Predicting Human Scores of Essay Quality Using Computational Indices of Linguistic and Textual Features

Scott A. Crossley[1], Rod Roscoe[2], and Danielle S. McNamara[2]

[1] Department of Applied Linguistics, Georgia State University, 34 Peachtree St. Suite 1200, One Park Tower Building, Atlanta, GA 30303, USA
scrossley@gsu.edu
[2] Institute for Intelligent Systems, The University of Memphis, FedEx Institute of Technology, Memphis, TN 38152
rdroscoe@memphis.edu, dsmcnamra1@gmail.com

Abstract. This study assesses the potential for computational indices to predict human ratings of essay quality. The results demonstrate that linguistic indices related to type counts, given/new information, personal pronouns, word frequency, conclusion n-grams, and verb forms predict 43% of the variance in human scores of essay quality.

1 Introduction

In educational settings, trained, professional readers (e.g., teachers) typically assess writing quality. These evaluations have important consequences for the writer because these judgments provide a source of feedback and determine passing or failing grades. The goal of this study is to investigate the linguistic and textual features in argumentative essays that influence human judgments of writing quality. This approach is in contrast to writing research that primarily investigates cognitive and behavioral processes that occur during writing (i.e., planning, translating, reviewing, and revising) but not the products of writing [1], such as the linguistic features of a text [2]. However, linguistic features at the word, syntactic, and discourse levels have been found to significantly influence essay quality, and can be important indicators of writing development [3].

A better understanding of the relationships between linguistic features and writing quality has several benefits. This knowledge may help writers to make more informed decisions about effective writing and composition. Such knowledge would also help readers and teachers make more accurate or specific evaluations of writing quality, which would enable them to provide more precise or targeted feedback.

In this study, we use computational linguistic indices to assess human ratings of essay quality. Because these linguistic and textual analyses are automated, they can be implemented within computer systems that automate the process of assessing writing and providing student feedback. Thus, this research informs both writing pedagogy and instructional technology (e.g., intelligent tutoring systems).

G. Biswas et al. (Eds.): AIED 2011, LNAI 6738, pp. 438–440, 2011.

2 Methodology

We collected 314 timed (25-minute) essays written by 314 college freshmen at a large university in the United States. All essays were written in response to two Scholastic Achievement Test (SAT) writing prompts. We separated the corpus into a training (n = 209) and test set (n = 105) based on a 67/33 split. The training set was used to select the computational indices for the initial statistical analyses (correlations and regression analyses). The test set was used to calculate the predictive ability of the selected variables in an independent corpus.

Expert raters rated the quality of the 314 essays in the corpus using a standardized SAT rubric for holistic quality. The final interrater reliability for all essays in the corpus was $r > .75$. We used the mean score between the raters as the final value for the quality of each essay unless the differences between the 2 raters was >= 2, in which case a third expert rater adjudicated the score.

The linguistic features of the essays were analyzed using Coh-Metrix indices [4]. We selected indices from Coh-Metrix with theoretical and empirical links to essay quality and writing proficiency. These indices were organized into broad measures that reflected general linguistic constructs: cohesion, lexical sophistication, syntactic complexity, rhetorical strategies, and text structure. Cohesion measures included causality, incidence of connectives, incidence of logical operators, lexical overlap, semantic co-referentiality, anaphoric reference, prompt overlap, and paragraph overlap. Lexical sophistication measures included word hypernymy, word polysemy, academic words, lexical diversity, word frequency and word information indices (e.g., word concreteness, familiarity, meaningfulness, and imagability). Syntactic complexity measures included syntactic similarity and phrase structure complexity. Rhetorical strategies measures included indirect pronouns, amplifiers, downtoners, exemplification and n-gram indices for rhetorical phrases common in high quality introductory, body, and concluding paragraphs.

3 Results

We selected the computational indices that demonstrated the highest Pearson correlation when compared to the human essay scores, and that did not demonstrate multicollinearity. This led to the selection of 26 variables.

A linear regression analysis was conducted with the 26 variables. These 26 variables were regressed onto the raters' score for the 209 essays in the training set, and were checked for outliers and multicollinearity. The linear regression yielded a significant model, $F(6, 200) = 23.202$, $p < .001$, $r = .641$, $r^2 = .410$. Six variables were significant predictors: total types, LSA given/new, incidence of personal pronouns, word frequency, all n-grams (conclusion paragraphs), and incidence of verb base form. The model for the test set using these variables yielded $r = .655$, $r^2 = .429$.

4 Discussion

This study demonstrated that a combination of computational indices related to type counts, given/new information, incidence of personal pronouns, word frequency,

incidence of n-grams related to conclusion quality, and incidence of verb base form explained 43% of the variance in human judgments of essay quality. This is a two-fold increase in predictive power over previous findings [3] and provides further evidence that computational indices can be used to assess essay quality.

These linguistic indices allow us to better understand how textual features influence human judgments of writing quality. As in past studies, longer essays (i.e., greater number of word types) that use more sophisticated vocabulary (i.e., less frequent words), and more complex grammar (i.e., fewer base verb forms) were judged higher in quality. In contrast to past studies, higher quality essays in this analysis also displayed more cohesion such that essays judged higher in quality maintained stronger links to previously given information. Our model also reported a positive relationship between essay quality and the incidence of conclusion n-grams (e.g., concluding phrases, conditionals, and modals) indicating that the presence of rhetorical elements is important in judgments of essay quality. Additionally, lower quality essays used more personal pronouns suggesting that weaker writers relied more on writer-based prose than reader-based prose.

Advancing research on automated linguistic analysis enhances our ability to detect and understand the textual features that contribute to effective writing. In turn, this empowers us to teach developing writers how to harness such knowledge to further their academic and professional goals, both via traditional feedback given by teachers, and by automated feedback and strategies taught by intelligent tutoring systems.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

References

1. Abbott, R., Berninger, V., Fayol, M.: Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. Journal of Educational Psychology 102, 281–298 (2002)
2. Berninger, V., Mizoka, D., Bragg, R.: Theory-based diagnosis and remediation of writing disabilities. Journal of School Psychology 29, 57–79 (1991)
3. McNamara, D.S., Crossley, S.A., McCarthy, P.M.: Linguistic features of writing quality. Written Communication 27, 57–86 (2010)
4. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behavioral Research Methods, Instruments, and Computers 36, 193–202 (2004)