Paragraph Specific N-Gram Approaches to Automatically Assessing Essay Quality

Scott Crossley
Georgia State University
34 Peachtree Ave, Ste 1200
Atlanta, GA 30303
01+404-413-5179
scrossley@gsu.edu

Caleb DeFore
Georgia State University
34 Peachtree Ave, Ste 1200
Atlanta, GA 30303
01+404-413-5200
cdefore1@student.gsu.edu

Kris Kyle
Georgia State University
34 Peachtree Ave, Ste 1200
Atlanta, GA 30303
01+404-413-5200
kkyle3@student.gsu.edu

Jianmin Dai
Arizona State University
PO Box 872111
Tempe, AZ 85287
01+404-413-5200
Jianmin.Dai@asu.edu

Danielle S. McNamara
Arizona State University
PO Box 872111
Tempe, AZ 85287
01+404-413-5200
dsmcnamara1@gmail.com

ABSTRACT

In this paper, we describe an n-gram approach to automatically assess essay quality in student writing. Underlying this approach is the development of n-gram indices that examine rhetorical, syntactic, grammatical, and cohesion features of paragraph types (introduction, body, and conclusion paragraphs) and entire essays. For this study, we developed over 300 n-gram indices and assessed their potential to predict human ratings of essay quality. A combination of these n-gram indices explained over 30% of the variance in human ratings for essays in a training and testing corpus. The findings from this study indicate the strength of using n-gram indices to automatically assess writing quality. Such indices not only explain text-based factors that influence human judgments of essay quality, but also provide new methods for automatically assessing writing quality.

Keywords

Essay quality, computational linguistics, corpus linguistics, automatic feedback, intelligent tutoring systems.

1. INTRODUCTION

Academic success often depends on a student's writing proficiency [1]. Unfortunately, for many students, such proficiency is often difficult to attain and frequently remains elusive throughout schooling [5]. One major problem in the teaching of writing skills is that students have limited opportunities to write and receive feedback from teachers and peers. Such a problem is related to time constraints inside and outside of the classroom [5], which minimize opportunities for students and teachers to interact one on one. A potentially profitable approach to providing students with greater access to writing opportunities and ensuring that students receive feedback on their writing is through the use of automatic writing evaluation (AWE) systems that provide students with the opportunities to write essays and automatically receive feedback on the quality of their writing.

However, AWE systems often lack the sensitivity to respond to a number of features in student writing and, more specifically, to those features that relate to instructional efficacy [8]. Our goal in this study is to investigate the potential for n-gram indices related to paragraph types (i.e., introduction, body, and conclusion paragraphs) to predict human judgments of essay quality. We are interested in paragraph specific indices because developing writers need to focus on and learn strategies for building quality introduction, body, and conclusion paragraphs. If we can identify n-grams in quality essays that relate to paragraph building strategies and to human judgments of writing quality, then we can use these n-gram indices to assign automatic scores to essays. In addition, such indices may prove beneficial in providing automated formative feedback to users that directly link to instructional strategies (i.e., strategies for building stronger paragraphs).

1.1 The Writing Pal

The Writing Pal (W-Pal) is an intelligent tutoring system (ITS) that contains an AWE system in order to provide summative and formative feedback to users [4]. However, unlike strict AWE systems, W-Pal adopts a pedagogical focus by providing writing strategy instruction to users. Thus, unlike AWE systems, which focus on essay practice with some support instruction, W-Pal emphasizes strategy instruction and targeted strategy practice prior to whole-essay practice. The writing strategies cover the three phases of the writing process: prewriting, drafting, and revising. Each of the writing phases is further subdivided into instructional modules. These modules include *Freewriting* and *Planning* (prewriting); *Introduction Building*, *Body Building*, and *Conclusion Building* (drafting); and *Paraphrasing*, *Cohesion Building*, and *Revising* (revising). In W-Pal, students view lessons on each strategy, play practice games, and then write practice essays for each of the modules.

Essay writing is an essential component of W-Pal. As a result, the system includes an essay-writing interface, which allows students to compose essays. These essays are then analyzed by the W-Pal AWE system, which is used to provide automated formative and summative feedback to the participants based upon natural language input and hierarchical classification as compared to regression analyses. Such hierarchical classification affords the opportunity to provide feedback at different conceptual levels on a variety of linguistic and rhetorical features [2].

In general, the feedback in W-Pal focuses on the strategies taught in W-Pal lessons (e.g., *Conclusion Building*, P*araphrasing*, and *Cohesion Building*) and practice games and is primarily based on linguistic features reported by the AWE scoring model. For instance, if a student produces an essay that is too short, the system will provide feedback to the student suggesting the use of idea generation techniques such as those found in the freewriting module. If a student essay does not meet the paragraph threshold, the W-Pal feedback system will suggest techniques to plan and organize the essay more effectively including outlining and focusing on structural elements such as positions, arguments, and evidence (all elements taught in the instructional modules). Such feedback can be general (e.g., asking students to condense similar sentences, restructure sentences, and improve cohesion), but it can also be more specific and remind students to preview their thesis statements and arguments in the introduction paragraph, write concise topic sentences and present evidence in body paragraphs, and provide conclusion statements and restate the thesis in the concluding paragraph. The feedback system in W-Pal has proven effective in prior writing studies [6] demonstrating that essays revised using W-Pal feedback are scored significantly higher than their original drafts (as assessed by an automatic scoring algorithm).

However, in practice, the feedback provided by the W-Pal AWE system to W-Pal users can be repetitive and overly broad [6]. For instance, students often receive the same feedback from the AWE as they continue to submit drafts and revisions of papers over time. The repetition in the AWE systems is a product of the general nature of much of the feedback provided by the system and is a direct reflection of the specificity of many of the linguistic indices found in the NLP scoring algorithms used by W-Pal. These algorithms are often informed by linguistic features that, while predictive of essay quality, are not highly useful in providing feedback to users. For instance, the current algorithm includes many indices related to lexical sophistication and syntactic complexity, both of which are important indicators of essay quality [3]. However, feedback at such a fine-grain level of linguistic analysis (e.g., use more infrequent words or produce more sentences that include infinitive forms) is not very practical, helpful, or formative. As a result, much of the feedback given to W-Pal users is necessarily general in nature and could potentially hinder students' ability to utilize the feedback effectively.

2. METHODS

Our goal in this study is to develop paragraph specific n-gram indices to automatically assess the essay quality of student writers in the ITS W-Pal. The purpose of these indices is to provide potentially stronger links between the instructional modules in W-Pal and the automatic scores assigned to essays by the AWE system. If practical and specific elements of texts related to essay quality can be developed, then these elements, in turn, could also inform feedback mechanisms and potentially provide better connections between the instructional modules in W-Pal (i.e., Introduction Building, Body Building, and Conclusion Building) and formative feedback concerning these modules.

2.1 Corpus

The corpus we used to develop the n-gram indices comprised 1123 argumentative (persuasive) essays. Because our interest is in developing automated indices that are predictive across a broad range of prompts, grade levels, and temporal conditions, we selected a general corpus that contained 16 different prompts, three different grade levels (10^{th} grade, 11^{th} grade, and college freshman), and two different temporal conditions (essay that were untimed and essays that were written in 25-minute increments).

Not all the essays from this corpus were used to develop the n-gram indices. Only those essays that contained at least three paragraphs were selected to develop the n-gram indices. Such essays provide some evidence that the writer had produced an introduction, body, and conclusion paragraph affording the opportunity to examine paragraph specific n-grams. After removing all essays that contained fewer than 3 paragraphs, we were left with 971 essays. We used these essays to develop the n-gram indices. We used the essays in the entire corpus (N =1123) to train a regression model.

We tested the training regression model on a test set of argumentative essays that were not used in the developmental process. The essays were written by participant in a W-Pal study. They ranged in grade level from 9^{th} to 12^{th} (M = 10.2, SD = 1.0). Each participant wrote a pretest and a posttest essay (N = 128). The essays were written within the W-Pal essay-writing interface.

2.2 Human judgments

Each essay in the developmental corpus and the test set was scored independently by two expert raters using a 6-point rating scale developed for the Scholastic Aptitude Test. The rating scale was used to holistically assess the quality of the essays and had a minimum score of 1 and a maximum score of 6.

2.3 N-gram indices

To develop the n-gram indices, we first separated the paragraphs in all the essays that contained three or more paragraphs based on sequential positioning. All initial paragraphs were classified as introductory paragraphs; all middle paragraphs were classified as body paragraphs; and all final paragraphs were classified as conclusion paragraphs. Each paragraph was further classified as low quality (i.e., average essay score of 3 or less) or high quality (i.e., average essay score of 3.5 or greater).

The paragraphs for each position and quality rating were then analyzed using WordSmith [7] to identify key n-grams (unigrams, bigrams, and trigrams). Two expert raters then identified linguistic patterns among the key n-grams and used these linguistic patterns to classify the n-grams into linguistic groupings related to rhetorical, grammatical, syntactic, and cohesion features. N-grams were organized in the groupings based on strength of keyness. However, if a unigram was a keyword and that unigram was also included within a key bi-gram or tri-gram, the bi-gram or tri-gram was removed if it had a lower keyness value. The selected n-gram groupings are briefly discussed below.

2.3.1 Introductory Paragraphs

Twenty groupings of n-grams were identified for the introductory paragraphs. These groupings were based mostly on rhetorical features, but also include cohesion, syntactic, and grammatical features.

2.3.2 Body Paragraphs

Twenty-seven groupings of n-grams were identified for the body paragraphs. These groupings were based mostly on rhetorical features, but also include cohesion, syntactic, and grammatical features.

2.3.3 Conclusion Paragraphs

Twenty-five groupings of n-grams were identified for the conclusion paragraphs. These groupings were based mostly on rhetorical features, but also include cohesion and syntactic features.

2.4 Analyses

For each n-gram grouping, we calculated an incidence score and a proportion score for the n-grams in the grouping for each paragraph type (i.e., introduction, body, and conclusion paragraphs) and for the essay as a whole. We also combined all of the positive and all of the negative n-grams into separate indices and computed their incidence in the paragraph types and for the essays as a whole. These incidence and proportion scores became our automated indices for the subsequent regression analysis. Within each essay, all body paragraphs were pooled and treated as a single entity.

We used the essays in the entire corpus to create regression models to predict the human ratings for the essays. We first conducted correlations between the index scores and the human ratings of essay quality. We selected all those variables that demonstrated at least a small effect size ($r > .10$) and did not demonstrate strong multicollinearity with one another or with text length ($r < .899$). The model from this regression analysis was then extended to the essays in the testing corpus to examine how well the model predicted essay quality in an independent corpus.

3. Results

3.1 Multiple Regression All Essays

Of the 316 n-gram grouping indices calculated for this study, 163 of the indices demonstrated at least a small effect size with the human ratings of essay quality ($p < .001$) for all the essays in the corpus. Of these, four demonstrated strong correlations with text length and were removed. Lastly, six indices demonstrated strong multicollinearity with other indices and were removed, leaving 153 indices.

The linear regression using the selected variables yielded a significant model, $F(20, 1102) = 32.925$, $p < .001$, $r = .612$, $r^2 = .374$. Twenty variables were significant predictors in the regression. The remaining variables were not significant predictors and were either not included in the model or were removed in the steps of the model (in the case the index *Body all positive* grouping index). The regression model is presented in Table 1. We used the B weights and the constant from the regression analysis to assess the model on an independent data set (the 128 essays from the W-Pal efficacy study). The model for the test set yielded $r = .576$, $r^2 = .332$.

4. Discussion

This study demonstrates that n-gram indices related to rhetorical, grammatical, and cohesion feature of a text can be strongly predictive of human judgments of essay quality. These n-grams were calculated at the paragraph level and at the text level. The indices were tested on essays that contained as few as 1 to 2 paragraphs and on essays that contained only 3 or more paragraphs. The results of this study provide models of essay quality that could be implemented in an AWE system to provide increased accuracy of summative feedback (i.e., holistic scores). Because many of the n-gram indices are paragraph specific and many of them are related to rhetorical or cohesion patterns (as compared to syntactic and grammatical patterns), the indices are expected to provide more specific feedback to users within the W-Pal system that will be both more practical and more useful. The feedback that is based on these indices can be linked to instructional modules within the W-Pal system.

The regression model demonstrated that the combination of the 20 variables accounts for 37% of the variance in the human evaluations of overall writing quality. The most predictive indices were generally the combined n-gram indices that integrated all the

Table 4: Linear regression results for all essays

Entry	Variable Added/Removed	Correlation	R-Squared	*B*	SE	B
Entry 1	Body all positive	0.474	0.225	Removed	Removed	Removed
Entry 2	Body all positive proportion	0.510	0.260	0.766	0.174	0.199
Entry 3	Conclusion all positive	0.527	0.278	0.067	0.010	0.223
Entry 4	Conclusion all negative	0.548	0.300	-0.013	0.005	-0.087
Entry 5	Body adverbs positive proportion	0.562	0.316	0.397	0.074	0.132
Entry 6	Body connectives positive essay	0.568	0.322	0.017	0.004	0.130
Entry 7	Remove body all positive	0.566	0.321	-	-	-
Entry 8	Introduction stance negative	0.573	0.328	-0.089	0.028	-0.088
Entry 9	Conclusion all negative proportion	0.578	0.334	-0.823	0.210	-0.149
Entry 10	Introduction choice negative	0.583	0.339	-0.250	0.084	-0.079
Entry 11	Body general references positive essay	0.588	0.345	0.041	0.012	0.090
Entry 12	Body 3rd person negative	0.590	0.348	-0.019	0.007	-0.079
Entry 13	Introduction all negative proportion	0.593	0.351	-0.406	0.172	-0.112
Entry 14	Body casual positive	0.595	0.354	0.232	0.095	0.059
Entry 15	Body quantity positive essay	0.597	0.356	0.016	0.006	0.070
Entry 16	Introduction totality positive essay	0.599	0.359	-0.033	0.013	-0.075
Entry 17	Conclusion set membership positive essay	0.602	0.362	0.054	0.023	0.060
Entry 18	Body tense positive	0.604	0.364	0.041	0.017	0.063
Entry 19	Introduction 2nd person negative	0.606	0.367	0.038	0.015	0.071
Entry 20	Conclusion 1st person positive essay	0.608	0.369	-0.020	0.010	-0.052
Entry 21	Introduction conditionals negative	0.610	0.372	-0.067	0.032	-0.059
Entry 22	Introduction comparison positive essay	0.612	0.374	0.158	0.076	0.055

Notes: Estimated Constant Term is 2.563; *B* is unstandardized Beta; SE is standard error; B is standardized Beta
Note: Essay is n-gram count across the entire essay. All other n-gram counts across the paragraph types.

positive or negative n-grams for the paragraph type. For instance, positive body n-grams and positive and negative conclusion n-grams were the strongest predictors of essay quality (predicting almost 30% of the variance in the human ratings alone) followed by negative introduction n-grams. The remaining indices were more specific in nature and included six introduction n-gram indices (related to stance, choice, totality, 2nd person, conditionals, and comparison), seven body n-gram indices (related to adverbs, connectives, general reference, 3rd person, causality, quantity, and tense), and two conclusion n-gram indices (related to set membership and first person). The majority of these indices were measured at the paragraph level with 7 of the 20 indices measured across the text. Because this analysis included essays with only 1 or 2 paragraphs, we presume that conclusion n-gram indices were less predictive insomuch as many essays would not contain a second or third paragraph that would act as a conclusion.

From a linguistic perspective, this study has demonstrated that rhetorical features of paragraphs are important indicators of essay quality. The majority of the n-gram indices that loaded into our regression models were rhetorical in nature. For instance, the use of adverbs such as *yet, unfortunately,* and *completely* are important indicators of writing proficiency demonstrating that better writers use a greater number of such adverbs. High quality essays also contain fewer negative stance n-grams in the introduction (e.g., *I think, know, feel that*). Good writers also use more general reference terms such as *these* and *those*, indicating that referencing previous noun phrases is an important indicator of writing quality. Such an index may also relate to the cohesive properties of the text and, in support, this study also reports that other cohesive features loaded into our regression models. For instance, positive n-gram connectives (i.e., *however, and*) found in the body are significant predictors. Unlike rhetorical and cohesive n-gram indices, no syntactic indices loaded into our regression model and only one grammatical n-gram index loaded (positive body tense n-grams). Such a finding does not diminish the importance of syntactic and grammatical features in essay writing, but rather demonstrates that an n-gram approach likely does not capture the complexity needed to assess such features.

We envision that these n-gram indices could be used to provide formative feedback to users in an ITS. For instance, these n-gram indices directly overlap with instruction modules in W-Pal (i.e., introduction building, body building, and paragraph building) and would thus link with the writing strategies with which users become familiar during training. The indices are also much more paragraph specific than current feedback algorithms in W-Pal, which focus on general feedback concerning relevance to topic, essay structure, paragraph structure, and revising strategies. For example, the current feedback reminds users to attend to structural elements in paragraphs such as positions, arguments, and evidence. However, the feedback algorithms do not provide specific linguistic features to which to attend. We envision that the n-gram indices discussed in this study could provide useful and specific formative feedback to assist in student essay revision. For instance, users could be given specific feedback about their use of adverb, general reference, connective, quantity, and tense n-grams in their body paragraphs. Users could also receive direct and specific feedback on their use of set membership words and 1st persons in their conclusion. This feedback would be based on concrete linguistic features in the text and would provide rhetorical, cohesion, and grammatical information to the user that could be exploited during the revision process.

5. Conclusion

While strongly predictive, the n-gram indices investigated here should be examined in conjunction with more traditional linguistic indices that have demonstrated predictive power in explaining essay quality (i.e., lexical, syntactic, and cohesive features of text; [3]). Such an analysis would assess how predictive the n-gram indices are when combined with other variables. More importantly, the indices should be tested to examine the degree to which they are able to provide more direct and specific formative feedback and the effects of such feedback on essay revision and quality.

6. 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.

7. REFERENCES

[1] Kellogg, R. and Raulerson, B. 2007. Improving the writing skills of college students. *Psychonomic Bulletin and Review. 14*, 237-242.

[2] Crossley, S., Roscoe, R., and McNamara, D. (in press). Using natural language processing algorithms to detect changes in student writing in an intelligent tutoring system. Manuscript submitted to the *26th International Florida Artificial Intelligence Research Society Conference*.

[3] McNamara, D., Crossley, S., and Roscoe, R. 2013. Natural language processing in an intelligent writing strategy tutoring system. *Behavioral Research Methods, Instruments and Computers.* Advance online publication.

[4] McNamara, D., Raine, R., Roscoe, R., Crossley, S., Jackson, G., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Carney, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P., and Graesser, A. 2012. The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 298-311). Hershey, P.A.: IGI Global.

[5] National Commission on Writing. 2003. *The Neglected "R."* College Entrance Examination Board, New York.

[6] Roscoe, R., Kugler, D., Crossley, S., Weston, J., and McNamara, D. S. 2012. Developing pedagogically-guided threshold algorithms for intelligent automated essay feedback. In P. McCarthy & G. Youngblood (Eds.), *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference* (pp. 466-471). Menlo Park, CA: The AAAI Press.

[7] Scott, M. 2008. WordSmith Tools version 5, Liverpool: Lexical Analysis Software.

[8] Shermis, M.D., Burstein, J.C. and Bliss, L. 2004. The impact of automated essay scoring on high stakes writing assessments. *Paper presented at the annual meeting of the National Council on Measurement in Education*, April 2004, San Diego, CA