(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.

Automated Scoring Using A Hybrid Feature Identification Technique

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu†
Martin Chodorow‡, Lisa Braden-Harder‡‡, and Mary Dee Harris‡‡‡

†Educational Testing Service, Princeton NJ, ‡Hunter College, New York City, NY,
‡‡Butler-Hill Group, Reston, VA, and ‡‡‡Language Technology, Inc, Austin, TX

Abstract

This study exploits statistical redundancy inherent in natural language to automatically predict scores for essays. We use a hybrid feature identification method, including syntactic structure analysis, rhetorical structure analysis, and topical analysis, to score essay responses from test-takers of the Graduate Management Admissions Test (GMAT) and the Test of Written English (TWE). For each essay question, a stepwise linear regression analysis is run on a training set (sample of human scored essay responses) to extract a weighted set of predictive features for each test question. Score prediction for cross-validation sets is calculated from the set of predictive features. Exact or adjacent agreement between the Electronic Essay Rater (*e-rater*) score predictions and human rater scores ranged from 87% to 94% across the 15 test questions.

1. Introduction

This paper describes the development and evaluation of a prototype system designed for the purpose of automatically scoring essay responses. The paper reports on evaluation results from scoring 13 sets of essay data from the Analytical Writing Assessments of the Graduate Management Admissions Test (GMAT) (see the GMAT Web site at http://www.gmat.org/ for sample questions) and 2 sets of essay data from the Test of Written English (TWE) (see http://www.toefl.org/tstprpmt.html for sample TWE questions).

Electronic Essay Rater (*e-rater*) was designed to automatically analyze essay features based on writing characteristics specified at each of six score points in the scoring guide used by human raters for manual scoring (also available at http://www.gmat.org/). 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). Lower scores are assigned to essays as these characteristics diminish.

One of our main goals was to design a system that could score an essay based on features specified in the scoring guide for manual scoring. *E-rater* features include rhetorical structure, syntactic structure, and topical analysis. For each essay question, a stepwise linear regression analysis is run on a set of training data (human-scored essay responses) to extract a weighted set of predictive features for each test question. Final score prediction for cross-validation uses the weighted predictive feature set identified during training. Score prediction accuracy is determined by measuring agreement between human rater scores and *e-rater* score predictions. In accordance with human interrater "agreement" standards, human and *e-rater* scores also "agree" if there is an exact match or if the scores differ by no more than one point (adjacent agreement).

2. Hybrid Feature Methodology

E-rater uses a hybrid feature methodology that incorporates several variables either derived statistically, or extracted through NLP techniques. The final linear regression model used for predicting scores includes syntactic, rhetorical and topical features. The next three sections present a conceptual rationale and a description of feature identification in essay responses.

2.1 Syntactic Features

The scoring guides indicate that one feature used to evaluate an essay is *syntactic variety*. All sentences in the essays were parsed using the Microsoft Natural Language Processing tool (MSNLP) (see MSNLP (1997)) so that syntactic structure information could be accessed. The identification of syntactic structures in essay responses yields information about the syntactic variety in an essay with regard to the identification of clause or verb types.

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 also used as measures of *syntactic variety*.

2.2 Rhetorical Structure Analysis

GMAT essay questions are of two types: *Analysis of an Issue* (issue*) and Analysis of an Argument* (argument). The GMAT 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 GMAT 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 indicate that an essay will receive a score based on the examinee's demonstration of a well-developed essay. In this study, we try to identify organization of an essay through automated analysis and identification of the rhetorical (or argument) structure of the essay.

Argument structure in the rhetorical sense may or may not correspond to paragraph divisions. One can make a point in a phrase, a sentence, two or three sentences, a paragraph, and so on. For automated argument identification, *e-rater* identifies 'rhetorical' relations, such as Parallelism and Contrast that can appear at almost any level of discourse. This is part of the reason that human readers must also rely on cue words to identify new arguments in an essay.

Literature in the field of discourse analysis supports our approach. It points out that rhetorical 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), and Knott (1996)). *E-rater* follows this approach by using rhetorical cue words and structure features, in addition to other topical and syntactic information. We adapted 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. Cue words such as "perhaps," and "possibly" are considered to be "belief" words used by the writer to express a belief in developing an argument in the essay. Words like "this" and "these" may often be used to flag that the writer has not changed topics (Sidner (1986)). We also observed that in certain discourse contexts structures such as infinitive clauses mark the beginning of a new argument.

E-rater's automated argument partitioning and annotation program (APA) outputs an annotated version of each essay in which the *argument units* of the essays are labeled with regard to their status as "marking the beginning of an argument," or "marking argument development." APA also outputs a version of the essay that has been partitioned "by argument", instead of "by paragraph," as it was originally partitioned by the test-taker. APA uses rules for argument annotation and partitioning based on syntactic and paragraph-based distribution of cue words, phrases and structures to identify rhetorical structure. Relevant cue words and terms are stored in a cue word lexicon.

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 resemble other good essays in its choice of words and, conversely, a poor essay to resemble other poor ones. *E-rater* evaluates the lexical and 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 programs were implemented that compute measures of content similarity, one based on word frequency (*EssayContent*) and the other on word weight (*ArgContent*), as in information retrieval applications (Salton (1988)).

In *EssayContent,* the vocabulary 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, is computed separately by *ArgContent* 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. For each argument in the test essay, a vector of word weights is also constructed. Each argument 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.

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 greatest agreement was obtained from an adjusted mean of the argument scores that compensated for an effect of the number of arguments in the essay. For example, essays

(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.

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 =
(Σarg_scores + n_args) / (n_args + 1)

3. Training and Testing

In all, *e-rater*'s syntactic, rhetorical, and topical analyses yielded a total of 57 features for each essay. The training sets for each test question consisted of 5 essays for score 0, 15 essays for score 1, and 50 essays each for scores 2 through 6. 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. For example, **Figure 1,** below, shows the predictive feature set generated for the ARG1 test question (see results in **Table 1**). The predictive feature set for ARG1 illustrates how criteria specified for manual scoring described earlier, such as argument topic and development (using the *ArgContent* score and argument development terms), syntactic structure usage, and word usage (using the *EssayContent* score), are represented by *e-rater*. 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. This prediction was then compared to the scores assigned by two human raters to check for exact or adjacent agreement.

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 1: 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. There was an average of 638 response essays per test question. *E-rater* and human rater mean agreement across the 15 data sets was 89%. In many cases, agreement was as high as that found between the two human raters.

The items that were tested represented a wide variety of topics (see http://www.gmat.org/ for GMAT sample questions and http://www.toefl.org/tstprpmt.html for sample TWE questions). The data also represented a wide variety of English writing competency. In fact, the majority of test-takers from the 2 TWE data sets were nonnative English speakers. Despite these differences in topic and writing skill *e-rater* performed consistently well across items.

Table 1: Mean Percentage and Standard Deviation for *E-rater (E)* and Human Rater (H) Agreement & Human Interrater Agreement For 15 Cross-Validation Tests

	H1~H2	**H1~*E***	**H2~*E***
Mean	90.4	89.1	89.0
S.D	2.1	2.3	2.7

To determine the features that were the most reliable predictors of essay score, we examined the regression models built during training. A feature type was considered to be a reliable predictor if it proved to be significant in at least 12 of the 15 regression analyses. Using this criterion, the most reliable predictors were the *ArgContent* and *EssayContent* scores, the number of cue words or phrases indicating the development of an argument, the number of syntactic verb and clause types, and the number of cue words or phrases indicating the beginning of an argument.

4. Discussion and Conclusions

This study shows how natural language processing methods and statistical techniques can be used for the evaluation of text. The study indicates that rhetorical, syntactic, and topical information can be automatically extracted and used for machine-based score prediction of essay responses. These three types of information model features specified in the manual scoring guides. This study also shows that *e-rater* adapts well to many different topical domains and populations of test-takers.

The information used for automated score prediction by *e-rater* can also be used as building blocks for automated generation of diagnostic and instructional summaries. Clauses and sentences annotated by APA as "the beginning of a new argument" might be used to identify main points of an essay (Marcu (1997)). In turn, identifying the main points in the text of an essay could be used to generate feedback reflecting essay topic and organization. Other features could be used to automatically generate statements that explicate the basis on which *e-rater* generates scores. Such statements could supplement manually created qualitative feedback about an essay.

6. 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.

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/

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

(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.

Instructional Text: A Case Study in Purpose Relation." *Computational Linguistics 21(1)*, 29-57.