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

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.

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...
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. 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.

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. 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. 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 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.

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.

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. 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. 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

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.

**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**


Selected CS articles and columns are also available for free at [http://ComputingNow.computer.org](http://ComputingNow.computer.org).