AUTOMATED EVALUATION OF ESSAYS AND SHORT ANSWERS

Jill Burstein
Claudia Leacock
Richard Swartz

Automated Evaluation of Essays and Short Answers

Jill Burstein (jburstein@etstechnologies.com)
Claudia Leacock (cleacock@etstechnologies.com)
Richard Swartz (rswartz@etstechnologies.com)

ETS Technologies, Inc.
A Subsidiary of Educational Testing Service
Princeton, NJ
http://www.etstechnologies.com

Abstract

Essay questions designed to measure writing ability, along with open-ended questions requiring short answers, are highly-valued components of effective assessment programs, but the expense and logistics of scoring them reliably often present a barrier to their use. Extensive research and development efforts at Educational Testing Service (ETS) over the past several years (see http://www.ets.org/research/erater.html) in natural language processing have produced two applications with the potential to dramatically reduce the difficulties associated with scoring these types of assessments.

The first of these, e-rater™, is a software application designed to produce holistic scores for essays based on the features of effective writing that faculty readers typically use: organization, sentence structure, and content. The e-rater software is "trained" with sets of essays scored by faculty readers so that it can accurately "predict" the holistic score a reader would give to an essay. ETS implemented e-rater as part of the operational scoring process for the Graduate Management Admissions Test (GMAT) in 1999. Since then, over 750,000 GMAT essays have been scored, with e-rater and reader agreement rates consistently above 97%. The e-rater scoring capability is now available for use by institutions via the Internet through the Criterion Online Writing Evaluation service at http://www.etstechnologies.com/criterion. The service is being used for both instruction and assessment by middle schools, high schools, and colleges in the U.S.

ETS Technologies is also conducting research that explores the feasibility of automated scoring of short-answer content-based responses, such as those based on questions that appear in a textbook's chapter review section. If successful, this research has the potential to evolve into an automated scoring application that would be appropriate for evaluating short-answer constructed responses in online instruction and assessment applications in virtually all disciplines.

E-rater History and Design

Educational Testing Service (ETS) has pursued research in writing assessment since its founding in 1947. ETS administered the Naval Academy English Examination and the Foreign Service Examination as early as 1948 (ETS Annual Report, 1949-50), and the Advanced Placement (AP) essay exam was administered in Spring of 1956. Some of the earliest research in writing assessment (see Coward, 1950 and Huddleston, 1952) laid the foundation for holistic scoring which continues to be used by ETS for large-scale writing assessments.

Currently several large-scale assessment programs contain a writing measure: the Graduate Management Admissions Test (GMAT), the Test of English as a Foreign Language (TOEFL), the Graduate Record Examination (GRE), Professional Assessments for Beginning Teachers (PRAXIS), the College Board's SAT II Writing Test and Advanced Placement (AP) exam, and the College-Level Examination Program (CLEP) English and Writing Tests. Some of these tests have moved to computer-based delivery, including the GMAT AWA, TOEFL, and GRE. The migration to computer-based delivery of these tests, along with the collection of examinee essay data in digital form, has permitted the exploration and use of automated methods for generating essay scores.

In February 1999, ETS began to use *e-rater* for operational scoring of the GMAT Analytical Writing Assessment (see Burstein, et al and Kukich, 2000). The GMAT AWA has two test question types (prompts): the issue prompt and the argument prompts. Prior to the use of *e-rater*, both the paper-and-pencil, and initial computer-based versions of the GMAT AWA were scored by two human readers on a six-point holistic scale. A final score was assigned to an essay response based on the original two reader scores if these two scores differed by no more than one score point. If the two readers were discrepant by more than one point, a third reader score was introduced to resolve the final score.

Since February 1999, an *e-rater* score and one human reader assigned a score to an essay. Using the GMAT score resolution procedures for two human readers, if the *e-rater* and human reader scores differed by more than one-point, a second human reader resolved the discrepancy. Otherwise, if the *e-rater* and human reader score agreed within one-point, these two scores were used to compute the final score for the essay.

Since *e-rater* was made operational for GMAT AWA scoring, it has scored over 750,000 essays – approximately 375,00 essays per year. The reported discrepancy rate between *e-rater* and one human reader score has been less than three percent. This is comparable to the discrepancy rater between two human readers.

E-rater Design and Holistic Scoring

Holistic essay scoring has been researched since the 1960's (Godshalk, 1966) and departs from the traditional, analytical system of teaching and evaluating writing. In the holistic scoring approach, readers are told to read quickly for a total impression and to take into account all aspects of writing as specified in the scoring guide. The final score is based on the readers total impression (Conlan, 1980).

From *e-rater*'s inception, it has always been a goal that the features used by the system to assign an essay score be related to the holistic scoring guide features. Generally speaking, the scoring guide indicates that an essay that stays on the topic of the question, has a strong, coherent and well-organized argument structure, and displays a variety of word use and syntactic structure will receive a score at the higher end of the six-point scale (5 or 6). *E-rater* features include discourse structure, syntactic structure, and analysis of vocabulary usage (topical analysis).

Natural Language Processing (NLP) in *E-rater*

Natural language processing (NLP) is the application of computational methods to analyze characteristics of electronic files of text or speech. In this section, only text-based applications are discussed. Methods used are either statistical, or linguistic-based analyses of language features. NLP applications utilize tools such as syntactic parsers, to analyze the syntactic form of a text (Abney, 1996); discourse parsers, to analyze the discourse structure of a text (Marcu, 2000); lexical similarity measures, to analyze word use of a text (Salton, 1989).

E-rater and NLP

E-rater uses a corpus-based approach to model building. The corpus-based approach uses actual essay data to analyze the features in a sample of essay responses. This approach is in contrast to a theoretical approach in which feature analysis and linguistic rules might be hypothesized a priori based on the kinds of characteristics one might expect to find in the data sample – in this case, a corpus of first-draft, student essay responses.

When using a corpus-based approach to build NLP-based tools for text analysis, researchers and developers typically use copyedited text sources. The corpora often used include text from newspapers, such as the Wall Street Journal, or the Brown

corpus, which contains 1 million words of text across genres (for example, newspapers, magazines, excerpts from novels, and technical reports). For instance, an NLP tool known as a part-of-speech tagger (Brill, 1997) is designed to label each word in a text with its correct part-of-speech (e.g., Noun, Verb, Preposition). Text that has been automatically tagged (labeled) with part-of-speech identifiers can be used to develop other tools, such as syntactic parsers, in which the part-of-speech tagged text is used to generate whole syntactic constituents. These constituents detail how words are connected into larger syntactic units, such as noun phrases, verb phrases, and complete sentences. The rules that are used in part-of-speech taggers to determine how to label a word are developed from copyedited text sources such as those mentioned above. By contrast, *e-rater* feature analysis and model building (described below) are based on unedited text corpora representing the specific genre of first-draft essay writing.

E-rater Details: Essay Feature Analysis and Scoring

The *e-rater* application currently has five main independent modules. The application is designed to identify features in the text that reflect writing qualities specified in human reader scoring criteria. The system has three independent modules for identifying scoring guide relevant features from the following categories: syntax, discourse, and topic. Each of the feature recognition modules described below identifies features that correspond to scoring guide criteria features which can be correlated to essay score, namely, *syntactic variety*, *organization of ideas*, and *vocabulary usage.* *E-rater* uses a fourth independent model building module to select and weight predictive features for essay scoring. The model building module reconfigures the feature selections and associated regression weightings given a sample of human reader scored essays for a particular test question. A fifth module is used for final score assignment.

Syntactic Module

E-rater's syntactic analyzer (parser) works in the following way to identify syntactic constructions in essay text.[1] *E-rater* tags each word for part-of-speech (Brill, 1997), uses a syntactic "chunker" (Abney, 1996) to find phrases, and assembles the phrases into trees based on subcategorization information for verbs (Grishman, et al, 1994). The parser identifies various clauses, including infinitive, complement, and subordinate clauses. The ability to identify such clause types allows *e-rater* to capture *syntactic variety* in an essay.

Discourse Module

[1] The parser used in *e-rater* was designed by Claudia Leacock, Tom Morton and Hoa Trang Dang.

E-rater identifies discourse cue words, terms, and syntactic structures, and these are used to annotate each essay according to a discourse classification schema (Quirk, et al, 1985). The syntactic structures, such as complement clauses, are outputs from the syntactic module described in Section 2.3.1. Such syntactic structures are used to identify, for example, the beginning of a new argument based on their position within a sentence and within a paragraph.

Generally, *e-rater's* discourse annotations denote the beginnings of arguments (the main points of discussion), or argument development within a text, as well as the classification of discourse relations associated with the argument type (e.g., *parallel relation*). Discourse features based on the annotations have been shown to predict the holistic scores that human readers assign to essays, and can be associated with *organization of ideas* in an essay.

E-rater uses the discourse annotations to partition essays into separate arguments. These argument partitioned versions of essays are used by the topical analysis module to evaluate the content individual arguments (Burstein, et al, 1998; Burstein & Chodorow, 1999). *E-rater's* discourse analysis produces a flat, linear sequence of units. For instance, in the essay text *e-rater*'s discourse annotation indicates that a contrast relationship exists, based on discourse cue words, such as *however*. Hierarchical discourse-based relationships showing intersentential relationships are not specified. Other discourse analyzers do indicate such relationships (Marcu, 2000).

Topical Analysis Module

Vocabulary usage is another criterion listed in human reader scoring guides. To capture use of vocabulary, or identification of topic *e-rater* includes a topical analysis module. The procedures in this module are based on the vector-space model, commonly found in information retrieval applications (Salton, 1989). These analyses are done at the level of the essay (big bag of words) and the argument.

For both levels of analysis, training essays are converted into vectors of word frequencies, and the frequencies are then transformed into word weights. These weight vectors populate the training space. To score a test essay, it is converted into a weight vector, and a search is conducted to find the training vectors most similar to it, as measured by the cosine between the test and training vectors. The closest matches among the training set are used to assign a score to the test essay**.**

As already mentioned*, e-rater* uses two different forms of the general procedure sketched above. For looking at topical analysis at the essay level, each of the training essays (also used for training *e-rater*) is represented by a separate vector in the training space. The score assigned to the test essay is a weighted mean of the scores for the 6 training essays whose vectors are closest to the vector of the test essay.

In the method used to analyze topical analysis at the argument level, all of the training essays are combined for each score category to populate the training space with just 6 "supervectors", one each for scores 1-6. The argument partitioned version of the essays generated from the discourse module are used in the set of test essays. Each test essay is evaluated one argument at a time. Each argument is converted into a vector of word weights and compared to the 6 vectors in the training space. The closest vector is found and its score is assigned to the argument. This process continues until all the arguments have been assigned a score. The overall score for the test essay is based on a mean of the scores for all arguments (see Burstein and Marcu, 2000 for details).

Model Building

The syntactic, discourse, and topical analysis modules each yield feature information that can be used for model building, and essay scoring. As mentioned earlier, a corpus-based linguistics approach is used for *e-rater* model building. To build models, a training set of human scored sample essays is collected that is representative of the range of scores in the scoring guide. As discussed earlier, this type of essay is generally scored on a 6-point scale, where a "6" indicates the score assigned to the most competent writer, and a score of "1" indicates the score assigned to the least competent writer. Optimal training set samples contain 265 essays that have been scored by two human readers.[2] The data sample is distributed in the following way with respect to score points: 15 1's, and 50 in each of the score points 2 through 6.

The model building module is a program that runs a forward-entry stepwise regression. Syntactic, discourse, and topical analysis information for the model building sample (training) are used as input to the regression program. This regression program automatically selects the features that are predictive for a given set of training data based on one test question. The program outputs the predictive features and their associated regression weightings. This output composes the model that is then used for scoring.

Scoring

In an independent scoring module, a linear equation is used to compute the final essay score. To compute the final score for each essay, the sum of the product of each regression weighting and its associated feature integer is calculated.

[2] E-rater models have been successfully built and used in operational scoring in *Criterion* with training set sizes smaller than the optimal 265 essays. ETS Technologies continues research to reduce the training set sizes required for model building, so that increasingly more test questions can be introduced into *Criterion*.

*Criterion*SM On-line Writing Evaluation Service: *E-rater* for Different Writing Levels

E-rater is currently embedded in Criterion, an on-line essay evaluation product of ETS Technologies, Inc., a subsidiary of ETS. The Criterion version of e-rater is web-based. This essay evaluation system is being used by institutions for writing assessment, and for classroom instruction. Using a web-based, real-time version of the system, instructors and students can see the *e-rater* score for an essay response within seconds. In this application, essay responses receive only an *e-rater* score.

Our current research in automated essay scoring has indicated that *e-rater* performs comparably to human readers at different grade levels. *Criterion* has scoring *e-rater* models based on prompts and data samples for grades 4, 8, and 12, using national standards prompts; for undergraduates, using English Proficiency Test (EPT) and PRAXIS prompts; and, for non-native English speakers, using TOEFL prompts. Both the TOEFL and GMAT programs are currently using *Criterion* for low-stakes, practice tests.

E-rater Targeted Advisories

Since one of *Criterion*'s primary functions is to serve as an instructional tool, we have also developed a feedback component that is referred to as the *advisory component*.[3] The advisories are generated based on statistical measures that evaluate word usage in essay responses in relation to the stimuli, and a sample of essay responses to a test question. The advisories are completely independent from the *e-rater* score, and only provide additional feedback about qualities of writing related to topic and fluency.

This advisory component includes feedback to indicate the following qualities of an essay response: a) the text is too brief to be a complete essay (suggesting that the student write more), b) the essay text does not resemble other essay written about the topic (so implying that perhaps the essay is *off-topic*), and c) the essay response is overly repetitive (suggesting that the student use more synonyms).

Summary and Future Directions

The current *e-rater* scoring technology can score essays at a number of different grade levels: elementary school, middle school, high school, college, and graduate school. The technology bases its scoring decisions on samples of data scored by human reader experts. *E-rater* uses NLP tools and statistical techniques to model the expert scoring decisions, and score test-taker essays. To date, *e-rater* has scored over 750,000 high-stakes essays since it began scoring GMAT essays in early 1999. *E-rater*

[3] This advisory component was designed and implemented by Martin Chodorow and Chi Lu.

scores show only a three percent discrepancy rate with a single human reader. This is the same discrepancy rate that occurs between two single human readers. In the *Criterion* application, in addition to a numerical score, *e-rater* generates a number of advisories that provide test-takers with information related to brevity, repetitiveness of response, and off-topicness of responses.

As indicated throughout this paper, the success and acceptance of automated essay scoring has permitted it to become integrated into many on-line writing assessments. In addition to a numerical, holistic rating, systems for evaluating writing need to provide feedback that reflects characteristics specific to each individual's writing. Such feedback can be used by students to help them in the essay revision process – thereby allowing them to develop their first-draft to more refined writing. There are many factors that contribute to overall improvement of developing writers. These factors include, for example, refined sentence structure, variety of appropriate word usage, and organizational structure.

Some of our current implementations with regard to *grammatical feedback* include: the identification of sentence types, such as simple and complex sentences, and sentence fragments; confusable word usage errors, such as between affect and effect, and who's and whose; and, grammar errors, such as the use of 'should of,' instead of 'should have' (Chodorow and Leacock, 2000).

The improvement of organizational structure is believed to be critical in the essay revision process toward overall improvement of essay quality. Therefore, it would be desirable to have a system *discourse-based feedback*. Such a system could present to students a guided list of questions to consider about the quality of the discourse. For instance, it has been suggested by writing experts that if the *thesis statemen*t of a student's essay could be automatically provided, the student could then use this information to reflect on the thesis statement and its quality. In addition, such an instructional application could utilize the thesis statement to discuss other types of discourse elements in the essay, such as the relationship between the *thesis statement* and the *conclusion*, and the connection between the *thesis statement* and the *main points* in the essay.

In the teaching of writing, students are often presented with a 'Revision Checklist.' The 'Revision Checklist' is intended to facilitate the revision process. This is a list of questions posed to the student that help the student reflect on the quality of their writing. So, for instance, such a list might pose questions as in the following. a) *Is the intention of my thesis statement clear?,* b) *Does my thesis statement respond directly to the essay question?,* c) *Are the main points in my essay clearly stated?,* d) *Do the main points in my essay relate to my original thesis statement?* If these questions are expressed in general terms, they are of little help; to be useful, they need to be grounded and need to refer explicitly to the essays students write (Scardamalia and Bereiter, 1985; White 1994). The ability to automatically identify, and present to students the discourse elements in their essays can help them to focus and reflect on

the critical discourse structure of the essay. In addition, the ability for the application to indicate to the student that a discourse element could not be located, perhaps due to the 'lack of clarity' of this element could also be helpful. Assuming that such a capability was reliable, this would force the writer to think about the clarity of a given discourse element, such as a thesis statement.

Currently, we have implemented software that automatically identifies essay-based discourse elements in student essays. The current version of the application can identify the thesis statement, topic sentence of each main idea, idea development, and concluding statement of an essay. Current evaluations indicate that the selections by the algorithm agree exactly with a human judge on the selection of a particular discourse element (such as thesis statement), as often as two human judges agree with each other (Burstein, Marcu, Andreyev, and Chodorow, submitted). This software is being developed into an application for student essay revision. The current essay evaluation technologies, that is essay scoring, combined with feedback about features of writing could give students an tremendous opportunity to spend more time practicing writing, and developing their writing skills.

C-rater™

An additional area of inquiry for ETS Technologies is the feasibility of automating the scoring of short answer content-based questions such as those that appear in a textbook's chapter review section. To date, we have developed an automated scoring prototype, c-rater™, using natural language processing technology, and evaluated its effectiveness at producing "credit/no credit" ratings (see Leacock and Chodorow, 2000). Results of an initial, small-scale study with a university virtual learning program were encouraging: c-rater achieved over 80 percent agreement with the score assigned by an instructor. This research has the potential to evolve into an automated scoring application that would be appropriate for evaluating user-constructed responses in online instruction and assessment applications.

C-rater is related to e-rater in that it uses many of the same natural language processing tools and techniques, but the two differ in some important ways.

- Holistic scoring versus content scoring: E-rater assigns a holistic score. That is, it assigns a score for writing skills rather than for specific content. There is no correct answer in a holistic scoring rubric, only a description of how to identify good writing. Concept-rater needs to score a response as being either correct or incorrect and to do this, it must identify whether a response contains specific information in the form of some particular concepts. If the response expresses these concepts it is correct, and if it does not, it is incorrect, without regard to writing skills.
- Rhetorical structure versus predicate-argument structure: E-rater identifies, and gives a grade based, in part, on the rhetorical structure of an essay. Rhetorical structure shapes and organizes the main points of the essay. Concept-rater, on the

other hand, needs to identify specific content. In order to do this, it generates a fine-grained analysis of the predicate-argument structure, or logical relations between the syntactic components (e.g. subject, verb, object) for each sentence in the response.

- Training materials: E-rater is trained on a collection of 270 essays that have been manually scored by trained human raters. Concept-rater does not require a large collection of graded answers for training. Instead, it uses the single correct answer that is found in an instructor's guide or answer key. C-rater takes this approach because it is unrealistic to require extensive data collection for the purpose of grading relatively low stakes quizzes, especially given that there is often a set of short questions at the end of each chapter in a textbook.

Conclusion

The value and effectiveness of the e-rater automated essay scoring technology has been well demonstrated over the past two years through its use as part of the operational scoring process for the GMAT Analytical Writing Assessment. As the volume of online student writing increases, and as the quality of natural language processing tools and technologies improves, the quality and utility of e-rater feedback will continue to evolve and improve.

The ability to use automatic essay scoring in operational scoring environments reduces the time and costs associated with having multiple human readers score essay responses. The agreement between two human readers, and between *e-rater* and one human reader has been noted to be comparable (Burstein, et al 1998). *E-rater* scores are comparable to human reader scores, and automated scoring procedures can reduce the time and costs involved with manual essay scoring. Therefore, automated essay scoring would appear to be a favorable solution toward the introduction of more writing assessments on high-stakes standardized tests, and in a lower stakes environment -- for classroom instruction. Moreover, the availability of these technologies may well provide incentive for making more assessment and instructional materials available online.

References

Abney, Steven. (1996) Part-of-speech tagging and partial parsing. In Church, Young and Bloothooft (eds), *Corpus-based Methods in Language and Speech*. Dordrecht: Kluwer.

Brill, E. (to appear). Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging, Natural Language Processing Using Very Large Corpora. Dordrecht: Kluwer Academic Press.

Burstein, J., Marcu, D., Andreyev, S., and Chodorow, M. (submitted). Towards Automatic Classification of Discourse Elements in Essays. To appear in Proceedings of the Annual Meeting of the Association for Computational Linguistics, Toulouse, France, July, 2001.

Burstein, Jill and Daniel Marcu (2000). Towards Using Text Summarization for Essay-Based Feedback. Le 7e Conference Annuelle sur Le Traitement Automatique des Langues Naturelles TALN'2000, Lausanne, Switzerland, pp: 51-59.

Burstein, J., Kukich, K. Wolff, S. Lu, C. Chodorow, M, Braden-Harder, L. and Harris M.D. (1998). Automated Scoring Using A Hybrid Feature Identification Technique. *Proceedings of ACL,* 206-210.

Chodorow, M. and Leacock, C. (2000). An unsupervised method for detecting grammatical errors. In Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics, 140-147.

Conlan, G.(1980). Comparison of Analytic and Holistic Scoring. Unpublished Report.

Coward, Ann F. (1950). The Method of reading the Foreign Service Examination in English Composition. ETS RB-50-57, Educational Testing Service, Princeton, NJ.

ETS Annual Report, 1949-50, Educational Testing Service, Princeton, NJ.

Godshalk, Fred I., Swineford, F. and Coffman, W.E. (1966). The Measurement of Writing Ability. New York, NY. College Entrance Exam Board.

Grishman, R., Macleod, C., and Meyers, A. (1994). "COMLEX Syntax: Building a Computational Lexicon", Proceedings of Coling, Kyoto, Japan. (available for download at:http://cs.nyu.edu/cs/projects/proteus/comlex/)

Huddleston, Edith M. (1952). Measurement of Writing Ability at the College-Entrance Level: Objective Vs. Subjective Testing Techniques. ETS RB-52-57.

Jing, Hongyan and McKeown, K. (2000). Cut and Paste Text Summarization, In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pp. 178-185.

Knight, K. (1997). Automating Knowledge Acquisition for Machine Translation. AI Magazine 18(4).

Kukich, K. (2000). Beyond Automated Essay Scoring. IEEE Intelligent Systems, pp: 22-27.

Leacock, Claudia and Chodorow, M. (2000). Automated Scoring of Short-Answer Responses. ETS Technologies Research Report.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization.* The MIT Press.

Quirk, R., Greenbaum, S., Leech, S., and Svartik, J. (1985). A Comprehensive Grammar of the English Language. Longman, New York.

Salton, Gerard. (1989). Automatic text processing : the transformation, analysis, and retrieval of information by computer. Addison-Wesley, Reading, Mass.

Scardamalia, M. and C. Bereiter (1985). Development of Dialectical Processes in Composition. In Olson, D. R., Torrance, N. and Hildyard, A. (eds), *Literacy, Language, and Learning: The nature of consequences of reading and writing.* Cambridge University Press.

Teufel, S. and, Moens, M. (1999): Argumentative classification of extracted sentences as a first step towards flexible abstracting. In: I. Mani, M. Maybury (eds.), *Advances in automatic text summarization*, MIT Press.

White E.M. (1994). Teaching and Assessing Writing. Jossey-Bass Publishers, 103-108.