Automated Essay Scoring for Nonnative English Speakers

Jill Burstein
Educational Testing Service
Princeton, New Jersey 08540
jburstein@ets.org

Martin Chodorow
Hunter College, CUNY
New York, New York
martin.chodorow@hunter.cuny.edu

Abstract

The *e-rater* system^{TM} [1] is an operational automated essay scoring system, developed at Educational Testing Service (ETS). The average agreement between human readers, and between independent human readers and *e-rater* is approximately 92%. There is much interest in the larger writing community in examining the system's performance on nonnative speaker essays. This paper focuses on results of a study that show *e-rater*'s performance on Test of Written English (TWE) essay responses written by nonnative English speakers whose native language is Chinese, Arabic, or Spanish. In addition, one small sample of the data is from US-born English speakers, and another is from non-US-born candidates who report that their native language is English. As expected, significant differences were found among the scores of the English groups and the nonnative speakers. While there were also differences between *e-rater* and the human readers for the various language groups, the average agreement rate was as high as operational agreement. At least four of the five features that are included in *e-rater*'s current operational models (including discourse, topical, and syntactic features) also appear in the TWE models. This suggests that the features generalize well over a wide range of linguistic variation, as *e-rater* was not confounded by non-standard English syntactic structures or stylistic discourse structures which one might expect to be a problem for a system designed to evaluate native speaker writing.

Introduction

Research and development in automated essay scoring has begun to flourish in the past five years or so, bringing about a whole new field of interest to the NLP community (Burstein, et al (1998a, 1998b and 1998c), Foltz, et al (1998), Larkey (1998), Page and Peterson (1995)). Research at Educational Testing Service (ETS) has led to the recent development of *e-rater*, an operational automated essay scoring system. *E-rate*r is based on features in holistic scoring guides for human reader scoring. Scoring guides have a 6-point score scale. Six's are assigned to the "best" essays, and "1's" to the least well-written. Scoring guide criteria are based on structural (syntax and discourse) and vocabulary usage in essay responses (see http://www.gmat.org).

E-rater builds new models for each topic (prompt-specific models) by evaluating approximately 52 syntactic, discourse and topical analysis variables for 270 human reader scored training essays. Relevant features for each model are based on the predictive feature set identified by a stepwise linear regression. In operational scoring, when compared to a human reader,

[1] The *e-rater* system^{TM} is a trademark of Educational Testing Service. In the paper, we will refer to the *e-rater* system^{TM} as *e-rater*.

e-rater assigns an exactly matching or adjacent score (on the 6-point scale) about 92% of the time. This is the same as the agreement rate typically found between two human readers. Correlations between *e-rater* scores and those of a single human reader are about .73; correlations between two human readers are .75.

The scoring guide criteria assume standard written English. Non-standard English may show up in the writing of native English speakers of non-standard dialects. For general NLP research purposes, it is useful to have computer-based corpora that represent language variation (Biber (1993)). Such corpora allow us to explore issues with regard to how the system will handle responses that might be written in non-standard English. Current research at ETS for the Graduate Record Examination (GRE) (Burstein, et al, 1999) is making use of essay corpora that represent subgroups where variations in standard written English might be found, such as in the writing of African Americans, Latinos and Asians (Breland, et al (1995) and Bridgeman and McHale (1996)). In addition, ETS is accumulating essay corpora of nonnative speakers that can be used for research.

This paper focuses on preliminary data that show *e-rater*'s performance on Test of Written English (TWE) essay responses written by nonnative English speakers whose native language is Chinese, Arabic, or Spanish. A small sample of the data is from US-born English speakers and a second small sample is from non-US-born candidates who report that their native language is English. The data were originally collected for a study by Frase, et al (1997) in which analyses of the essays are also discussed. The current work is only the beginning of a program of research at ETS that will examine automated scoring for nonnative English speakers. Overall goals include determining how features used in automated scoring may also be used to (a) examine the difficulty of an essay question for speakers of particular language groups, and (b) automatically formulate diagnostics and instruction for nonnative English speakers, with customization for different language groups.

1. E-rater Feature Identification, Model Building and Scoring

The driving concept that underlies *e-rater* is that it needs to evaluate the same kinds of features that human readers do. This is why from the beginning of its development, we made it a priority to use features from the scoring guide and to eliminate any direct measures of essay length. Even though length measures can be shown to be highly correlated with human reader essay scores, length variables are not scoring guide criteria (Page and Peterson, 1995). The features currently used by the system are syntactic features, discourse cue words, terms and structures, and topical analysis, specifically, vocabulary usage at the level of the essay (big bag of words) and at the level of the argument. Argument, in this case, refers generally to the different discussion points made by the writer.

1.1 Syntactic Structure and Syntactic Variety

The holistic rubric criteria specify that the *syntactic variety* used by a candidate should be considered with regard to essay score. The *e-rater* system uses an ETS-enhanced version of the CASS syntactic chunker (Abney (1996)), referred to here as the parser. The parser identifies several syntactic structures in the essay responses, such as subjunctive auxiliary verbs (e.g., would, should, might), and complex clausal structures, such as complement, infinitive, and subordinate clauses. Recognition of such features in an essay yields information about its syntactic variety.

1.2 Discourse Cues and Organization of Ideas

Organization of ideas is another criterion that the scoring guide asks human readers to consider in assigning essay score. *E-rater* contains a lexicon based on the conceptual framework of conjunctive relations from Quirk, et al (1985) in which cue terms, such as "In summary" and "In conclusion," are classified as conjuncts used for summarizing. The conjunct classifiers contain information about whether or not the item is a kind of discourse development term (e.g., "for example" and "because"), or whether it is more likely to be used to begin a discourse statement (e.g., First, Second, or Third*).* *E-rater* also contains heuristics that define the syntactic or essay-based structures in which these terms must appear to be considered as discourse markers. For example, for the word "first" to be considered a discourse marker, it must not be a nominal modifier, as in the sentence, "The first time I went to Europe was in 1982," in which "first" modifies the noun "time." Instead, "first" must occur as an adverbial conjunct to be considered a discourse marker, as in the sentence, "First, it has often been noted that length is highly correlated with essay score." The cue term lexicon and the associated heuristics are used by *e-rater* to automatically annotate a high-level discourse structure of each essay. These annotations are also used by the system to partition each essay into separate arguments which are input to the system's topical analysis component, described below, for analyzing topical content.

1.3 Topical Analysis and Vocabulary Usage

Vocabulary usage is listed as another criterion on human reader scoring guides. Good essays are relevant to the assigned topic. They also tend to use a more specialized and precise vocabulary in discussing the topic than poorer essays do. We should therefore expect a good essay to resemble other good essays in its choice of words and, conversely, a poor essay to resemble other poor ones. To capture use of vocabulary or identification of topic, *e-rater* uses content vector analyses that are based on the vector-space model commonly found in information retrieval applications.

Training essays are converted into vectors of word frequencies, and the frequencies are then transformed into word weights.[2] 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.

E-rater uses two different forms of the general procedure sketched above. In one form, for looking at topical analysis at the essay level, each of the 270 training essays 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. This

[2] Word (or term) weight reflects not only a word's frequency in the essay but also its distribution across essays. E-rater's formula for the weight of word *w* in essay *j* is:

$$\text{weight}_{wj}=(\text{freq}_{wj}/\text{maxfreq}_{j}) * \log(\text{nessays}/\text{essays}_{w})$$

where freq_{wj} is the frequency of word *w* in essay *j*, maxfreq_{j} is the frequency of the most frequent word in essay *j*, nessays is the total number of training essays, and essays_{w} is the number of training essays that contain *w*. The first part of the formula measures the relative importance of the word in the essay. The second part gauges its specificity across essays, so that a word that appears in many essays will have a lower weight than one which appears in only a few. In the extreme case, a word that appears in all essays (e.g., "the") has a weight of 0.

score is computed using the following formula, rounded to the nearest integer:

Score for test essay t =

$$\Sigma(\text{cosine}_{tj} * \text{score}_j)/\Sigma\ \text{cosine}_{tj}$$

where j ranges over the 6 closest training essays, score_j is the human rater score for training essay j, and cosine_{tj} is the cosine between test essay t and training essay j.

The other form of content vector analysis that the system uses combines all of the training essays for each score category and populates the training space with just 6 "supervectors", one each for scores 1-6. This method is used to evaluate the vocabulary usage at the argument level. The 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 an adjusted mean of the argument scores using the following formula, rounded to the nearest integer:

Score for test essay t =

$$(\Sigma\text{argscore}_j + \text{nargs}_t)/(\text{nargs}_t + 1)$$

where j ranges over the arguments in test essay t, argscore_j is the score of argument j, and nargs_t is the number of arguments in t. Using this adjusted mean has the overall effect of reducing, slightly, the score for essays with few arguments, and of increasing somewhat the score of essays with many arguments.

2. Model Building and Essay Scoring

E-rater builds a new model for each test question (prompt). In pre-operational trials, a set of 270 essays scored by at least two human readers has been shown to be optimal for training. The distribution at each score point in the 270 training essays is as follows: five 0's, fifteen 1's, and fifty 2's through 6's.[3]

The syntactic, discourse, and topical analysis features are identified for each of the 270 essays. Vectors of raw counts of occurrences of syntactic and discourse structure information, and scores generated for the two topical analysis components are submitted to a stepwise linear regression. For each prompt, the regression selects the subset of predictive features. Typically, 8 to 12 features are selected. Although every model has a different combination of features, in the 75 models that we are currently running, the five most frequently occurring features are: 1) the topical analysis score by argument, 2) the topical analysis score by essay, 3) the number of subjunctive auxiliary words, 4) the ratio of subjunctive auxiliary words to total words in the essay, and 5) the total number of argument development terms.

The coefficient weightings for each of the predictive features generated from the regression for each prompt are then used to score new essays for that prompt.

3. *E-rater* Agreement Performance on Nonnative Speaker Data

Some questions that will now be addressed in looking at *e-rater* system performance on nonnative speaker essay data are: (1) How does performance for nonnative speakers on TWE compare with performance in operational scoring? (2) How does the system's agreement with human readers differ for each of the language groups in this

[3] To date, this training sample composition has given us the best cross-validation results. Some previous studies experimenting with smaller training samples with this fairly flat distribution, or samples which reflect more directly the natural distribution of the data at each score point have shown lower performance in scoring cross-validation sets of 500 – 900 essays.

study? (3) How does *e-rater*'s agreement with human readers differ for the nonnative speaker language groups as compared to the English speaking language groups? (4) Is there a significant difference between the features used most often in models for operational prompts as compared to the TWE prompts?

3.1 Data sample

For this study, two prompts from the Test of Written English were used. These prompts (TWE1 and TWE2) ask candidates to read and think about a statement, and then to agree or disagree with the statement, and to give reasons to support the opinion given by the candidate. The scoring guides for these essays have a 6-point scale, where a "6" is the highest score and a "1" is the lowest score. They are holistic guides, though the criteria are more generally stated than in the scoring guides used to build *e-rater*.

For each of the prompts a total of 255 essays were used for training. Fifty training essays were randomly selected from each of the score categories 2-6. Because of the small number of essays with a score of 1, only five 1's were included in each training set. The remainder of the essays were used for cross-validation purposes.

4. Results

Tables 1-3 show overall and language specific scoring results for TWE1 and TWE2 cross-validation data. The data are presented in terms of mean score and also as percent agreement between *e-rater* and human readers, where agreement is defined as exactly matching or adjacent scores on the 6-point scale. In previous studies of holistically scored essays (Burstein, et al (1998a, 1998b and 1998c)), we have examined *e-rater*'s agreement with two individual human readers. For these TWE data, only a final human reader score (labeled GDF in the Tables) was available. The final score reflects the average of two or three human reader scores. A third human reader is typically used if the first two humans disagree by more that a single point. For the operational essay data, the mean agreement between *e-rater* and the final human reader score is 90%, about the same as the mean agreement between two individual human readers at about 92%.[4] For the same data, Pearson correlations between *e-rater* and final human reader scores, and between two human readers are about the same at .75. Table 1 shows that, for TWE essays, overall *e-rater* agreement with the human reader final score is high. The values are comparable to those for the operational essays although the correlations are somewhat lower.

[4] Baseline agreement for the TWE data is approximately 84%. This is determined by calculating agreement if the most common score, "4", is assigned to all essays. Using the same technique for GMAT essays showed baseline agreement to be about 83%.

Table 1: Comparison of Human Readers Final Score (GDF) & *e-rater* Score (E) Over All Language Groups in TWE1 and TWE2

Prompt	n=	%Agreement (Exact+Adjacent)	Pearson r	GDF Score		E Score	
				Mean	S.D.	Mean	S.D.
TWE1	562	91.1	.667	4.16	.974	4.08	1.041
TWE2	576	93.4	.718	4.16	.936	4.07	.989
Mean		**92.3**	**.693**	**4.16**	**.955**	**4.08**	**1.015**

An analysis of variance was performed on the essay scores, using Reader (GDF, E) as a within factor and Prompt (TWE1, TWE2) and Language Group as between factors. Although small, the difference in mean score between GDF and *e-rater* was statistically significant ($F_{(1,1128)}$ = 5.469, p < .05). There was no significant main effect for Prompt, and no interactions between Prompt and the other factors. Tables 2 and 3 show the results for TWE1 and TWE2 by Language Group and Reader.

Table 2: Comparison of Human Readers Final Score (GDF) & *e-rater* Score (E) By Language Groups in TWE1

Language Group	n=	%Agreement (Exact+Adjacent)	Pearson r	GDF Score		E Score	
				Mean	S.D.	Mean	S.D.
Arabic	146	89.0	.645	3.83	.973	3.67	.947
Chinese	153	88.2	.543	4.09	.884	4.12	1.00
Spanish	131	92.4	.644	3.96	.986	3.70	.915
US-English	97	96.9	.632	4.96	.624	4.93	.814
Non-US English	35	91.4	.544	4.31	.900	4.51	.981

Table 3: Comparison of Human Readers Final Score (GDF) & *e-rater* Score (E) By Language Groups in TWE2

Language Group	n=	%Agreement (Exact+Adjacent)	Pearson r	GDF Score		E Score	
				Mean	S.D.	Mean	S.D.
Arabic	151	96.4	.783	3.85	.959	3.70	.909
Chinese	139	91.0	.707	3.92	.957	4.04	1.03
Spanish	138	93.5	.616	4.07	.845	3.69	.733
US-English	103	92.0	.519	4.83	.613	4.95	.759
Non-US English	45	93.3	.465	4.68	.732	4.60	.780

The main effect for Language Group was significant ($F_{(4,1128)}$ = 76.561, p < .001). As expected, the two English groups scored substantially higher than the nonnative speakers. Finally, the interaction of Language Group by Reader was also significant ($F_{(4,1128)}$ = 12.397, p < .001), reflecting higher scores for GDF than for *e-rater* in some groups (e.g., Spanish) and lower scores for GDF than for *e-rater* in others (e.g., Chinese).

Despite the score differences, χ^2 analyses showed no significant differences on the Agreement measure for Language Group in either TWE1 or TWE2. There was however an effect of Prompt in the analysis of Agreement for Arabic speakers, where Agreement levels in TWE1 and TWE2 were significantly different ($\chi^2(1) = 6.607$, $p < .01$); no other group differences in Agreement were found between the two prompts.

5. Discussion and Conclusions

In this study we have evaluated the performance and effects of *e-rater* on two sets of nonnative speaker essay responses, approximately 1100 essays. The results show that overall system performance is quite good and highly comparable to results for scoring the primarily native speaker data found in operational essays. The models that *e-rater* built to score TWE1 and TWE2 contain 7 or 8 features, and these include syntactic, discourse and topical analysis features. Importantly, at least 4 of the top 5 features that are included in the current operational models also appear in the models for TWE1 and TWE2. It is useful to know that even when 75% of essays used for model building were written by nonnative English speakers (as in this study), the features selected by the regression procedure were largely the same as those in models based on operational writing samples in which the majority of the sample were native English speakers. This suggests that the features that the system considers are generalizable from native speaker writing to nonnative speaker writing. Further, *e-rater* was not confounded by non-standard English syntactic structures or stylistic discourse structures, which one might expect to be a problem for a system designed to evaluate native speaker writing.

Although there were significant differences between final human reader score and *e-rater* score across language groups, in absolute terms the differences were small (only a fraction of a score point) and did not produce significant differences in agreement. For one group, prompt made a difference. It would be useful to analyze the essays in more detail to see what features are responsible for the score variations and how essay topic might explain any differences due to prompt. We are currently investigating the use of tree-based regression models to supplement linear regression (Sheehan, 1997). Preliminary analyses of tree-based regressions, however, do not show an improvement in e-rater performance. This may be explained by the fact that the most predictive features in *e-rater* are linearly related to score.

In future studies, we will have sufficient data to build individual models for different language groups to examine how this affects *e-rater*'s performance. In addition, we hope to learn about how building language-specific models can be used for automated generation of diagnostic and instructional feedback -- perhaps customized for different language groups.

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.

Biber, D. (1993) Using register-diversified corpora for general language studies. *Computational Linguistics*, 19, 219-241.

Breland, Bonner and Kubota (1995), Factors in Performance on Brief, Impromptu Essay

Examinations, College Board Report No. 95-4.

Bridgeman and McHale (1996). Gender and Ethnic Group Differences on the GMAT Analytical Writing Assessment. ETS RR-96-2.

Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris (1998). Automated Scoring Using A Hybrid Feature Identification Technique. In the *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, August, 1998. Montreal, Canada.

Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu, & Martin Chodorow (1998). Enriching Automated Scoring Using Discourse Marking. In the *Proceedings of the Workshop on Discourse Relations & Discourse Marking, Annual Meeting of the Association of Computational Linguistics*, August, 1998. Montreal, Canada.

Burstein, Jill C., Lisa Braden-Harder, Martin Chodorow, Shuyi Hua, Bruce Kaplan, Karen Kukich, Chi Lu, James Nolan, Don Rock and Susanne Wolff (1998). Computer Analysis of Essay Content for Automated Score Prediction. ETS RR 98-15.

Foltz, P. W., W. Kintsch, and T. K. Landauer. (1998). The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse Processes*, 25(2&3), 285-307.

Frase, L. T., Faletti, J., Ginther, A., & Grant, L. (1997). Computer Analysis of the TOEFL Test of Written English (TWE). Educational Testing Service, Princeton, NJ.

Larkey, L. (1998). Automatic Essay Grading Using Text Categorization Techniques. *Proceedings of the 21st ACM-SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 90-95.

Page, E. B. and N. Petersen. (1995). The computer moves into essay grading: updating the ancient test. *Phi Delta Kappan*. March, 561-565.

Sheehan, K (1997). A Tree-Based Approach to Proficiency Scaling and Diagnostic Assessment. *Journal of Educational Measurement*. 34(4), 333-352.

Acknowledgements

We owe considerable thanks to Karen Kukich, who designed the architecture for the operational version of *e-rater*, and who led the team at ETS that refined and implemented the production version of the system. The production version makes the process of continued research in automated scoring much smoother. Thanks to Susanne Wolff for her invaluable collaboration on the *e-rater* system from its inception. We are also grateful to Steve Abney for allowing us to use CASS in the operational version of *e-rater*; to Claudia Leacock for leading the design and development effort to create the ETS-enhanced version of CASS; to Thomas Morton and Hoa Trang Dang for the implementation of the ETS-enhanced version of CASS; and to Daniel Zuckerman, who made it possible to run CASS on both Unix and NT platforms. We thank Chi Lu for implementations and re-implementations of statistical programming embedded in *e-rater*, and Magdalena Wolska for the work she did implementing tools that allowed us to conduct this study with ease. Finally, we are grateful to ETS test developers for all of the dialogue that has helped us understand the educational direction of automated essay scoring, and all the people at ETS and GMAC who supported this research from its inception.