Toward Evaluation of Writing Style: Finding Overly Repetitive Word Use in Student Essays

Jill Burstein
Educational Testing Service
Princeton, New Jersey 08541, USA
jburstein@ets.org

Magdalena Wolska
Universität des Saarlandes
Saarbücken, Germany
magda@coli.uni-sb.de

Abstract

Automated essay scoring is now an established capability used from elementary school through graduate school for purposes of instruction and assessment. Newer applications provide automated diagnostic feedback about student writing. Feedback includes errors in grammar, usage, and mechanics, comments about writing style, and evaluation of discourse structure. This paper reports on a system that evaluates a characteristic of lower quality essay writing style: *repetitious word use*. This capability is embedded in a commercial writing assessment application, *Criterion*^{SM}. The system uses a machine-learning approach with word-based features to model repetitious word use in an essay. System performance well exceeds several baseline algorithms. Agreement between the system and a single human judge exceeds agreement between two human judges.

1 Introduction

Automated evaluation of student essay writing is a rapidly growing field. Over the past few years, at least four commercially automated essay scoring systems have been made available (PEG;Page 1966; *e-rater*®Burstein et al., 1998; Intelligent Essay Assessor™;Foltz, Kintsch, and Landauer 1998; and, Intellimetric™; Elliot, 2003). In addition, based on the demands of users of the automated scoring technology, tools have been developed that perform more detailed evaluations of student writing. One such application is *Critique* Writing Analysis Tools. *Critique* and *e-rater* are embedded in a broader writing instruction application, *Criterion*^{SM} Online Essay Evaluation (see http://www.etstechnologies.com). *Critique* performs a number of evaluations on a student essay related to errors in grammar (Chodorow and Leacock, 2000), usage, and mechanics, comments on style, and analysis of essay-based discourse (organization and development) (Burstein et al, 2001 and Burstein and Marcu, 2003, and Burstein, Marcu and Knight , forthcoming).

Many of these capabilities use machine-learning approaches to model each particular kind of analysis. To develop such tools requires large sets of human annotated data, where judges have annotated information required to train a system to evaluate a particular kind of essay characteristic. For example, to build a capability to identify sentence fragments, a corpus of essay data needs to be annotated for this kind of ungrammatical sentence. A capability exists that identifies essay-based discourse elements in essays, for example, *thesis statements*, and *conclusions.* To do this, human judges annotated a corpus of essays for these particular kinds of discourse elements.

The judges' annotations were used to build an essay-based discourse analysis system.

Annotation protocols are required for each task. For identification of sentence fragments, this is reasonably straightforward. In terms of essay-based discourse analysis, it is fairly clear-cut. Though there is a certain amount of debate, annotators can be trained to have a reasonable amount of agreement in classifying essay-based discourse elements. *Style*, in contrast to grammar usage and discourse strategy, is tricky in terms of getting people to agree. It is a strongly subjective measure.

We discuss a system that identifies a specific characteristic of undesirable writing style --- *overly repetitious word usage.* Unlike identification of sentence fragments, and essay-based discourse strategy, there are no hard-and-fast rules that tell us how often a word must be used in an essay to be considered overly repetitious. The results reported in this paper indicate that even for a subjective style measure, human judges annotations can be modeled. The system can label repetitive words with precision, recall, and F-measures upwards of 0.90. It clearly outperforms all baseline methods described in the paper.

In earlier work with the writing instruction application, "Writer's Workbench," some features associated with style were evaluated, including: average word length, the distribution of sentence lengths, grammatical types of sentences (e.g., simple and complex), the percentage of passive voice verbs, and the percentage of nouns that are nominalizations (see MacDonald et al, 1982 for a complete description of the Writer's Workbench). In contrast to a subjective measure such as, repetitive word usage, the stylistic features in the Writer's Workbench are not subjective.

2 Approach

Since we want this system to model human judgements about overly repetitious word use, two human annotators labeled a corpus of essays. The decision-based machine learning algorithm, C5.0[1], was used to model the human judgements.

2.1 Human Annotation of Repetitious Word Use

As noted in the Introduction, the identification of good or bad writing style is highly subjective. With regard to word overuse in an essay, what one person may find irritating may not really bother someone else. Our goal in developing this tool was to indicate to students the cases in which word overuse might affect the rating of the paper with regard to its overall quality.

In the annotation protocol, the central guideline for the two human judges was to label as *repetitious* only those cases where the repetition of a word interfered with the overall quality of the essay. Both annotators were expert essay graders. They used a PC-based graphical user interface to label occurrences of repetitious words in a corpus containing 296 essays[2]. These essay data were randomly selected from a larger set of 5,000 essays. The final set contained essays from across several populations (6th grade through college freshman), and 11 test question topics.

2.2 Decision-Based Approach

We hypothesized, *a priori*, a number of features that could reasonably be associated with word overuse, such that the overuse interfered with a smooth reading of the essay. Our hypotheses were based on general discussions with the annotators before the annotation process began. The annotators are part of a team of experts who are critical in the decision-making process with regard to what kinds of feedback are helpful to students. We have on-going discussions with them that provide us with information about the kinds of

[1] For details about this software, see http://www.rulequest.com.

[2] Practical constraints (e.g., time and costs) did not allow for additional annotation.

issues that they are concerned about in student essay writing. Based on our hypotheses, we found that 7 features could be used in combination to reliably predict the word(s) in a student's essay that should be labeled as repetitious. These features are described below in Figure 1.

For each *lemmatized word token* in an essay, a vector was generated that contained the values for the 7 features. A stoplist is used, so that function words were excluded. A decision-based machine learning algorithm, C5.0, was used to model repetitious word use, based on human judge annotations.

1) Absolute Count: Total number of occurrences.
2) Essay Ratio: Proportional occurrence of the word in the essay (based on the total number of words in the essay).
3) Paragraph Ratio: Average proportional occurrence of the word in a paragraph (based on the average number of words in all paragraphs in the essay).
4) Highest Paragraph Ratio: Proportional occurrence of the word in the paragraph where it appears with the highest frequency (based on the number of words in the paragraph where it occurs most frequently).
5) Word Length: Total number of characters in a word.
6) Is Pronoun: Is the word a pronoun?
7) Previous Occurrence Distance: The distance between the word and its previous occurrence (based on number of words.)

Figure 1: Word-Based Features

3 Results

3.1 Human Performance

The results in Table 1a show agreement between the two human judges based on essays marked with repetition by one of the judges, at the word level. So, this includes cases where one judge annotated some repeated words and the other judge annotated no words as repeated. Each judge annotated overly repetitious word use in about 25% of the essays. In Table 1a, "J1 with J2" agreement indicates that Judge 2 annotations were the basis for comparison; and, "J2 with J1" agreement indicates that Judge 1 annotations were the basis for comparison. The Kappa between the two judges was 0.5 based on annotations for all words (i.e., repeated + non-repeated). Kappa indicates the agreement between judges with regard to chance agreement (Uebersax, 1982). Research in content analysis (Krippendorff, 1980) suggests that Kappa values higher than 0.8 reflect very high agreement, between 0.6 and 0.8 indicate good agreement, and values between 0.4 and 0.6 show lower agreement, but still greater than chance.

Figures 2 and 3 in the Appendix show annotated essays by each judge. These figures illustrate the kinds of disagreement on repeated words that exist between judges. The sample in Figure 2 shows annotations made by Judge 1, but not by Judge 2. Figure 3 shows an example where Judge 2 annotated words as repeated, but Judge 1 did not.

		Precision	**Recall**	**F-measure**
J1 with J2[3]	70 essays			
Repeated words	1,315	0.55	0.56	0.56
Non-repeated words	42,128	0.99	0.99	0.99
All words	43,443	0.97	0.97	0.97
J2 with J1[4]	74 essays			
Repeated words	1,292	0.56	0.55	0.56
Non-repeated words	42,151	0.99	0.99	0.99
All words	43,443	0.97	0.97	0.97

Table 1a: Precision, Recall, and F-measures Between Judge 1 (J1) and Judge 2 (J2)

[3] Precision = Total number J1 + J2 agreements ÷ total number J1 labels; Recall = Total number J1 + J2 agreements ÷ total number J2 labels; F-measure =2 * P * R ÷ (P + R).

[4] Precision = Total number J1 + J2 agreements ÷ total number J2 labels; Recall = Total number J1 + J2 agreements ÷ total number J1 labels; F-measure =2 * P * R ÷ (P + R).

In Table 1a, agreement on "Repeated words" between judges is somewhat low. *How can we build a system to reliably identify overly repetitious words if judges cannot agree?* If we look in the total set of essays identified by either judge as having some repetition, we find an overlapping set of 40 essays where both judges annotated the essay as having some sort of repetition. We call this the *agreement subset*.

Of the essays that Judge 1 annotated as having repetition, approximately 57% (40/70) agreed with Judge 2 as having some sort of repetition; of the essays that Judge 2 annotated with repetitious word use, about 54% (40/74) agreed with Judge 1. If we look at the total number of "Repeated words" labeled by each judge for all essays in Table 1a, we find that these 40 essays contain the majority of "Repeated words" for each judge: 64% (838/1315) for Judge 2, and 60% (767/1292) for Judge 1.

It is possible that even for the essays where judges both agree that there is some kind of repetitive word use, they do not agree on what the repetition is. Therefore, we want to answer the following question: *On the subset of essays where judges agree that there is repetition, do they agree on the same words as being repetitious?*

The core agreement with regard to "Repeated words" appears to be in these 40 essays. Table 1b shows high agreement between the two judges for "Repeated words" in the agreement subset. The Kappa between the two judges for "All words" (repeated + non-repeated) on this subset is 0.88. Figure 4 in the Appendix shows an example of an essay where both judges annotated the same words as repeated words.

		Precision	**Recall**	**F-measure**
J1 with J2	40 essays			
Repeated words	838	0.87	0.95	0.91
Non-repeated words	4,977	0.99	0.98	0.98
All words	5,815	0.97	0.97	0.97
J2 with J1	40 essays			
Repeated words	767	0.95	0.87	0.90
Non-repeated words	5,048	0.98	0.99	0.98
All words	5,815	0.97	0.97	0.97

Table 1b: Precision, Recall, and F-measure Between Judge 1 (J1) and Judge 2 (J2): "Essay-Level Agreement Subset"

3.2 System Performance

Table 2 shows agreement for repeated words between several baseline systems, and each of the two judges. Each baseline system uses one of the 7 word-based features used to select repetitious words (see Figure 1). Baseline systems label all occurrences of a word as repetitious if the criterion value for the algorithm is met. After several iterations using different values, the *final criterion value* (V) is the one that yielded the highest performance. The final criterion value is shown in Table 2. Precision, Recall, and F-measures are based on comparisons with the same sets of essays and words from Table 1a. Comparisons between Judge 1 with each baseline algorithm are based on the 74 essays where Judge 1 annotated repetitious words, and likewise, for Judge 2, on this judge's 70 essays annotated for repetitious words.

Using the baseline algorithms in Table 2, the F-measures for non-repeated words range from 0.96 to 0.97, and from 0.93 to 0.94 for all words (i.e., repeated + non-repeated words). The exceptional case is for Highest Paragraph Ratio Algorithm with Judge 2, where the F-measure for non- repeated words is 0.89, and for all words is 0.82.

To evaluate the system in comparison to each of the human judges, for each *feature combination algorithm*, a 10-fold cross-validation was run on each set of annotations for both judges. For each cross-validation run, a unique nine-tenths of the data were used for training, and the remaining one-tenth was used for cross-validating that model. Based on this evaluation, Table 3, shows agreement at the word level between each judge and a system that uses a different combination of features. Agreement refers to the mean agreement across the 10-fold cross-validation runs.

All systems clearly exceed the performance of the 7 baseline algorithms in Table 2. The best system is *All Features*, in which all 7 features are used. These results are indicated in ***italicized boldface*** in Table 3. It also indicates that building a model using the annotated sample from human judges 1 or 2 yielded indistinguishable results. For this reason, we arbitrarily used the data from one of the judges to build the final system.

When the *All Features* system is used, the F-measure = 1.00 for non-repeated words, and for all words for both "J1 with

Baseline Systems [5]	**V**	**J1 with System**			**J2 with System**		
		Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Absolute Count	19	0.24	0.42	0.30	0.22	0.39	0.28
Essay Ratio	0.05	0.27	0.54	0.36	0.21	0.44	0.28
Paragraph Ratio	0.05	0.25	0.50	0.33	0.24	0.50	0.32
Highest Paragraph Ratio	0.05	0.25	0.50	0.33	0.11	0.76	0.19
Word Length	8	0.05	0.14	0.07	0.06	0.16	0.08
Is Pronoun	1	0.04	0.06	0.04	0.02	0.03	0.02
Distance	3	0.01	0.11	0.01	0.01	0.10	0.01

Table 2: Precision, Recall, and F-measures Between Human Judges (J1 & J2)
& Highest Baseline System Performance for Repeated Words

Feature Combination Algorithms	**J1 with System**			**J2 with System**		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Absolute Count + Essay Ratio + Paragraph Ratio + Highest Paragraph Ratio (Count Features)	0.95	0.72	0.82	0.91	0.69	0.78
Count Features + Is Pronoun	0.93	0.78	0.85	0.91	0.75	0.82
Count Features + Word Length	0.95	0.89	0.92	0.95	0.88	0.91
Count Features + Distance	0.95	0.72	0.82	0.91	0.70	0.79
All Features: Count Features + Is Pronoun + Word Length + Distance	***0.95***	***0.90***	***0.93***	***0.96***	***0.90***	***0.93***

Table 3: Precision , Recall, and F-measure Between Human Judges (J1 & J2)
& 5 Feature Combination Systems for Predicting Repeated Words

[5] Precision = Total judge+ system agreements ÷ total system labels;
Recall = Total judge + system agreements ÷ total judge labels; F-measure = 2 * P * R ÷ (P + R).

System" and "J2 with System." Using *All Features*, agreement for repeated words more closely resembles inter-judge agreement for the *agreement subset* in Table 1b. It seems that the machine learning algorithm is capturing the patterns of repetitious word use in that set of 40 essays. Perhaps, an additional explanation as to why each judge has high agreement with the system, is that each judge is internally consistent.

4 Discussion and Conclusions

Teachers would generally prefer that students try to use synonyms in their writing, instead of the same word, repeatedly. Feedback about word overuse is helpful in terms of getting students to refine the use of vocabulary in their writing. Therefore, writing teachers would agree that it is an important capability in an automated essay evaluation system.

The evaluations presented in this paper show that a reliable repetitive word detection system can be built to model human annotations, even though this is a highly subjective writing style measure. An evaluation of our system indicates that it outperforms all baseline systems. It also has agreement with a single judge upward of 0.90 with regard to Precision, Recall and F-measures.

As research continues in automated essay scoring, it is standard to try to incorporate in a scoring system, any new features of writing that can be captured automatically. This new capability to identify repetitious word usage is currently being evaluated in terms of how it can contribute to better accuracy in an automated scoring system. Preliminary results indicate that the ability to detect if a writer is overusing certain vocabulary can contribute to the overall accuracy of the score from an automated essay scoring system. We are experimenting with the information about repetitious word usage in different discourse elements in an essay, e.g., *thesis statements*. In this case, the detection of repetitious words in these elements could contribute to a method for rating the overall quality of a particular element.

The repetitious word detection system was trained on annotated data across 11 test question topics; however, informal evaluations indicate that the system makes reasonable decisions on any topic. Though more systematic testing still needs to be done, the system appears to be topic-independent.

5 Acknowledgements

The authors would like to thank Claudia Leacock for advice on earlier versions of this paper. This work was completed while both authors were affiliated with ETS Technologies, Inc, formerly a wholly-owned subsidiary of Educational Testing Service. ETS Technologies is currently an internal division of Educational Testing Service.

References

Burstein, Jill, Marcu, Daniel, and Knight, Kevin (forthcoming). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. Special Issue on Natural Language Processing of IEEE Intelligent Systems, January/February, 2003.

Burstein, J. and Marcu D. (2003). Developing Technology for Automated Evaluation of Discourse Structure in Student Essays. In M. Shermis and J. Burstein (eds.), *Automated essay scoring: A cross-disciplinary perspective*, Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Burstein, J., Marcu, D., Andreyev, S., and Chodorow, M. (2001). Towards Automatic Classification of Discourse Elements in Essays. *In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, July, 2001.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., and Harris M. D. 1998. Automated Scoring Using A Hybrid Feature Identification Technique. *Proceedings of 36th Annual Meeting of the Association for*

Computational Linguistics, 206-210. Montreal, Canada.

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

Elliott, S. (2003). Intellimetric: From Here to Validity. In M. Shermis and J. Burstein (eds.) *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Foltz, P. W., Kintsch, W., and Landauer, T. K. 1998. Analysis of Text Coherence Using Latent Semantic Analysis. *Discourse Processes* 25(2-3):285-307.

Krippendorff K. (1980). Content Analysis: An Introduction to Its Methodology. Sage Publishers.

MacDonald, N. H., Frase, L.T., Gingrich P.S., and Keenan, S.A. (1982). The Writer's Workbench: Computer Aids for Text Analysis. IEEE Transactions on Communications. 30(1):105-110.

Page, E. B. 1966. The Imminence of Grading Essays by Computer. *Phi Delta Kappan*, 48:238-243.

Uebersax, J.S. (1982) "A Generalized Kappa Coefficient," Educational and Psychological Measurement, Vol. 42, pp. 181-183.

Appendix: Sample Human Judge Annotations for Repeated Words, In UPPER CASE BOLDFACE

THE BEST PET

Did **YOU** ever have a pet that **YOU** thought was the best thing that **YOU** ever had. I am going to tell **YOU** about a pet that I thought was the best.

The best pet I thought was the best was a pit bull. **THEY** are very easy to tran, **THEY** are competetive. **THEY** are very strong, and good pets. Thet do not turn on you if you fight them. **THEY** can protect things very well. **THEY** are alwas good to have.

Figure 2: Sample Annotated Essay from Judge 1 Which Judge 2 Did Not Identify

SHORTS

The question here is what I think about, not being allwoed to wear **SHORTS**. I think we should be allowed to wear **SHORTS**. Imean what is the big deal. I know us girls can get our **SHORTS** pretty **SHORT**, but we can also get skirts pretty **SHORT** too. So we should just have the same rules for skirts. Pretty soon we can't wear skirts. Well this get's me on another thing. We can't wear capris! I know this isn't about capris, but still they go down to your knees that dosn't make since.

Boys should be able to wear those long **SHORTS** that dosn't show anything. Well I don't know. Maybe it's good we can't wear **SHORTS**. I don't know, Im just a teenager.

Figure 3: Sample Annotated Essay from Judge 2 Which Judge 1 Did Not Identify

One major **SCHOOL** issue that we students face daily is the subject of **SCHOOL** safety. Many **SCHOOLS** across the country have encountered **SCHOOL VIOLENCE**. I think that most **SCHOOL VIOLENCE** starts with the **SCHOOL** and the community. Students who engage in **SCHOOL VIOLENCE** are usually made fun of or are insecure about themselves. Some ways that I think that we can stop **SCHOOL** follow. I think that in order to stop **SCHOOL VIOLENCE** in and around our communities we have to get the community involved in sharing and making it aware to other cities and towns that **SCHOOL VIOLENCE** is very real, and we face it everyday. One way I think that we can cut down on **SCHOOL VIOLENCE** is to have striter disapline policies. When students in a **SCHOOL** joke around or threaten other students about killing them, or bringing weapons to **SCHOOL**, the staff of that **SCHOOL** needs to take action. When a student has thought out a plan to kill others, they obviously need to be talked to. I hope that by reading these ways to stop **SCHOOL VIOLENCE** we can all take action to make our **SCHOOLS** safer. We can not stop **SCHOOL VIOLENCE** until we stop blaming others, and see that we too have overlooked **SCHOOL VIOLENCE**. **SCHOOL VIOLENCE** is a major **SCHOOL** issue that everyone can stop, if we all try to help.

Figure 4: Sample Essay Where Both Judges Agree On Repeated Words