International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 02 | Feb 2020 www.irjet.net p-ISSN: 2395-0072

Modeling Student's Vocabulary Knowledge with Natural Language Processing Tools

Aman Rawani[1], Ashutosh Pratap Singh[2], Navneet Gupta[3]

[1,2]*Student, Dept. of ECE, Arya Institute of Engineering & Technology, Jaipur, Rajasthan, India*
[3]*Associate Professor, Dept. of ECE, Arya Institute of Engineering & Technology, Jaipur, India*

Abstract - *This particular study aims to investigate the degree to which the lexical properties of students' essays can inform stealth assessments of their vocabulary knowledge. In particular, we used scores calculated with the natural language processing tool, TAALES, to predict students' performance on a measure of their vocabulary knowledge. To this end, two corpora were collected which contained essays from sophomores and high school students, respectively. The lexical properties of their essays were then calculated using TAALES. The results of this study indicated that two of the linguistic indices were able to account for 44% of the variance in the sophomores' vocabulary knowledge scores. Additionally, these significant indices from this first corpus analysis were able to account for a significant portion of the variance in the high school students' vocabulary scores. Overall, these results suggest that natural language processing techniques can be used to inform stealth assessments and help to improve student models within computer-based learning environments.*

Key Words: Artificial Intelligence, Intelligent Tutoring Systems, Natural Language Processing(NLP), Tool for the Automatic Analysis of Lexical Sophistication (TAALES), Primary Corpus.

1. INTRODUCTION

Writing maybe a complex cognitive and social process that is important for both academic and professional success. As contemporary societies are growing increasingly reliant on text sources to communicate ideas, the importance of developing proficiency in this area is becoming more important than ever. Unfortunately, acquiring writing skills is not a simple task – as evidenced by the many students who underachieve each year on national and international assessments of writing proficiency. Indeed, this text production process is quite complex and relies on the development of both lower and higher-level knowledge and skills, ranging from a strong knowledge of vocabulary to strategies necessary for tying their ideas together.

To develop those skills that are required to produce high quality texts, students are needed to be provided with comprehensive instruction that targets their individual strengths and weaknesses. In particular, these instructions should explicitly describe and demonstrate the skills and strategies that will be necessary during each of the phases of writing process. Additionally, it should offer student opportunities to receive the summative and formative feedback on his work, while engaging in deliberate practice. This deliberate kind of practice is an important factor in students' development of strong writing skills because it is promoting self-regulation of the planning, generation, and reviewing processes. Unfortunately, however, deliberate practice inherently relies on individualized writing feedback. This is often difficult for teachers to provide, as they are also facing large class sizes and do not have the time to provide thorough analysis on every essay that each student writes.

As a result of these classroom needs, researchers have developed computer-based writing systems which can provide students with feedback on their writing skills. These systems have been used for classroom assignments as well as high-stakes writing assessments to ease the burden of each student's essay scoring. Specifically, automated essay scoring (AES) systems evaluate the linguistic properties of student's essay to assign them holistic score.

These systems use a multitude of natural language processing (NLP) and machine learning techniques to provide the essay scores. To provide student with greater context for the score on his essay, AES systems are most commonly incorporated into educational learning environments, such as automated writing evaluation (AWE) system and intelligent tutoring systems (ITSs). These systems not only provide student with summative feedback on his essay (i.e., his holistic scores), they also provide formative feedback and writing instructions. In order to be successful, these systems must contain algorithms that is able to provide individualized feedback that is relevant to student's skills.

Importantly, these computer-based writing environments rely on linguistic features to assess the quality of the individual essays submitted to the systems. One way to accommodate the individual differences in their scores is to develop user models based on student's characteristics, beyond simply their scores on essays. These models are able to provide more specific instruction and feedback that are tailored to student's strengths and weaknesses. In this paper, we examine the efficacy of NLP techniques to inform stealth assessments of this knowledge. In particular, we examine whether the lexical properties of student's essays can accurately model their scores on a standardized measure of vocabulary knowledge. Ultimately, our aim is to use these measures to provide more individualized tutoring to students.

1.1 Stealth Assessments

In order to provide a more personalized learning experience, computer-based learning environment must rely on repeated assessments of performance of each student. These measures can provide important information about student's knowledge state and learning trajectories, which can help to increase the adaptiveness of these systems. Despite the importance of these assessments, they are not particularly conducive to robust student learning.

Within the context of computer-based learning environment, these stealth assessments can be informed by a wealth of information that can be easily logged in the systems. These data can range from the speed of someone's typing to the trajectories of his mouse movements. Snow and colleagues stated that this measure of behavior patterns could serve as a stealth assessment of agency in adaptive learning environments. Overall, stealth assessments can serve as a viable solution to the assessment problem, as they can be informed by a wide variety of data patterns to model the characteristics of student users.

1.2 Natural Language Processing

Natural language processing (NLP) tools provide a way through which researchers can develop stealth assessments of student's characteristics. In addition, these tools can help researchers to investigate the relationships between individual differences and the learning process at a more sophisticated size. By calculating indices related to multiple levels of the content (e.g., lexical, syntactic, discourse), researchers can look beyond simple measures of holistic quality (i.e., essay scores) and start to examine and model the components of the writing procedure more thoroughly. These models of student's performance can then allow researchers and educators to provide students with more effective instruction that specifically targets their individual's needs.

Broadly, NLP involves around the automated calculation of linguistic text features using a computer program. Thus, the focus of NLP primarily depends on the use of computers to understand, process, and produce natural language text for the purpose of automating certain communicative acts and for studying communicative processes. This technique can serve as a powerful methodological approach for researchers who are interested in examining particular aspects and getting the results of the writing process or for many other domains in which usually students produce natural language.

Researchers have employed NLP techniques within a variety of domains and contexts for mainly the purpose of developing a better understanding the learning process. For example, Varner, Jackson and colleagues (2013) used NLP tools to calculate the extent to which student's self-explanation of complex science texts contained cohesive elements. Results from this study indicated that the better readers produced more cohesive self-explanation than less skilled readers, indicating that the automated indices of cohesion could potentially serve as a proxy for the coherence of student's mental text representations.

In another study, Graesser and colleagues (2011) developed the multiple components of text readability using the NLP tools. These components related to different dimensions of text complexity, such as narration, concreteness and referential cohesion. Through the use of NLP tools, these researchers were able to develop the components that provide multidimensional information about texts and specific properties that influence student's ability to comprehend these texts successfully.

1.3 The Writing Pal

The Writing Pal (W-Pal) is an intelligent tutoring system (ITS) that was designed to provide explicit writing strategy instruction and practice to high school and early college students. Unlike typical AWE systems, W-Pal places a strong emphasis on the instruction of writing strategies, as well as multiple forms of practice (i.e., strategy-specific practice and holistic essay writing practice). The strategy instruction in W-Pal covers all three phases of the writing process: prewriting, drafting, and revising. Within W-Pal, these strategies are taught in individual instructional modules, which include: Freewriting and Planning (prewriting); Introduction Building, Body Building, and Conclusion Building (drafting); and Paraphrasing, Cohesion Building, and Revising. In these videos, the agent describes and provides examples of specific strategies that are important for writing. After viewing these lesson videos, students unlock multiple mini- games, which allow them to practice the strategies in isolation before applying them to complete essays. Within the W-Pal system, students can engage with identification mini-games, where they are asked to select the best answer to a particular question, or generative mini-games, where they produce natural language (typed) responses related to the strategy they are practicing. One of the key features of the W-Pal system is its AWE component (i.e., the essay practice component). This system contains a word processor where students can write essays in response to a number of SAT-style prompts (teachers also have the option of adding in their own prompts to assign to students). Once a student has completed an essay, it is submitted to the W- Pal system. The W-Pal algorithm then calculates a number of linguistic features related to the essay and provides summative and formative feedback to the student (see Figure 2 for a screenshot of the W-Pal feedback screen). The summative feedback in W-Pal is a holistic essay score that ranges from 1 to 6. The formative feedback in W-Pal provides information about strategies that students can employ in order to improve their essays. Once they have read the feedback, students have the option to revise their essays based on the feedback that they were assigned.

2. CURRENT STUDY

The purpose of the current study is to investigate the degree to which the lexical properties of students' essays can inform stealth assessments of their vocabulary knowledge. Ideally, these assessments will serve to inform student models in the Writing Pal system and contribute to its adaptability in the form of more sophisticated scoring algorithms, feedback, and adaptive instruction. To this end, two corpora were collected which contained essays from early college and high school students, respectively. The lexical properties of these essays were then calculated using the Tool for the Automatic Analysis of Lexical Sophistication (TAALES).

TAALES is an automated text analysis tool that provides linguistic indices related to the lexical sophistication of texts. We used this tool in the current study so that we could investigate the relationships between students' vocabulary knowledge and the lexical properties of the essays. We hypothesized that these lexical indices would be significantly related to vocabulary knowledge and that they would provide reliable measures of vocabulary knowledge across two distinct student populations.

2.1 Primary Corpus

The primary corpus for this study is comprised of 108 essays written by college students from a large university campus in Southwest United States. These students were, on average, 19.75 years of age (range: 18-37 years), with the majority of students reporting a grade level of college freshman or sophomores. Of the 108 students, 52.9% were male, 53.7% were Caucasian, 22.2% were Hispanic, 10.2% were Asian, 3.7% were African-American, and 9.3 % reported other ethnicities. All students wrote a timed (25-minute), prompt based, persuasive essay that resembled what they would see on an SAT. Students were not allowed to proceed until the entire 25 minutes had elapsed. These essays contained an average of 410.44 words (SD = 152.50), ranging from a minimum of 84 words to a maximum of 984 words.

2.2 Vocabulary Knowledge Assessment

Students' vocabulary knowledge was assessed using the Gates- MacGinitie (4th ed.) reading comprehension test (form S) level 10/12 [35]. This assessment is a 10-minute task, which is comprised of 45 simple sentences that each contains an underlined vocabulary word. Students were asked to read each sentence and then select the most closely related word (from a list of five choices) to the underlined word within the sentence.

2.3 Text Analyses

To assess the lexical properties of students' essays, we utilized the Tool for the Automatic Analysis of Lexical Sophistication (TAALES). TAALES is an automated text analysis tool that computes 135 indices that correspond to five primary categories of lexical sophistication: word frequency, range, n-gram frequencies, academic language, and psycholinguistic word information [34]. These categories are discussed in greater detail below (see 34 for more thorough information). Word frequency indices are indicative of lexical sophistication, because high frequency words are typically learned earlier in life, are processed more quickly, and are indicative of writing quality (i.e., with high frequency words indicating lower quality writing). There are two primary forms of frequency measures: frequency bands and frequency counts. Frequency bands measure the percentage of a text that occurs in particularly frequency bands (e.g., whether they are in the most frequent 1,000 words, 2,000 words in a frequency list, etc.). Frequency counts employ reference corpora and calculate the frequency of the words in a target text within the reference corpus.

Range indices are indicative of how widely used a particular word or family of words is. Thus, unlike frequency indices, range indices do not simply calculate a raw count of a word in a particular list or corpus. Rather, range indices measure the number of individual documents that contain that word in order to determine the extent that it is used broadly. Range has been used to successfully distinguish the frequent verbs produced by L2 speakers of English from the frequent verbs produced by native English speakers [36].

N-gram frequencies emphasize units of lexical items rather than single words. In particular, n-grams consist of combinations of n number of words (e.g., the bigram "years ago") that frequently occur together. Bigram lists have been shown to be predictive of a speaker or writer's native language, as well as the quality of a given text.

Academic language indices measure the degree to which a text contains words that are found infrequently in natural language corpora, but frequently in academic texts. A number of academic word lists have been calculated to measure the words that are commonly used in academic texts, such as textbooks and journal articles. Thus, these indices provide a measure of how academic a text is compared to more typical texts.

Psycholinguistic word indices provide information about the specific characteristics of the words used in texts. These properties have been shown to be related to lexical decision times, lexical proficiency, and writing quality. TAALES focuses on

five particular properties of words: concreteness (i.e., perceptions of how abstract a word is), familiarity (i.e., judgments of how familiar words are to adults), imageability (i.e., judgments of how easy it is to imagine a word), meaningfulness (i.e., judgments of how related a word is to other words), and age of acquisition (i.e., judgments of the age at which a word is typically learned).

2.4 Statistical Analyses

Statistical analyses were conducted to investigate the role of lexical properties in assessing and modelling students' vocabulary knowledge scores. Pearson correlations were first calculated between students' scores on a vocabulary knowledge measure and the lexical properties of their essays (as assessed by TAALES). The indices that demonstrated a significant correlation with vocabulary knowledge scores ($p < .05$) were retained in the analysis. Multicollinearity of these variables was then assessed among the indices ($r > .90$). When two or more indices demonstrated multicollinearity, the index that correlated most strongly with vocabulary knowledge scores was retained in the analysis. All remaining indices were finally checked to ensure that they were normally distributed.

A stepwise regression analysis was conducted to assess which of the remaining lexical indices were most predictive of vocabulary knowledge. For this regression analysis, a training and test set approach was used (67% for the training set and 33% for the test set) in order to validate the analyses and ensure that the results could be generalized to a new data set. To additionally avoid overfitting the model, we chose a ratio of 15 essays to 1 predictor, which allowed 7 indices to be entered, given that there were 108 essays included in the analysis.

A final linear regression analysis was conducted to determine the extent to which these indices could model the vocabulary knowledge of students in a different population. In particular, we investigated whether the lexical sophistication indices that were retained in the previous regression model (i.e., the regression model for the college students) accounted for a significant amount of the variance in a second set of students' (i.e., the high school students) vocabulary knowledge.

3. CONCLUSIONS

3.1 Vocabulary Knowledge Analysis for the Primary Corpus

Pearson correlations were calculated between the TAALES indices and students' Gates-MacGinitie vocabulary knowledge scores to examine the strength of the relationships among these variables. This correlation analysis revealed that there were 45 linguistic measures that demonstrated a significant relation with vocabulary knowledge scores and did not demonstrate multicollinearity with each other. To avoid overfitting the model, we only selected the 7 indices that were most strongly correlated with vocabulary knowledge. These 7 indices are listed in Table 1 (see Kyle & Crossley for explanations of each variable) [34].

A stepwise regression analysis was calculated with these 7 TAALES indices as the predictors of students' vocabulary knowledge scores for the students in the training set. This regression yielded a significant model, $F(2, 76) = 29.296$, $p < .001$, $r = .660$, $R2 = .435$. Two variables were significant predictors in the regression analysis and combined to account for 44% of the variance in students' vocabulary knowledge scores: mean age of acquisition log score [$\beta = .92$, $t(2, 76) = 6.423$, $p < .001$] and normed count for all academic word lists [$\beta = -.36$, $t(2, 76) = -2.539$, $p = .013$]. The regression model for the training set is presented in Table 2. The test set yielded $r = .600$, $R2 = .360$, accounting for 36% of the variance in vocabulary knowledge scores.

Table -1: Sample Table Correlations between Gates-MacGinitie vocabulary knowledge scores and TAALES linguistic scores

TAALES Variable	r	p
Mean age of acquisition log score	.614	<.001
Mean range (number of documents that a word occurs in) log score	.562	<.001
Spoken bigram proportion	.511	<.001
Mean unigram concreteness score	.492	<.001
Mean frequency score (bigrams)	.488	<.001
Mean frequency score (bigrams)	.476	<.001

Normed count for all academic word lists	.401	<.001

Table -2: TAALES regression analysis predicting Gates- MacGinitie vocabulary knowledge scores

Entry	Variable Added	R^2	ΔR^2
Entry 1	Mean age of acquisition log score	.387	.387
Entry 2	Normed count for all academic word lists	.401	<.001

The results of this regression analysis indicate that the students with higher vocabulary scores produced essays that were more lexically sophisticated. The essays contained words that were acquired at a later age, such as the words vociferous or ubiquitous, which are predicted to be learned later than words such as toy and animal. The essays also contained a greater proportion of academic words that are frequently found in academic texts, such as financier or contextualized, rather than household words such as bread and house. Hence, better writers use words that are found in academic, written language, rather than more common, mundane language. Notably, these two indices, age of acquisition, and academic words, are likely to correlate with indices related to the frequency or familiarity of words in language. However, in this case, they more successfully captured students' vocabulary knowledge from their writing samples compared to simple frequency or familiarity indices.

3.2 Generalization to a New Data Set

Our second analysis specifically tested the ability of the linguistic indices to predict the Gates-MacGinitie vocabulary knowledge scores of students in a completely separate population. To address this question, we collected a test corpus of essays written by high school students and analysed the lexical properties of these essays. Specifically, we calculated the mean age of acquisition log score and the normed count for all academic word lists, as these were the two indices retained in the previous regression model. These indices were then used as predictors in a regression model to predict students' vocabulary knowledge.

3.3 Test Corpus

The test corpus in this paper was collected as part of a larger study (n = 86), which compared the complete Writing Pal system to the AWE component of the system. Here, we focus on the pretest essays produced by these participants. All participants were high-school students recruited from an urban environment located in the southwestern United States. These students were, on average, 16.4 years of age, with a mean reported grade level of 10.5. Of the 45 students, 66.7% were female and 31.1% were male. Students self reported ethnicity breakdown was 62.2% were Hispanic, 13.3% were Asian, 6.7% were Caucasian, 6.7% were African-American, and 11.1% reported other. All students wrote a timed (25-minute), prompt-based, argumentative essay that resembled what they would see on the SAT. Students were not allowed to proceed until the entire 25 minutes had elapsed. These essays contained an average of 340.84 words (SD = 124.31), ranging from a minimum of 77 words to a maximum of 724 words. Finally, these students completed the same vocabulary knowledge assessment as the students in the previous corpus.

3.4 Generalization to a New Data Set

The two TAALES indices (i.e., mean age of acquisition log score and the normed count for all academic word lists) were entered as predictors of students' Gates-MacGinitie vocabulary knowledge scores. This regression yielded a significant model, $F(2, 83) = 8.521$, $p < .001$, $r = .413$, $R2 = .170$. Only one of the variables was a significant predictor in the regression analysis: mean age of acquisition log score [$\beta = .54$, $t(2, 83) = 3.666$, $p < .001$]. This model suggests that the regression model generated with the primary corpus partially generalized to a new data set. One of the indices accounted for a significant amount of the variance in students' vocabulary knowledge scores. However, this variance was smaller than the variance accounted for in the primary corpus.

REFERENCES

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

[2] Baer, J. D., and McGrath, D. 2007. The reading literacy of U.S. fourth-grade students in an international context: Results from the 2001 and 2006 Progress in International Literacy Study (PIRLS). National Center for Educational Statistics, Institute of Education Sciences, U.S. Department of Education.

[3] National Assessment of Educational Progress. 2009. The Nation's Report Card: Writing 2009. Retrieved Nov. 5, 2012, nces.ed.gov/nationsreportcard/writing.

[4] Allen, L. K., Snow, E. L., Crossley, S. A., Jackson, G. T., and McNamara, D. S. 2014. Reading comprehension components and their relation to the writing process. L'annᅠpsychologique/Topics in Cognitive Psychology, 114, (2014) 663 -691.

[5] Flower, L. and Hayes, J. 1981. Identifying the organization of writing processes. In L. Gregg and E. Steinberg (Eds.), Cognitive processes in writing. Erlbaum & Associates, Hillsdale, NJ, 3-30.

[6] Allen, L. K., Snow, E.L., and McNamara, D. S. 2014. The long and winding road: Investigating the differential writing patterns of high and low skilled writers. In J. Stamper, S. Pardos, M. Mavrikis, and B. M. McLaren (Eds.), Proceedings of the 7th International Conference on Educational Data Mining (London, UK, July 4 -7, 2014). Heidelberg, Berlin, Germany: Springer, 304-307.

[7] Johnstone, K.M., Ashbaugh, H., and Warfield, T.D. 2002. Effects of repeated practice and contextual writing experiences on college students' writing skills. Journal of Educational Psychology (2002), 94, 305–315.

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

[9] Allen, L. K., Jacovina, M. E., and McNamara, D. S. in press. Computer-based writing instruction. In C. A. MacArthur, S. Graham, and J. Fitzgerald (Eds.), Handbook of Writing Research.

[10] Dikli, S. 2006. An overview of automated scoring of essays.