EasyChair Preprint

№ 6196

A Differential N-gram Use Measure for Automated Essay Scoring

Qian Wan, Scott Crossley, Laura Allen and Danielle McNamara

July 31, 2021

A Differential N-gram Use Measure for Automated Essay Scoring

Qian Wan[1], Scott Crossley[1], Laura Allen[2] and Danielle McNamara[3]

[1]Department of Applied Linguistics and ESL, Georgia State University

[2]Department of Psychology, University of New Hampshire

[3]1Department of Psychology, Arizona State University

Author Note

The authors declare that there no conflicts of interest with respect to this preprint.

The research reported here was supported by the Chan Zuckerberg Initiative, the Bill & Melinda Gates Foundation, and Schmidt Futures through grants to Georgia State University. Additional funding was provided by the Institute of Education Sciences, U.S. Department of Education, and the Office of Naval Research, through Grants R305A180261, R305A180144, N00014-20-1-2623, N00014-19-1-2424 to Arizona State University and University of New Hampshire. The opinions expressed are those of the authors and do not represent views of the funding agencies.

Correspondence should be addressed to Qian Wan, Georgia State University, Department of Applied Linguistics and ESL; 15th floor, 25 Park Place, Atlanta, GA 30303, United States. Email: qwan1@gsu.edu

Abstract

This study implements and expands a differential word use (DWU) measure for automatically assessing the quality of independent and integrated student essays. For independent writing samples, the baseline unigram DWU measure successfully predicted human scores in a training corpus but was less generalizable to a test corpus. For integrated writing, tri-gram measures yielded the best performance, and using n-grams as a unit of feature development improved the performance on both training and testing corpora. This research contributes to our understanding of discourse by validating and expanding current DWU measure and by showing to which extent vocabulary use is related to the quality of written discourse in varied tasks.

Keywords: written discourse, automated essay evaluation, vocabulary-use

A Differential N-gram Use Measure for Automated Essay Scoring

Vocabulary use is considered an important element of essay quality; consequently, such measures are often incorporated into automated essay scoring systems (Attali, 2007; Attali & Burstein, 2006; Attali et al., 2010). Under the assumption that words appearing more frequently in high-quality essays are indicative of more sophisticated vocabulary and content, differential word use (DWU) metrics calculate the relative frequency of words in low- and high-quality essays. The average of these relative frequency values across all words in an essay can then be used to automatically calculate essay scores (Attali, 2011). However, this approach has only been used on specific prompts and not been generalized to multiple prompts by removing prompt-specific words in the development of the measures. Further, DWU measures have not been extended to multi-word units and part of speech (POS) tag, nor have the DWU measures been built and validated on datasets other than TOEFL or GRE essays.

The current study attempts to address these gaps by examining the performance of more generalizable DWU measures. We first explore the development of prompt-general DWU measures, then expand DWU measures from single words to bi-grams and tri-grams and POS tags of those n-grams (referred hereafter as DNU measures). Finally, we examine the extent to which these features correlate with essay quality for both independent and integrated writing in test sets that contain essays written by non-native English speakers.

Method

Data

Independent Essays. The training set of independent writing consists of 5852 persuasive essays written by mostly native speakers of English in response to 15 specific prompts. Most participants were students ranging in grade levels from 7-10 or freshman college students. The essays were evaluated by human raters following the SAT holistic rubric. The holistic scores (ranged from 0-6) were the average scores given by two raters. In most cases, when two human scores differed by 2, they adjudicated their scores.

A test set was used to evaluate the performance of the DWU/DNU measures, which contained 480 TOEFL independent essays written by non-native speakers of English in response to two different prompts.

Integrated Essays. The training set of integrated essays was comprised of 2429 source-based essays (written on 7 prompts) that were collected from a diverse group of participants who were generally native speakers of English. For each prompt, participants were given 4-7 source texts that provided information about the topics. The participants were asked to read the source texts and write an integrated essay to support their ideas. The essays were rated by two human raters based on a holistic rubric on a 1-6 scale. For each essay, the final holistic score was the average score provided by the two raters. Inter-rater agreement was higher than .6 for all the essays.

The test set of integrated essays contained 480 TOEFL essays (in response to two prompts) written by the same writers as in the independent test corpus. All the TOEFL essays were rated by at least two expert raters following TOEFL IBT writing rubrics on a 0-5 scale.

Differential Word/N-gram Measures

This paper followed the method in Attali (2011) to develop the DWU/DNU measures that calculate the relative frequency of a word (or n-grams/POS n-grams) in low- and high-scored essays (i.e., d-values) and then calculate a DWU/DNU score for each essay by adding the d-values of all the words (or n-grams/POS n-grams) appearing in the essay. In the current study, a low-quality essay is defined as scored between 0-2, and a high-quality essay being defined as scored between 4-6. To compute the differential frequency of the words (d-values), we count, for each word (indexed *i*), its frequency in the low- and high-scored essays (f_{il} and f_{ih}), and then compute the differences of log-transformed relative frequencies of the words:

$$d_i = \log (f_{ih} / f_h) - \log (f_{il} / f_l)$$

In this way, a d_i value of zero indicated that a word is equally likely to appear in a low- or high-scored essay. Then we calculate the average d_i values of all words that occurred in the target essay.

The d-values for all words were extracted from the independent and integrated training sets. DWU/DNU measures based on these d-values were then computed for essays in both the training and test sets.

We implemented the unigram DWU measures in Attali (2011) as a baseline. We then removed all prompt-specific words from the training sets and re-calculated the d-values and computed the generic unigram measures for all essays. Additionally, instead of using single words as a unit for d-values, we developed the DNU measures using bi-grams, tri-grams, and the POS tags of these n-grams, assuming the n-grams appearing more frequently in high-quality essays were indicative of more sophisticated discourse and higher essay quality.

Removal of Prompt-specific Words

To develop the generic unigram DWU measures, we automatically extracted prompt-based words by comparing the keyness values (Kilgarriff, 2001) of words among essays from different prompts such that the generic DWU measure ignored prompt-specific words.

Keyness values provided evidence of whether a word is more common in one corpus compared with the other corpus. To identify the prompt-based words in each group of essays written on the same prompt, we calculated raw frequency and normalized frequency (based on every 100,000 words scale) for the common words that both appeared in this group of essays and all of the other essays in the entire corpus. Keyness values for each common word were also calculated based on the frequency data. Generally, the threshold of keyness value is 3.84 (equivalent to $p < 0.05$). By comparing the frequency-related data in the two groups of essays, if a word had a keyness value greater than 3.84, we considered that word to be significantly more likely to occur in the corpus for one prompt than another. We calculated the keyness values based on the formulas from the research in Rayson and Garside (2000).

Among all the words that had keyness values greater than 3.84, we subtracted the 2500 most common words in English based on the frequency data of the CELEX dictionary[1] (Baayen et al., 1995). This was done to avoid mistakenly deleting non-prompt words from the original corpus. We considered the remaining words that had a keyness value greater than 3.84 and were not in the common word list as prompt-specific words. We then removed all prompt-specific words from the training corpora for the development of the generic unigram measure.

[1] https://catalog.ldc.upenn.edu/LDC96L14

Statistical Analyses

Correlations were calculated to examine relations between DWU/DNU scores and human scores and if removing prompt words or using n-grams (POS n-grams) improved the performance of the DWU/DNU measures.

To determine to what extent the DWU/DNU measures predicted human scores in tandem, stepwise linear regression analyses were conducted. Prior to regression analyses, DWU/DNU scores were checked for normality and multicollinearity, and highly correlated ($r > .699$) variables did not enter the models.

Results

Correlational Analyses

Summaries of correlations among human scores and DWU/DNU scores for the independent and integrated training and test sets are presented in Table 1, 2, 3, 4, respectively.

Table 1

Correlations of human scores and multiple DWU/DNU measures for independent training set

Measure	1	2	3	4	5	6	7
Human rater score	-	0.85	0.86	0.90	0.67	0.91	0.78
Unigram score	0.85	-	0.96	0.95	0.78	0.93	0.84
Generic unigram score	0.86	0.96	-	0.92	0.71	0.91	0.80
Bigram score	0.90	0.95	0.92	-	0.73	0.99	0.82
POS bigram score	0.67	0.78	0.71	0.73	-	0.71	0.92
Trigram score	0.91	0.93	0.91	0.99	0.71	-	0.81
POS trigram score	0.78	0.84	0.80	0.82	0.92	0.81	-

Note. 1 = Human rater score, 2 = Unigram score, 3 = Generic unigram score, 4 = Bigram score, 5 = POS bigram score, 6 = Trigram score, 7 = POS trigram score

Table 2

Correlations of human scores and multiple DWU/DNU measures for independent test set

Measure	1	2	3	4	5	6	7
Human rater score	-	0.24	0.17	0.17	0.08	0.11	0.05
Unigram score	0.24	-	0.57	0.74	0.42	0.53	0.33
Generic unigram score	0.17	0.57	-	0.38	0.12	0.24	0.06
Bigram score	0.17	0.74	0.38	-	0.44	0.66	0.37
POS bigram score	0.08	0.42	0.12	0.44	-	0.39	0.84
Trigram score	0.11	0.53	0.24	0.66	0.39	-	0.37
POS trigram score	0.05	0.33	0.06	0.37	0.84	0.37	-

Note. 1 = Human rater score, 2 = Unigram score, 3 = Generic unigram score, 4 = Bigram score, 5 = POS bigram score, 6 = Trigram score, 7 = POS trigram score

In the independent training set, the baseline unigram DWU measure was strongly correlated with human scores ($r = .85$, $p < .001$). The generic unigram, bi-gram and tri-gram

measures performed better than the baseline measure. For the test set, all of the DWU/DNU measures showed lower correlations. The correlation between the baseline unigram measure and human score ($r = .24$, $p < .001$) was 71% lower compared with the training set. Meanwhile, the correlations of generic unigram, bi-gram, tri-gram, POS bi-gram, and POS tri-gram measures descended in order.

Table 3

Correlations of human scores and multiple DWU/DNU measures for integrated training set

Measure	1	2	3	4	5	6	7
Human rater score	-	0.67	0.70	0.88	0.40	0.90	0.70
Unigram score	0.67	-	0.94	0.81	0.42	0.75	0.60
Generic unigram score	0.70	0.94	-	0.83	0.40	0.77	0.63
Bigram score	0.88	0.81	0.83	-	0.45	0.99	0.78
POS bigram score	0.40	0.42	0.40	0.45	-	0.43	0.59
Trigram score	0.90	0.75	0.77	0.99	0.43	-	0.78
POS trigram score	0.70	0.60	0.63	0.78	0.59	0.78	-

Note. 1 = Human rater score, 2 = Unigram score, 3 = Generic unigram score, 4 = Bigram score, 5 = POS bigram score, 6 = Trigram score, 7 = POS trigram score

Table 4

Correlations of human scores and multiple DWU/DNU measures for integrated test set

Measure	1	2	3	4	5	6	7
Human rater score	-	0.13	0.01	0.16	0.06	0.17	0.02
Unigram score	0.13	-	0.78	0.42	-0.05	0.16	0.12
Generic unigram score	0.01	0.78	-	0.41	-0.10	0.09	0.07
Bigram score	0.16	0.42	0.41	-	-0.02	0.39	0.12
POS bigram score	0.06	-0.05	-0.10	-0.02	-	0.04	0.35
Trigram score	0.17	0.16	0.09	0.39	0.04	-	0.03
POS trigram score	0.02	0.12	0.07	0.12	0.35	0.03	-

Note. 1 = Human rater score, 2 = Unigram score, 3 = Generic unigram score, 4 = Bigram score, 5 = POS bigram score, 6 = Trigram score, 7 = POS trigram score

For the integrated training set, the correlations between the human and DWU/DNU scores were lower than that of the independent training set. However, the trend was similar that the n-gram measures outperformed the unigram measures. The correlations for the test sets were weaker than those reported in the integrated training set. However, using n-grams instead of single words increased the correlations between the DNU and the human scores, which was not in line with what was found in the independent test set. Meanwhile, all correlations show that using POS tags as a unit of measurement was not as effective as the use of word n-grams.

Regression Analyses

For the independent test set, the regression model explained 5.5% of the variance (R^2 =.055, $F(1,478) = 27.88$, $p < .001$) of human scores. Unigram score was the only significant predictor of human score ($\beta = .57$, $p < .001$).

For the integrated test set, a significant regression model explained 2.8% of the variance (R^2 =.028, $F(1,478) = 13.84$, $p < .001$) of human scores. The tri-gram score was the only significant predictor ($\beta = .16$, $p < .001$).

Discussion

Overall, we did not find strong evidence that removing prompt-specific words would improve the performances of DWU measure. For both independent and integrated TOEFL test sets, the baseline unigram measures outperformed the generic unigram measures. This most likely reflects the notion that retaining prompt specific words provides content measures that are generally applicable across topics. Meanwhile, the use of POS n-grams did not seem to improve the performance of the DWU/DNU measures. However, using word n-grams features appears to increase the correlations between DWU/DNU and human scores in integrated writing. Specifically, in the integrated test set, the tri-gram measure reported the best performance, exceeding the bi-gram and unigram measures. This may indicate that high-quality integrated essays resemble each other by using more n-grams from the source texts regardless of specific topics. The regression models based on the DWU/DNU measures explained a relatively small amount of variance of the human scores in the test sets, though more variance was found in independent writing (5.5%) than integrated writing (2.8%).

This research contributes to our understanding of discourse by validating and expanding current DWU measure and by showing to which extent vocabulary use is related to the quality of written discourse in varied tasks. Since the current study has revealed the limitation of solely applying vocabulary features (i.e., the DWU measures) in evaluating the quality of written discourse, our future research will examine its effect in combination with other features (e.g., structural, and syntactic features) and across multiple contexts. By doing that, in the future, we hope to build a more robust understanding of written discourse, and aid in providing more accurate feedback to writers to help with the revision process and writing development in general.

References

Attali, Y. (2007). Construct validity of e-rater® in scoring TOEFL® essays. *ETS Research Report Series*, *2007*(1), i-22. https://doi.org/10.1002/j.2333-8504.2007.tb02063.x

Attali, Y. (2011). A differential word use measure for content analysis in automated essay scoring. *ETS Research Report Series*, *2011*(2), i-19. https://doi.org/10.1002/j.2333-8504.2011.tb02272.x

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650/1492

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning and Assessment*, *10*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1603/1455

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania*.

Rayson, P., & Garside, R. (2000, October). Comparing corpora using frequency profiling. In *The workshop on comparing corpora* (pp. 1-6). https://doi.org/10.3115/1117729.1117730

Kilgarriff, A., 2001. Comparing corpora. *International journal of corpus linguistics*, *6*(1), pp.97-133. https://doi.org/10.1075/ijcl.6.1.05kil