Computer Analysis of Essays

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

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

Abstract

Research into the use of advanced computational linguistics techniques recently culminated in the implementation of a prototype automatic essay scoring system at Educational Testing Service (ETS). In an evaluation study using data sets from thirteen different GMAT essay prompts, this system, *e-rater*, showed between 87% and 94% agreement with expert readers' scores, an accuracy comparable to that between two expert readers. This indicates that *e-rater* might be useful as a second reader for high stakes assessments, thus leading to a considerable reduction in essay scoring costs. We describe the design and evaluation of the *e-rater* automatic essay scoring system and discuss some implications of this research for assessment.

1. Introduction

Increased emphasis on performance-based testing has led to the inclusion of more constructed-response writing items on standardized tests. The Analytical Writing Assessments of the Graduate Management Admissions Test (GMAT) are typical examples (see the GMAT Web site at http://www.gmat.org for examples). But scoring essays manually is costly and time consuming. In addition, the subjectivity inherent in human reader scoring has been subject to criticism. For more than five years, ETS researchers have been exploring the use of advanced computational linguistics techniques for automatically scoring a variety of constructed writing responses. This research recently culminated in the implementation of a prototype automatic essay scoring system, dubbed *e-rater***.**

During the fall of 1997, *e-rater* was run in the background to compare its performance for scoring GMAT essays to that of expert readers. *E-rater's* scores showed between 87% and 94% agreement with expert readers' scores on thirteen different GMAT essay prompts,[1] an agreement rate comparable to that between two expert readers who scored the same essays. *E-rater* was also evaluated on two essay prompts from the Test of Written English (TWE) (see http://www.toefl.org for examples). For these two data sets, *e-rater* achieved agreement rates of 93% and 94% with expert readers. These results indicate that *e-rater* might be useful as a second reader for high stakes assessments, thus leading to a considerable reduction in essay scoring costs. This paper describes the design and evaluation of the *e-rater* automatic essay scoring system and discusses some implications of this research for assessment.

The *e-rater* system was designed to automatically analyze essay features based on writing characteristics specified in the holistic scoring guide used by expert readers for manual scoring of GMAT essays (also available at http://www.gmat.org). This scoring guide has a six-point scoring scale. It indicates that an essay that stays on the topic of the prompt, has a strong, coherent and well-organized argument structure, and displays variety in both syntactic structure and vocabulary usage 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.

[1] The term "prompt" is used to refer to the question that the examinee is asked to write an essay on.

One of our goals was to design a system that could score essays based on features similar to those used by human experts. We identified a wide variety of syntactic, rhetorical and topical features that might be viewed as evidence for the principles described in the scoring guide, and we implemented computational linguistics programs for quantifying the occurrence of these features in essays. For each essay prompt, we ran these feature extraction programs on a "training set" of essays scored by human experts. Then we used a stepwise linear regression to derive weights for those features that were most predictive of human expert scores, thus providing a scoring model for each essay prompt. Finally, we collected an additional set of cross-validation essays for each prompt, ran the *e-rater* scoring model programs to score them, and compared *e-rater's* scores to human experts' scores for the same essays.

We determined *e-rater's* score prediction accuracy for the cross-validation essay sets by measuring agreement between *e-rater's* scores and human reader scores. Human reader scores and machine scores are considered to "agree" if there is no more than a single point difference between the scores on the six-point scale. The same criterion is used to measure agreement between two human readers. When two human readers fail to agree, the essay is referred to a third expert reader. Additional measures of *e-rater's* scoring accuracy are discussed in Section 3 below. First we briefly present a conceptual rationale and a description of *e-rater's* evidentiary feature scoring methodology.

2. Evidentiary Feature Scoring

E-rater's evidentiary feature scoring methodology incorporates more than 60 variables that might be viewed as evidence that an essay exhibits writing characteristics described in the GMAT essay scoring guide. These variables comprise three general classes of features: syntactic, rhetorical and topical content features. The features are extracted from essay texts and quantified using computational linguistics techniques. In this section we briefly describe how features from each class are computationally quantified in essays.

2.1 Syntactic Structure Analysis

The scoring guide indicates that *syntactic variety* is an important feature in evaluating essays. Analysis of the syntactic structure of sentences in an essay can yield information about the essay's syntactic variety, such as quantity and ratio of simple, compound and complex sentences, types of dependent clauses, use of modal auxiliary verbs, and other features. *E-rater* employs a syntactic parser included in the Microsoft Natural Language Processing tool (MSNLP) (see MSNLP, 1997) to parse each of the sentences in an essay. Based on the information in these parses, other *e-rater* programs then quantify such features as number of complement clauses, subordinate clauses, infinitive clauses and relative clauses and occurrences of subjunctive modal auxiliary verbs such as *would*, *could*, *should*, *might* and *may*. Ratios of syntactic structure types per essay and per sentence are also computed as measures of *syntactic variety*.

2.2 Rhetorical Structure Analysis

GMAT essay prompts 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.[2]

The scoring guide indicates that an essay will receive a score based on the examinee's demonstration of a well-developed essay. For the argument essay, the scoring guide states specifically that a "6" essay "develops ideas cogently, organizes them logically, and connects them with clear transitions." The correlate to this for the issue essay is 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 language in holistic scoring guides, such as, "cogent", "logical," "insightful," and "well-organized" has "fuzzy" meaning because it is based on imprecise observation. Nolan uses "fuzzy logic" methods to automatically assign these kinds of "fuzzy" classifications to essays. In *e-rater*, we try to quantify evidence about how well organized an essay is through automated identification and analysis of the rhetorical (or argument) structure of the essay.

Argument structure in the rhetorical sense may or may not correspond to paragraph divisions. There is no particular text unit that corresponds to the stages, steps, or passages of an argument. One can make a point in a phrase, a sentence, two or three sentences, a paragraph, two or three paragraphs, several pages, and so on. For this reason, an essay reader must rely on cues of several types to identify the chunks of text that correspond to separate arguments and separate points within arguments. We found it to be useful to identify rhetorical relations, such as *Parallelism* and *Contrast*, and other coherence relations to try to identify the individual arguments and points within essays (Hobbs 1979, Polanyi 1988).

Literature in the field of discourse analysis points out that rhetorical relations can often be identified by the occurrence of cue words and specific syntactic structures (Cohen 1984, Mann and Thompson 1988, Hovy, et al. 1992, Hirschberg and Litman 1993, Van der Linden and Martin 1995, Knott 1996). *E-rater* follows this approach by identifying and quantifying an essay's use of cue words and other rhetorical structure features.

For example, we adapted the conceptual framework of conjunctive relations from Quirk, et al. (1985) in which phrases such as "In summary" and "In conclusion," are classified as conjuncts used for summarizing. *E-rater* identifies these phrases and others as cues for a *Summary* relation. Words such as "perhaps" and "possibly" are considered to be cues for a *Belief* relation, one used by the writer to express a belief while developing an argument in the essay. Words like "this" and "these" are often used within certain syntactic structures to indicate that the writer has not changed topics (Sidner 1986). In certain discourse contexts, structures such as infinitive clauses mark the beginning of a new argument.

[2] The TWE essays question types are similar to the GMAT issue and argument question types (see http://www.toefl.org).

E-rater contains a lexicon of relevant rhetorical cue words and phrases. It also contains a set of heuristic rules for identifying rhetorical relations based on syntactic and paragraph-based distribution of cue words, phrases and structures. These rhetorical analysis rules and lexicon are used by an argument partitioning and annotation program (APA) to produce a version of an essay that has been partitioned "by argument", instead of "by paragraph". In addition, APA annotates *argument units* within an essay to label their rhetorical relations as well as their function in marking the beginning of an argument or marking argument development.

2.3 Topical Content 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 might therefore expect a good essay to resemble other good essays in its vocabulary use patterns, and, similarly, a poor essay to resemble other poor ones. *E-rater* evaluates the topical content of an essay by comparing the patterns of words it contains to those found in manually graded training examples for each of the six score levels. Two different measures of content similarity are computed, one based on vocabulary use in the essay as a whole and another based on the specific vocabulary content of the *argument units* found in the essay, as determined by the APA program. We refer to the first as an *EssayContent* measure and the second as an *ArgContent* measure.

For the *EssayContent* measure, we first build a representative "supervector" for each of the six score levels by merging all the essays in the training set for that score level and computing the total frequency counts of all the words in those essays. Some function words (i.e., articles, prepositions, etc.) are removed prior to vector construction, and minimal suffix stripping of words is performed prior to the frequency computation.

To derive an *EssayContent* value for a new essay, *e-rater* computes cosine distances between a similarly constructed vector for the new essay and each of the supervectors representing the six score levels. The new test essay is assigned the score level of the closest matching supervector. An advantage of using the cosine distance measure 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. Unlike the *EssayContent* measure that uses raw word frequencies in the supervectors, *ArgContent* uses weighted frequency values in its supervectors. These weighted frequency values are computed based on a standard term weighting method used in information retrieval called inverse document frequency weighting (see Salton 1988).

To derive an *ArgContent* value for a new essay, each argument in the essay is first evaluated separately by computing cosine distances between its weighted vector and weighted supervectors for the six score levels. The score level of the most similar supervector is assigned to that argument. As a result of this analysis, *e-rater* has a set of scores (one per argument) for the new essay.

We were curious as to whether 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 experts than did 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 readers for the essay as a whole. The mode and the mean showed good agreement with human readers, 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 that contained only one or two arguments tended to receive slightly lower scores from the human readers than the mean of the argument scores, and essays that 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* measure, i.e.,

ArgContent =
(□ arg_scores + num_args) / (num_args + 1).

3. Training and Evaluation Results

In all, our syntactic, rhetorical, and topical content analyses yielded a total of 67 evidentiary features. To derive models capable of predicting scores assigned by human readers, we started with training sets of manually scored essays for each prompt. Each training set consisted of 5 essays for score level 0[3], 15 essays for score level 1 (a rating infrequently used by the human readers) and 50 essays each for score levels 2 through 6. For each essay in the training set, we ran *e-rater's* programs to compute values for all 67 evidentiary features. We then submitted the feature vectors to stepwise linear regression analyses to compute optimal weights for features. We refer to the resulting combination of significant features and their weights as the scoring model for the prompt.

After training, *e-rater* analyzed new test essays for each prompt (i.e., a cross-validation set) and used its scoring model to combine the significant features into a predicted score for each essay. We then compared *e-rater's* score predictions to the scores assigned by two (or sometimes three) human readers to check for agreement (i.e., scores that differ by no more than 1 point).

3.1 Agreement Results

Table 1 shows the agreement results for 8 GMAT *Argument* prompts, 5 GMAT *Issue* prompts and 2 TWE prompts. The number of essays, *n*, in the cross-validation sets is shown in column 2. The degree of agreement between *e-rater* and human readers ranged from 87% to 94% across the 15 prompts. In many cases, agreement was as high as that found between the two human readers.

The essay prompts in **Table 1** represent a wide variety of topics. Sample prompts that show topical variety in GMAT essays can be viewed at http://www.gmat.org. Topical variety in TWE essay prompts can be viewed at http://www.toefl.org/tstprpmt.html. The data also represent a wide variety of English writing competencies, in that the majority of test-takers from the two TWE data sets were non-native English speakers. Despite these differences in topic and writing skills, *e-rater* performed consistently well across all prompts.

[3] Human raters assigned a 0 to essays that either contained no response or were off-topic. Zeros were infrequent so training sets contained only 5 essays for the 0 score level.

Table 1: Percent Agreement, Mean Percentage & Standard Deviation between Human Reader and *E-rater* Essay Scores

Prompt	*n*=	HR1 ~ HR2	HR1 ~ *e-rater*	HR2 ~ *e-rater*
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%	------
Mean		90.4	89.1	89.0
S.D.		2.1	2.3	2.7

In Table 2, a *field score* is used to determine exact or adjacent agreement between human readers. The field score is derived by taking the average of the two human reader scores when the humans show exact or adjacent agreement, and rounding up (e.g., 3.5 becomes 4.0). The third reader score is used as the field score if the two original human readers disagree by more than a single point. **Table 2** shows summary results across prompts for agreement between the two human readers at each field score level.

Table 2: Percent Agreement Between Human Readers 1 and 2 Across All Prompts at Each Field Score Level

Field Score	% Agreement	Totals
0	100	188/188
1	92	149/161
2	86	730/848
3	88	1717/1943
4	90	2861/3168
5	92	2108/2285
6	92	402/433
Average	90	8155/9026

We then looked at *e-rater's* accuracy at each of the six score levels. In **Tables 3a and 3b**, a single human reader score is used as a baseline score in measuring exact or adjacent agreement between *e-rater* and human reader scores. **Table 3a** shows summary agreement data between *e-rater* and human reader 1, and **Table 3b** shows summary agreement data for *e-rater* and human reader 2 for each score level across all 13 GMAT prompts. From **Tables 3a** and **3b**, the occurrences of each score assignment can be observed for both *e-rater* and the human readers. Variation in percentage of exact or adjacent agreement exists depending on whether the human reader score (**HR1**/*e* and **HR2**/*e*), or the *e-rater* score *(e*/**HR1** and *e*/**HR2)**is used as a baseline score.

Table 3a: *E-rater* Agreement with Human Reader 1 at Each Score Level Across All Prompts

Score Level	%	Totals HR1/*e*	%	Totals *e*/HR1
0	96	182/188	98	107/109
1	78	250/317	91	252/276
2	85	951/1112	86	949/1098
3	92	2111/2272	87	2442/2783
4	92	2787/2998	92	2801/3044
5	84	1479/1749	89	1218/1368
6	65	256/390	70	247/348
Avg	88	8016/9026	88	8016/9026

Table 3b: *E-rater* Agreement with Human Reader 2 at Each Score Level Across All Prompts

Score Level	%	Totals HR2/*e*	%	Totals *e*/HR2
0	96	182/188	98	107/109
1	77	260/334	89	246/276
2	84	957/1132	86	945/1098
3	92	2043/2197	88	2459/2783
4	93	2841/3052	91	2794/3044
5	85	1508/1767	90	1235/1368
6	68	245/356	71	250/348
Avg	89	8036/9026	89	8036/9026

3.2 Reliabilities and Correlations

ETS statisticians computed several reliability statistics for the human and *e-rater* scores. **Table 4** shows single rater reliabilities based on correlations between two human readers and between *e-rater* and each human reader. **Table 5** shows mean reliabilities based on the means of two human readers in column 2 and based on the means of *e-rater* and each human reader in columns 3 and 4.

Table 4: Single Rater Reliabilities

Prompt	HR1 ~ HR2	HR1 ~ *e-rater*	HR2 ~ *e-rater*
Arg1	.74	.65	.70
Arg2	.76	.70	.69
Arg3	.71	.64	.68
Arg4	.70	.65	.68
Arg5	.79	.74	.74
Arg6	.65	.64	.66
Arg7	.72	.70	.63
Arg8	.67	.65	.66
Issue1	.73	.67	.70
Issue2	.72	.67	.66
Issue3	.64	.60	.59
Issue4	.75	.66	.66
Issue5	.72	.62	.63

Table 5: Mean Reliabilities

Prompt	m(HR1 ~ HR2)	m(HR1~ *e-rater*)	m(HR1~ *e-rater*)
Arg1	.85	.79	.82
Arg2	.86	.83	.81
Arg3	.84	.78	.81
Arg4	.82	.78	.80
Arg5	.88	.85	.85
Arg6	.79	.78	.79
Arg7	.84	.82	.78
Arg8	.80	.79	.79
Issue1	.84	.80	.82
Issue2	.84	.80	.79
Issue3	.78	.75	.73
Issue4	.86	.79	.79
Issue5	.84	.77	.77

Mean reliabilities are more encouraging, and as well, they are arguably more relevant, since it is a mean score of two readers that is considered in the field. For the majority of the prompts, mean reliability for two human readers appears to be comparable to mean reliability for a human reader and *e-rater*.

Furthermore, these data are conservative from the perspective that a third human reader is called in when the first two disagree on a score by more than a single point. Taking those third scores into account would increase mean reliabilities of human readers since in the majority of cases the third score falls between the first two. By the same token, an additional human reader would be used to resolve any scoring discrepancies greater than 1 point that arise between a human reader and *e-rater*.

Ordinary Pearson correlations between human reader scores and between human reader and *e-rater* scores are shown in **Table 6.** *E-rater's* agreement with each human reader is comparable to the agreement between two human readers for most prompts.

Another important observation is that the reported reliability and correlation figures do show some variation across prompts. Some of this variation may be attributed to topic variation across prompts. Indeed, *e-rater's* suite of text analysis tools might be useful in future research studies aimed at predicting and clarifying variation across prompts

Table 6: Ordinary Pearson Correlations

Prompt	HR1 ~HR2	HR1 -*e*	HR2 -*e*
Arg1	.87	.82	.85
Arg2	.88	.85	.84
Arg3	.85	.81	.83
Arg4	.84	.82	.83
Arg5	.89	.87	.87
Arg6	.82	.81	.82
Arg7	.86	.84	.81
Arg8	.83	.82	.82
Iss1	.86	.83	.84
Iss2	.86	.83	.82
Iss3	.82	.80	.79
Iss4	.87	.82	.82
Iss5	.86	.81	.81

3.3. Salient Evidentiary Features

To determine which evidentiary features were the most salient predictors of essay scores, we examined the regression models derived during training. A ranking of *e-rater's* ten most salient evidentiary features according to their prominence across all 15 scoring models is shown in **Table 7.**

Table 7: Occurrence of Evidentiary Features across 15 Scoring Models

Feature	Feature Class	Feature Counts
ArgContent	Topical/ Rhetorical	15/15
EssayContent	Topical	14/15
Total Argument Development Words	Rhetorical	14/15
Auxiliary Subjunctives	Syntactic	12/15
Paragraphs	Surface	8/15
Arg Initialization: Complement Clauses	Rhetorical	7/15
Arg Development: Rhet Ques Words	Rhetorical	6/15
Arg Development: Evidence Words	Rhetorical	6/15
Subordinate Clauses	Syntactic	4/15
Relative Clauses	Syntactic	4/15

A feature type was considered to be a salient predictor if it proved to be significant in at least 12 of the 15 regression analyses. Using this criterion, the most salient predictors were *ArgContent, EssayContent*, the total number of argument development words, and the total number of subjunctive auxiliary verbs. Apart from these four features, the individual features that were significant for each prompt varied greatly across the 15 different scoring models. This point is noteworthy in that it attests to the non-coachability of *e-rater's* scoring method.

3.4 A First Look at *E-rater* Misses

The results of *e-rater's* performance are quite promising with regard to exact or adjacent agreement with human readers. We are now beginning to explore the issue of when and why *e-rater* "misses". "Misses" are those essays for which *e-rater* assigned a score that disagreed with a human reader by more than a single point (e.g., the human reader score is a "4" and *e-rater* assigns a "6"). "Hits" are those essays that *e-rater* scored appropriately (i.e., exact or adjacent agreement with human readers). In this section we first compare human reader's pattern of misses to *e-rater's*. Then, we briefly compare feature patterns in *e-rater's* hits and misses to look for any obvious differences.

3.4.1 Is *E-rater* Missing when Human Readers are missing?

One frequently posed question with regard to *e-rater* misses, is "*Does e-rater miss when human readers have difficulty, too?*" One way to address this question is to compare the disagreement rates between two human readers to the disagreement rates between *e-rater* and human readers at each score level across all prompts. **Table 8** shows disagreement rates between two human readers at each field score level as described for **Table 2**.

Table 8: Percent Disagreement Between Human Readers 1 and 2 Across all Prompts at Each Field Score Level

Field Score	% Dis-agreement	Totals
0	0	0/188
1	8	12/161
2	14	118/848
3	12	226/1943
4	10	307/3168
5	8	177/2285
6	7	31/433
Average	10	871/9026

Tables 9a and 9b show the disagreement between human reader one and *e-rater*, and human reader two and *e-rater*, respectively. Again, the human reader score and the *e-*

rater score assignments are used as a baseline to show the variation in the results when alternate baseline scores are used. The overall rate of disagreement between human readers, and between human readers and *e-rater* is approximately equivalent. Differences exist at the different score points, however.

Table 9a: *E-rater* Disagreement with Human Reader 1 at Each Score Level Across All Prompts

Score	%	Totals *e* / HR1	%	Totals HR1 / *e*
0	3	6/188	2	2/109
1	21	67/317	9	24/276
2	14	161/1112	14	149/1098
3	7	161/2272	12	341/2783
4	7	211/2998	8	243/3044
5	15	270/1749	11	150/1368
6	34	134/390	29	101/348
Avg	11	1010/9026	11	1010/9026

Table 9b: *E-rater* Disagreement with Human Reader 2 at Each Score Level Across All Prompts

Score	%	Totals *e* / HR2	%	Totals HR2 / *e*
0	3	6/188	2	2/109
1	22	74/334	11	30/276
2	15	175/1132	14	153/1098
3	7	154/2197	12	324/2783
4	7	211/3052	8	250/3044
5	15	259/1767	10	133/1368
6	31	111/356	28	98/348
Avg	11	990/9026	11	990/9026

We can then compare *e-rater's* performance to the third human reader's score to ascertain how often *e-rater* agrees or disagrees with the third human reader. For each essay for which a third human reader was required to adjudicate an unresolved score, we computed agreement between the third human reader score and the *e-rater* score prediction. These results are in **Table 10**.

A low rate of agreement between *e-rater* and the third human reader would indicate that *e-rater* is also "missing" when human readers 1 and 2 disagree. Conversely, a high rate of agreement between *e-rater* and human rater 3 would indicate that *e-rater* is in agreement with the field score, hence is it not "missing" when human readers 1 and 2 disagree.

Table 10 indicates that it is only at score level 1 that *e-rater* tends to "miss" when human readers 1 and 2 disagree. At the remaining score levels, *e-rater* agrees fairly strongly with human reader 3, i.e., the field score; in these instances, *e-rater* does not tend to "miss" when human readers 1 and 2 disagree. On average, *e-rater* agrees with the human reader 3 score 86% of the time. One could speculate from this that, overall, *e-rater* and human readers 1 and 2 miss relatively infrequently on the same essays.

Table 10: Percent Agreement Between *E-rater* and Human Reader 3 at Each Score Level Across All Prompts

HR3 (Field) Score	% Agreement	Totals: *e* Agreement / HR3 Scores
1	17%	2/12
2	75%	88/118
3	86%	194/226
4	95%	292/307
5	84%	148/177
6	74%	23/31
Avg	86%	747/871

3.4.2 Possible Source of "Misses"

We also performed a preliminary analysis of the patterns of evidentiary features appearing in *e-rater's* hits and misses. For each syntactic and rhetorical feature, we calculated the average number of occurrences of that feature in both the set of essays that *e-rater* missed and the set of essays that *e-rater* scored correctly (hits) for each prompt. We found that, across all prompts, there was little or no difference in the average number of occurrences of syntactic or rhetorical evidentiary features between these two sets.

As is explained earlier, the *EssayContent* and *ArgContent* programs assign a score to an essay based on the relevant vocabulary in an essay. For these two topical content features, *EssayContent* and *ArgContent*, we calculated the percentages of the time that they were in exact or adjacent agreement with a human reader for both the set of *e-rater* missed and the set of essays *e-rater* hit.

We observed some differences in the amount of agreement with human reader scores between *EssayContent* and *ArgContent* scores. The differences between *EssayContent* scores and human reader scores over both *e-rater* hits and misses data sets were typically greater, on average, than those for *ArgContent*.

Overall, these analyses suggest that the greatest source of *e-rater* misses may be in the topical analysis components. We are currently exploring ways to address this in order to achieve greater agreement between *e-rater* and human reader scores, including the possible use of Latent Semantic Analysis techniques (Deerwester et al. 1990), more sophisticated lexical (stopwords) and morphological processing, and potential logical form and semantic processing.

4. Conclusion and Future Implications

The results of this study indicate that a combination of advanced computational linguistics and statistical analysis techniques has now put automated essay scoring into the arena of practical applications. With a modicum of developmental effort, an operational system for automated essay scoring could be deployed in a matter of months. Such a system might serve as a second reader for high stakes assessments, thus leading to considerable savings in today's essay scoring costs. These results also invoke a variety of additional research questions focusing on the likely consequences of computer scoring on test validity, writing instruction, and public understanding and acceptance.

Note that the architecture of the *e-rater* essay scoring system does **not** take the human reader out of the loop. Indeed, because the system requires initial training sets of manually scored essays, the scoring models derived by the system actually embody the judgements made by human readers.

One implication of this "human derived scoring model" architecture is its flexibility in allowing for a cost vs. reliability level trade-off. That is, for low stakes applications, such as practice essay writing systems, a training set of essays scored by a single human reader may suffice. For higher stakes assessments, a training set scored by two or three human readers would increase the reliability of derived scoring models. For the price of *n* additional human readers, derived computer scoring models might approximate "true" scores. Another variable under the

control of testing agencies is the number of human readers to be deployed in conjunction with the automatic scoring engine after the scoring model has been derived. A related variable is how often a scoring model might be "calibrated" with additional human scores.

Additional research could improve not only the accuracy but also the diagnostic and explanatory power of this automated scoring architecture. For example, *e-rater* currently employs only linear statistical analysis techniques to derive its scoring models. Non-linear techniques are currently being explored to improve scoring accuracy. Furthermore, the set of evidentiary features currently used by *e-rater* for score prediction is only a first approximation to those used in human judgements. Further extensions, revisions and clustering of evidentiary feature sets might eventually provide a greater insight into "what human experts are doing when they score an essay."

We believe that the information used for automated score prediction by *e-rater* can also be used as building blocks for automated generation of writing diagnostics and instructional feedback. For example, clauses and sentences annotated by the APA program as "the beginning of a new argument" could 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 that reflects essay topic and the organization of the text. Other features could be used to automatically generate statements that inform the test-taker of the basis on which essays are scored by the computer. They could also supplement manually generated qualitative feedback about an essay.

Acknowledgements

Lisa Braden-Harder, a Computational Linguist and consultant to Microsoft, also played an integral role in the design and development of the *e-rater* system. Don Rock, Bruce, Kaplan and Shuyi Hua, three of ETS's complement of world-class statisticians, graciously and patiently provided the statistical analyses and expertise involved in this research. We're grateful to work with all of these talented individuals.

References

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

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, R. Harshman (1990). "Indexing by Latent Semantic Analysis, *JASIS, 41(6)*, 391-407.

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

Hobbs, Jerry (1979). "Coherence and coreference", *Cognitive Science, 3(1)*, 67-90.

Hovy, Eduard, Julia Lavid, Elisabeth Maier (1992). "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, http://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.

Nolan, James (forthcoming). "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.

Polanyi, Livia (1988). "A formal model of discourse structure", *Journal of Pragmatics, 12*, 601-638.

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.