine learning to identify discourse elements based on an essay-annotation protocol.[3] Meanwhile, many researchers used natural language processing (NLP) and information retrieval (IR) techniques to extract linguistic features that might more directly measure essay qualities.

During the late 1990s, more systems were developed, including the Intelligent Essay Assessor (IEA), e-rater, and IntelliMetric. IntelliMetric successfully scored more than 370,000 essays in 2006 for the Analytical Writing Assessment (AWA) portion of the Graduate Management Admission Test (GMAT).

Intelligent Essay Assessor (IEA) uses latent semantic analysis (LSA) to analyze essay semantics.[4] The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. LSA captures transitivity relations and collocation effects among vocabulary terms, thereby letting it accurately judge the semantic relatedness of two documents regardless of their vocabulary overlap.[5]

IEA measures the content, style, and mechanics components separately and, whenever possible, computes each component in the same way so that score interpretation is comparable across applications. The system must be trained on a set of domain-representative texts to measure an essay's overall quality. For example, a biology textbook could be used when scoring biology essays. LSA characterizes student essays by representing their meaning and compares them with highly similar texts of known quality. It adds corpus-statistical writing-style and mechanics measures to help determine overall scoring, validate an essay as appropriate English (or other language), detect plagiarism or attempts to fool the system, and provide tutorial feedback.[6]

E-rater employs a corpus-based approach to model building, using actual essay data to examine sample essays. The features of e-rater include syntactic, discourse, and topical-analysis modules. The origin of the syntactic module is parsing. In discourse analysis, it assumes the essay can be segmented into sequences of discourse elements, which include introductory material, a thesis statement, main ideas, supporting ideas, and a conclusion.[7] To identify the various discourse elements, the system was trained on a large corpus of human-annotated essays. Finally, the topical-analysis module identifies vocabulary usage and topical content. In practice, a good essay must be relevant to the topic assigned. Moreover, the variety and type of vocabulary used in good essays differ from that of poor essays. The assumptions behind this module are that good essays resemble other good essays.

In recent years, many supervised-learning approaches on essay-scoring systems have been proposed. Lawrence M. Rudner and Tahung Liang used a Bayesian approach to perform AES, showing the effectiveness of the supervised-learning approach for essays.[8] Essentially, the supervised-learning model needs enough labeled data to construct the classification model. Our experiments indicate that such approaches require at least 200 scored essays, which make them inappropriate for environments where there are not enough scored essays.

References

1. E.B. Page, "The Imminence of Grading Essays by Computer," *Phi Delta Kappan*, vol. 47, 1966, pp. 238–243.
2. K. Kukich, "Beyond Automated Essay Scoring," *IEEE Intelligent Systems*, vol. 15, no. 5, 2000, pp. 22–27.
3. J. Burstein, D. Marcu, and K. Knight, "Finding the Write Stuff: Automatic Identification of Discourse Structure in Student Essays," *IEEE Intelligent Systems*, vol. 18, no. 1, 2003, pp. 32–39.
4. T. Landauer and S. Dumais, "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge," *Psychological Rev.*, vol. 104, no. 2, 1997, pp. 211–240.
5. M.A. Hearst, "The Debate on Automated Essay Grading," *IEEE Intelligent Systems*, vol. 15, no. 5, 2000, pp. 22–37.
6. T.K. Landauer, D. Laham, and P.W. Foltz, "The Intelligent Essay Assessor," *IEEE Intelligent Systems*, vol. 15, no. 5, 2000, pp. 27–31.
7. Y. Attali and J. Burstein, "Automated Essay Scoring with E-Rater v.2," *J. Technology, Learning and Assessment*, vol. 4, no. 3, 2006, http://escholarship.bc.edu/jtla/vol4/3.
8. L.M. Rudner and T. Liang, "Automated Essay Scoring Using Bayes' Theorem," *J. Technology, Learning and Assessment*, vol. 1, no. 2, 2002, http://escholarship.bc.edu/jtla/vol1/2.

an exact agreement rate of approximately 52 percent.

Overview of Unsupervised Learning

In supervised learning, we can regard the AES as a classification learner and the scores of the training essays as the training data categories. New essays will be classified into an appropriate category based on the features and the classification model. On the other hand, the training data in an unsupervised-learning classifier does not contain label information, so the classifier must determine how the data is organized from unlabeled examples. We propose a novel unsupervised-learning method and apply it to an essay-scoring application without scored essays as the training data.

In the first phase, the voting algorithm could be applied to the essays to determine the essays' initial scores. The second phase could include other natural language processing (NLP) or information retrieval (IR) techniques to adjust the scores.

The attack experiments show that it is not easy to fool the system unless the users use the terms appearing in high-scoring essays. Currently, the limitation of this approach is that the essays must be on the same topic. In addition, the bag-of-words model makes it inapplicable to creative writing essays.

THE AUTHORS

Yen-Yu Chen is an associate engineer in the Information and Communications Research Laboratories at the Industrial Technology Research Institute, Taiwan. His research interests include artificial intelligence, natural language processing, and automated essay scoring. Chen has an MS in computer science from National Chiao Tung University. Contact him at chenyy@itri.org.tw.

Chien-Liang Liu is a postdoc in the Department of Computer Science at National Chiao Tung University, Taiwan. His research interests include machine learning, natural language processing, and data mining. Liu has a PhD in computer science from National Chiao Tung University. Contact him at clliu@mail.nctu.edu.tw.

Chia-Hoang Lee is a professor in the Department of Computer Science and a senior vice president at National Chiao Tung University, Taiwan. His research interests include artificial intelligence, human–machine interface systems, and natural language processing. Lee has a PhD in computer science from the University of Maryland, College Park. Contact him at chl@cs.nctu.edu.tw.

Tao-Hsing Chang is an assistant professor in the Department of Computer Science and Information Engineering at National Kaohsiung University of Applied Sciences, Taiwan. His research interests include artificial intelligence in education, natural language processing, and automated essay scoring. Chang has a PhD in computer science from National Chiao Tung University. Contact him at changth@cc.kuas.edu.tw.

Acknowledgments

The data we analyzed here were collected by the Research Center for Psychological and Educational Testing at National Taiwan Normal University. This work was supported in part by the National Science Council under grants NSC-98-2221-E-009-141 and NSC-98-2811-E-009-038.

References

1. J. Wang and M.S. Brown, "Automated Essay Scoring versus Human Scoring: A Comparative Study," *J. Technology, Learning and Assessment*, vol. 6, no. 2, 2007, http://escholarship.bc.edu/jtla/vol6/2.
2. C.T. Meadow, B.R. Boyce, and D.H. Kraft, *Text Information Retrieval Systems*, 2nd ed., Academic Press, 2000.