Enriching Automated Essay Scoring Using Discourse Marking

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu† and Martin Chodorow‡

† Educational Testing Service, Princeton NJ
‡ Hunter College, New York, New York

Abstract

Electronic Essay Rater (*e-rater*) is a prototype automated essay scoring system built at Educational Testing Service (ETS) that uses discourse marking, in addition to syntactic information and topical content vector analyses to automatically assign essay scores. This paper gives a general description of *e-rater* as a whole, but its emphasis is on the importance of discourse marking and argument partitioning for annotating the argument structure of an essay. We show comparisons between two content vector analysis programs used to predict scores, *EssayContent* and *ArgContent*. *EssayContent* assigns scores to essays by using a standard cosine correlation that treats the essay like a "bag of words," in that it does not consider word order. *ArgContent* employs a novel content vector analysis approach for score assignment based on the individual arguments in an essay. The average agreement between *ArgContent* scores and human rater scores is 82%, as compared to 69% agreement between *EssayContent* and the human raters. These results suggest that discourse marking enriches *e-rater's* scoring capability. When *e-rater uses* its whole set of predictive features, agreement with human rater scores ranges from 87% - 94% across the 15 sets of essay responses used in this study

1. Introduction

The development of Electronic Essay Rater (*e-rater*), an automated prototype essay scoring system, was motivated by practical concerns of time and costs that limit the number of essay questions on current standardized tests. Literature on automated essay scoring shows that reasonably high agreement can be achieved between a machine score and a human rater score simply by doing analyses based on the number of words in an essay (Page and Peterson (1995)). Scoring an essay based on the essay length is not a criterion that can be used to define competent writing. In addition, from a practical standpoint, essay length is a highly coachable feature. It doesn't take examinees long to figure out that a computer will assign a high score on an essay based on a pre-specified number of words.

E-rater's modules extract syntactic and discourse structure information from essays, as well as information about vocabulary content in order to predict the score. The 57 features included in *e-rater* are based on writing characteristics specified at each of the six score points in the scoring guide used by human raters for manual scoring (also available at http://www.gmat.org/). For example, the scoring guide indicates that an essay that stays on the topic of the test 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). Lower scores are assigned to essays as these characteristics diminish.

Included in *e-rater*'s feature set are features derived from discourse structure, syntactic structure, and topical analysis as they relate to the human scoring guide. For each essay question, *e-rater* is run on a set of training data (human-scored essay responses) to extract features. A stepwise linear regression analysis is performed on the features extracted from the training set to determine which ones have significant weights (the predictive features). Final score prediction for cross-validation sets is performed using these predictive features identified in the training sets. Accuracy is determined by measuring agreement between human rater assigned

scores and machine predicted scores, which are considered to "agree" if there is no greater than a single point difference on the six-point scale. This is the same criterion used to measure agreement between two human raters.

Among the strongest predictive features across the essay questions used in this study are the scores generated from *ArgContent* (a content vector analysis applied to discourse chunked text), and discourse-related surface cue word and non-lexical features. On average, *ArgContent* alone has 82% agreement with the human rater score as compared to *EssayContent*'s 69%. *EssayContent* is a content vector analysis program that treats an essay like a "bag of words." This suggests two things. First, the discourse markers detected by the argument annotation and partitioning program, *APA,* are helpful for identification of relevant units of discourse in essay responses. Second, the application of content vector analysis to those text units appears to increase scoring performance. Overall, it appears that discourse marking provides feature information that is useful in *e-rater*'s essay score predictions.

A long-term goal of automated essay scoring is to be able to generate diagnostic or instructional information, along with a numeric score to a test-taker or instructor. Information about the discourse structure of essays brings us closer to being able to generate informative feedback to test-takers about the essay's cohesion.

We report on the overall evaluation results from *e-rater*'s scoring performance on 13 sets of essay data from the Analytical Writing Assessments of the Graduate Management Admissions Test (GMAT) (see http://www.gmat.org/) and 2 sets of essay data from the Test of Written English (TWE) (see http://www.toefl.org/tstprpmt.html for sample TWE questions). The paper devotes special attention to *e-rater*'s discourse marking and analysis components.

2. Hybrid Feature Methodology

E-rater uses a hybrid feature approach in that it incorporates several variables that are derived statistically, or extracted through NLP techniques. The following sections describe the features used in this study.

2.1 Syntactic Features

The scoring guides indicate that one feature used to evaluate an essay is syntactic variety. Syntactic structures in essays are identified using NLP techniques. All sentences are parsed with the Microsoft Natural Language Processing tool (MSNLP) (see MSNLP (1997)). Examination of the parse trees yields information about syntactic variety with regard to what kinds of clauses or verb types were used by a test-taker.

A program was implemented to identify the number of complement clauses, subordinate clauses, infinitive clauses, relative clauses and occurrences of the subjunctive modal auxiliary verbs, *would*, *could*, *should*, *might* and *may*, for each sentence in an essay. Ratios of syntactic structure types per essay and per sentence were calculated as possible measures of syntactic variety.

2.2 Discourse Structure Analysis

GMAT essay questions are of two types: Analysis of an Issue (issue) and Analysis of an Argument (argument). The issue essay asks the writer to respond to a general question and to provide "reasons and/or examples" to support his or her position on an issue introduced by the test question. The argument essay focuses the writer on the argument in a given piece of text, using the term *argument* in the sense of a rational presentation of points with the purpose of persuading the reader. The scoring guides used for manual scoring indicate that an essay will receive a score based on the examinee's demonstration of a well-developed essay. For the argument essay, for instance, the scoring guide states that a "6" essay "develops ideas cogently, organizes them logically, and connects them with clear transitions." The correlate to this for the issue essay would appear to be that a "6" essay "...develops a position on the issue with insightful reasons..." and that the essay "is clearly well-organized." Nolan (1997) points out that terms in holistic scoring guides, such as "cogent," "logical," "insightful," and "well-organized" have "fuzzy" meaning, since they are based on imprecise observation. Nolan uses methods of "fuzzy logic" to automatically assign these kinds of "fuzzy" classifications to essays. In this study, we try to identify organization of an essay through automated

analysis and identification of the essay's argument structure through discourse marking.

Since there is no particular text unit that reliably corresponds to the stages, steps, or passages of an argument, readers of an essay must rely on other things such as surface cue words to identify individual arguments. We found that it was useful to identify rhetorical relations such as *Parallelism* and *Contrast*, and content or coherence relations that have more to do with the discourse involved. These relations can appear at almost any level -- phrase, sentence, a chunk consisting of several sentences, or paragraph. Therefore, we developed a program to automatically identify the discourse unit of text using surface cue words and non-lexical cues.

As literature in the field of discourse analysis points out, surface cue words and structures can be identified and used for computer-based discourse analysis (Cohen (1984), (Mann and Thompson (1988), Hovy, et al (1992), Hirschberg and Litman (1993), Vander Linden and Martin (1995), Knott (1996) and Litman (1996)). *E-rater's APA* module uses surface cue words and non-lexical cues (i.e., syntactic structures) to denote discourse structure in essays. We adapted the conceptual framework of conjunctive relations from Quirk, et al (1985) in which terms, such as "In summary" and "In conclusion," which we consider to be surface cue terms, are classified as conjuncts used for summarizing. Cue words such as "perhaps" and "possibly" are considered to be Belief words used by the writer to express a belief with regard to argument development in essays. Words like "this" and "these" may often be used to flag that the writer is developing on the same topic (Sidner (1986)). We also observed that, in certain discourse contexts, non-lexical, syntactic structure cues, such as infinitive or complement clauses, may characterize the beginning of a new argument.

The automated argument partitioning and annotation program (*APA*) was implemented to output a discourse-marked annotated version of each essay in which the discourse marking is used to indicate new arguments (arg_init), or development of an argument (arg_dev). An example of *APA* annotations is shown in Figure 1.

New Paragraph:
…

Sentence 1: *It is also assumed that shrinking high school enrollment may lead to a shortage of qualified engineers.*

arg_init#PARALLEL = *also*
arg_init#CLAIM_THAT = *that*
arg_aux#SPECULATE = *may*
…
Sentence 3: *It is conceivable that other programs such as arts, music or social sciences will be most affected by this drop in high school population.*
…
arg_dev#SAME_TOPIC = *It*
arg_dev#CLAIM_THAT = *that*
arg_dev#DETAIL = *such_as*

Figure 1: APA Output for 2 Essay Sentences

APA's heuristic rules for discourse marker annotation and argument partitioning are based on syntactic and paragraph-based distribution of surface cue words, phrases and non-lexical cues corresponding to discourse structure. Relevant cue words and terms are contained in a specialized surface cue word and phrase lexicon. In Figure 1, the annotations, arg_init#PARALLEL, and arg_dev#DETAIL indicate the rhetorical relations of Parallel structure and Detail information, respectively, in arguments. The arg_dev#SAME_TOPIC label denotes the pronoun "it" as indicating the writer has not changed topics. The labels arg_init#CLAIM_THAT and arg_dev#CLAIM_THAT indicate that a complement clause was used to flag a new argument, or argument development. Arg_aux#SPECULATE flags subjunctive modals that are believed to indicate a writer's speculation. Preliminary analysis of these rules indicates that some rule refinements might be useful; however, more research needs to be done on this.[1] Based on the arg_init flags in the annotated essays, *APA* outputs a version of the essay partitioned "by argument". The argument-partitioned versions of essays are input to *ArgContent*, the discourse-driven, topical analysis program described below.

2.3 Topical Analysis

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

[1] We thank Mary Dee Harris for her analysis of *APA* annotated outputs.

resemble other good essays in its choice of words and, conversely, a poor essay to resemble other poor ones. *E-rater* evaluates the topical content of an essay by comparing the words it contains to the words found in manually graded training examples for each of the six score categories. Two measures of content similarity are computed, one based on word frequency and the other on word weight, as in information retrieval applications (Salton, 1988). For the former application (*EssayContent*), content similarity is computed over the essay as a whole, while in the latter application (*ArgContent*) content similarities are computed for each argument in an essay.

For the frequency based measure (the *EssayContent* program), the content of each score category is converted to a single vector whose elements represent the total frequency of each word in the training essays for that category. In effect, this merges the essays for each score. (A stop list of some function words is removed prior to vector construction.) The system computes cosine correlations between the vector for a given test essay and the six vectors representing the trained categories; the category that is most similar to the test essay is assigned as the evaluation of its content. An advantage of using the cosine correlation is that it is not sensitive to essay length, which may vary considerably.

The other content similarity measure, *ArgContent*, is computed separately for each argument in the test essay and is based on the kind of term weighting used in information retrieval. For this purpose, the word frequency vectors for the six score categories, described above, are converted to vectors of word weights. The weight for word i in score category s is:

$$w_{i,s} = (freq_{i,s} / max\_freq_s) * \log(n\_essays_{total} / n\_essays_i)$$

where $freq_{i,s}$ is the frequency of word *i* in category *s*, max_freq_s is the frequency of the most frequent word in *s* (after a stop list of words has been removed), n_essays_{total} is the total number of training essays across all six categories, and n_essays_i is the number of training essays containing word *i*.

The first part of the weight formula represents the prominence of word *i* in the score category, and the second part is the log of the word's inverse document frequency (IDF). For each argument *a* in the test essay, a vector of word weights is also constructed. The weight for word *i* in argument *a* is

$$W_{i,a} = (freq_{i,a}/max\_freq_a) * \log(n\_essays_{total} / n\_essays_i)$$

where $freq_{i,a}$ is the frequency of word *i* in argument *a*, and max_freq_a is the frequency of the most frequent word in *a* (once again, after a stop list of words has been removed). Each argument (as it has been partitioned by *APA*) is evaluated by computing cosine correlations between its weighted vector and those of the six score categories, and the most similar category is assigned to the argument. As a result of this analysis, *e-rater* has a set of scores (one per argument) for each test essay.

We were curious to find out if an essay containing several good arguments (each with scores of 5 or 6) and several poor arguments (each with scores of 1 or 2) produced a different overall judgment by the human raters than an essay consisting of uniformly mediocre arguments (3's or 4's), or if perhaps humans were most influenced by the best or poorest argument in the essay. In a preliminary study, we looked at how well the minimum, maximum, mode, median, and mean of the set of argument scores agreed with the judgments of human raters for the essay as a whole. The mode and the mean showed good agreement with human raters, but the greatest agreement was obtained from an adjusted mean of the argument scores which compensated for an effect of the number of arguments in the essay. For example, essays which contained only one or two arguments tended to receive slightly lower scores from the human raters than the mean of the argument scores, and essays which contained many arguments tended to receive slightly higher scores than the mean of the argument scores. To compensate for this, an adjusted mean is used as *e-rater*'s *ArgContent*,

$$ArgContent = ((arg\_scores + n\_args) / (n\_args + 1)$$

3. Training and Testing

In all, *e-rater*'s syntactic, discourse, and topical analyses yielded a total of 57 features for each essay. The majority of the features in the overall feature set are discourse-related (see Table 3 for some examples). To predict the score assigned by human raters, a stepwise linear regression analysis was used to compute the optimal weights for these predictors

based on manually scored training essays. The training sets for each test question consisted of a total of 270 essays, 5 essays for score 0[2], 15 essays for score 1 (a rating infrequently used by the human raters) and 50 essays each for scores 2 through 6. After training, *e-rater* analyzed new test essays, and the regression weights were used to combine the measures into a predicted score for each one. *E-rater* predictions were compared to the two human rater scores to measure exact and adjacent agreement (see Table 1). **Figure 2** shows the predictive feature set identified by the regression analysis for one of the example test questions, ARG1, in **Tables 1** and **2**.

1. *ArgContent* Score
2. *EssayContent* Score
3. Total Argument Development Words/Phrases
4. Total Pronouns Beginning Arguments
5. Total Complement Clauses Beginning Arguments
6. Total Summary Words Beginning Arguments
7. Total Detail Words Beginning Arguments
8. Total Rhetorical Words Developing Arguments
9. Subjunctive Modal Verbs

Figure 2: Predictive Feature Set for ARG1 Test Question

3.1 Results

Table 1 shows the overall results for 8 GMAT argument questions, 5 GMAT issue questions and 2 TWE questions. The level of agreement between *e-rater* and the human raters ranged from 87% to 94% across the 15 tests. Agreement appears to be comparable to that found between the human raters.

Table 1: *E-rater* (*E*) and Human Rater (HR) Percentage Agreement & Human Interrater Percentage Agreement For Cross-Validation Tests

Question	*n*=	**HR ~ HR2**	**HR1 ~ *E***	**HR2 ~ *E***
Arg1	552	92	87	89
Arg2	517	93	91	89
Arg3	577	87	87	89
Arg4	592	91	92	93
Arg5	634	92	91	91
Arg6	706	87	87	88
Arg7	719	90	91	88
Arg8	684	89	89	90
Issue1	709	90	89	90
Issue2	747	92	89	90
Issue3	795	88	87	86
Issue4	879	92	87	87
Issue5	915	93	89	89
TWE1	260	------	93	------
TWE2	287	------	94	------

Table 2 shows that scores generated by *ArgContent* have higher agreement with human raters than do scores generated by *EssayContent*. This suggests that the discourse structures generated by *APA* are useful for score prediction, and that the application of content vector analysis to text partitioned into smaller units of discourse might improve *e-rater*'s overall scoring accuracy.

Table 2: Percentage Agreement Between *EssayContent (EC)* or *ArgContent (AC)* and Human Rater Score

Question	***n*=**	***HR1~ HR2***	***EC***	***AC***
Arg1	552	92	69	73
Arg2	517	93	68	75
Arg3	577	87	72	76
Arg4	592	91	70	81
Arg5	634	92	72	81
Arg6	706	87	67	82
Arg7	719	90	68	80
Arg8	684	89	62	80
Issue1	709	90	67	82
Issue2	747	92	65	83
Issue3	795	88	64	84
Issue4	879	92	69	83
Issue5	915	93	69	85
TWE1	260	------	77	88
TWE2	287	------	77	91
Average	638	90	69	82

Results for the essay questions in Tables 1 and 2 represent a wide variety of topics. (Sample questions that show topical variety in GMAT essays can be viewed at http://www.gmat.org/. Topical variety in TWE questions can be reviewed at http://www.toefl.org/tstprpmt.html.) The data also

[2] 0's either contain no text or the response is off-topic.

represented a wide range of English writing competency. The majority of test-takers from the two TWE data sets were nonnative English speakers. Despite these differences in topic and writing skill, *e-rater*, as well as *EssayContent*, and *ArgContent* performed consistently across items. In fact, over the 15 essay questions, the discourse features output by *APA* and scores output by *ArgContent* (based on discourse-chunked text) account for the majority of the most frequently occurring predictive features. These are shown in Table 3.

Table 3: Most Frequently Occurring Predictive Features Across 15 Essay Questions

Feature	Feature Class	Feature Counts
ArgContent	Topical/ Discourse	15/15
EssayContent	Topical	14/15
Total Argument Development Words	Discourse	14/15
Auxiliary Modals	Syntactic	12/15
Arg Init: Complement Clauses	Discourse	7/15
Arg Development: Rhetorical Question Words	Discourse	6/15
Arg Development: Evidence Words	Discourse	6/15
Subordinate Clauses	Syntactic	4/15
Relative Clauses	Syntactic	4/15

4. Discussion and Conclusions

The study indicates that discourse, syntactic, and topical information can be reliably used for machine prediction of essay scores. The results suggest that *e-rater*'s discourse marking is informative to the scoring process. *ArgContent*, the statistical, topical discourse analyzer, appears to be the most predictive feature. Other highly ranked features include surface cue words and non-lexical discourse cues.

One line of future research will examine the effects of various term weighting schemes on the performance of both *ArgContent* and *EssayContent*. Another study will compare the argument boundaries assigned by *APA* and the positions which human readers judge to be beginnings and ends of arguments.

We believe that the discourse related features used by *e-rater* might be the most useful building blocks for automated generation of diagnostic and instructional summaries about essays. For example, sentences indicated as "the beginning of an argument" could be used to flag main points of an essay (Marcu (1997)). *ArgContent*'s ability to generate "scores" for each argument could provide information about the relevance of individual arguments in an essay, which in turn could be used to generate helpful diagnostic or instructional information.

5. References

Cohen, Robin (1984). "A computational theory of the function of clue words in argument understanding." In Proceedings of 1984 International Computational Linguistics Conference. California, 251-255..

Hirschberg, Julia and Diane Litman (1993). "Empirical Studies on the Disambiguation of Cue Phrases." Computational Linguistics (19)3, 501-530.

Hovy, Eduard, Julia Lavid, Elisabeth Maier, "Employing Knowledge Resources in a New Text Planner Architecture," In Aspects of Automated NL Generation, Dale, Hovy, Rosner and Stoch (Eds), Springer-Verlag Lecture Notes in AI no. 587, 57-72.

GMAT (1997). http://www.gmat.org/

Knott, Alistair. (1996). "A Data-Driven Methodology for Motivating a Set of Coherence Relations." Ph.D. Dissertation, available at www.cogsci.edu.ac.uk/~alik/publications.
html, under the Heading, Unpublished Stuff.

Litman, Diane, J. (1996). "Cue Phrase Classification Using Machine Learning." Artificial Intelligence, 5, 53-94.

Mann, William C. and Sandra A. Thompson (1988). "Rhetorical Structure Theory: Toward a functional theory of text organization." Text 8(3), 243-281.

Marcu, Daniel. (1997). "From Discourse Structures to Text Summaries.", In Proceedings of the Intelligent Scalable Text Summarization Workshop, Association for Computational Linguistics,

Universidad Nacional de Educacion a Distancia, Madrid, Spain.

MSNLP (1997) http://research.microsoft.com/nlp/

Nolan, James (1997). The Architecture of a Hybrid Knowledge-Based System for Evaluating Writing Samples. In A. Niku-Lari (Ed.) Expert Systems Applications and Artificial Intelligence Technology Transfer Series, EXPERSYS-97. Gournay S/M, France: IITT International.

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

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartik (1985). A Comprehensive Grammar of the English Language. Longman, New York.

Sidner, Candace. (1986). Focusing in the Comprehension of Definite Anaphora. In Readings in Natural Language Processing, Barbara Grosz, Karen Sparck Jones, and Bonnie Lynn Webber (Eds.), Morgan Kaufmann Publishers, Los Altos, California, 363-394.

Salton, Gerard. (1988). Automatic text processing : the transformation, analysis, and retrieval of information by computer. Addison-Wesley, Reading, Mass.

TOEFL (1997). http://www.toefl.org/tstprpmt.html

Vander Linden, Keith and James H. Martin (1995). “Expressing Rhetorical Relations in Instructional Text: A Case Study in Purpose Relation.” Computational Linguistics 21(1), 29-57.