Predicting Off-task Behaviors in an Adaptive Vocabulary Learning System

SungJin Nam
School of Information
University of Michigan
Ann Arbor, MI 48109
sjnam@umich.edu

ABSTRACT
In many studies, engagement has been considered as an important aspect of effective learning. Retaining student engagement is thus an important goal in intelligent tutoring systems (ITS). My current studies with collaborators on Dynamic Support of Contextual Vocabulary Acquisition for Reading (DSCoVAR) include building prediction models for students' off-task behaviors. By extracting linguistically meaningful features and historical context information from interaction log data, these studies illustrate how some types of off-task behavior can be modeled from behavioral logs. The results of this research contribute to existing studies by providing examples of how to extract behavioral measures and predict off-task behaviors within a vocabulary learning system. Identifying off-task behaviors can improve students' learning by providing personalized learning materials: for example, off-task behavior classifiers can be used to achieve more accurate predictions of the student's vocabulary mastery level, which in turn can improve the system's adaptive performance. Toward our goal of developing highly effective personalized vocabulary learning systems, this research would benefit from expert feedback on issues that include: principled approaches for adaptive assessment and feedback in a vocabulary learning system; and alternative methods for defining and generating off-task labels.

Keywords
Engagement, off-task behaviors, prediction model, log data, intelligent tutoring system, adaptive system

1. INTRODUCTION
Engagement has long been considered as an important aspect of learning [17, 16]. Engagement is a comprehensive behavior that reflects an integration of different aspects of a person's cognitive state [11, 6, 7]. A student's engagement level while using the system can vary with time, and it can be influenced by many factors, such as the difficulty of questions, prior experience with similar technology, and individual interests or motivation [14, 1]. Thus, measures related to engagement need to consider the multidimensional construct of engagement and clarify which types of engagement are going to be measured in the study [18].

Other studies based on digital learning environments tend to capture engagement based on behavioral signals. Studies on intelligent tutoring systems (ITS) often used features like response time, number of erroneous attempts, and frequent accessing of hint messages to predict students' engagement [2, 4]. Studies in Massive Online Open Courses (MOOC) included features like the number of lecture videos seen, participation in pop-up quizzes, and social interactions like frequency of article posting or comments in the discussion forum, to predict the student's overall participation level [10, 15]. These studies showed that data traces of observable behavior can be used to predict student engagement, often operationalized as a classroom attitude observed from instructors or a survival rate of enlisted courses in a MOOC.

The purpose of this research topic is to model a particular subset of students' off-task behaviors while they use a vocabulary learning system, based on observations of their interaction from log data. In our study, each student response to an assessment question posed by the system was defined as an off-task behavior if it contained less serious, patterned, or repetitive errors [13, 12]. Key research questions on this topic that I will explore include: (1) identifying important predictive features of off-task behaviors in vocabulary learning systems that can be collected from log data, (2) evaluating different modeling methods that can help to develop more accurate prediction models for off-task behaviors, and (3) suggesting effective adaptive strategies for vocabulary learning systems that will help to sustain student's engagement and thus improve their learning outcomes and experience. The results from our current studies are expected to be used maximize the efficiency and long-term effectiveness of student learning outcomes.

2. CURRENT WORK AND RESULTS
Currently, I am working on developing a contextual word learning (CWL) system called Dynamic Support of Contextual Vocabulary Acquisition for Reading (DSCoVAR)[1]. DSCoVAR is an online vocabulary learning system that teaches K-12 students how to figure out the meaning of a word they don't know (sometimes called the *target word*) by using clues from the target word's surrounding context[8].

The DSCoVAR curriculum consists of three sessions: pre-

[1]http://dscovar.org

test, training, and post-test sessions. Questions in the pre- and post-test sessions include multiple types of questions measuring the student's knowledge on vocabulary before and after the training session. The training session consists of an instructional video and practice questions that teach the student different strategies for figuring out the meaning of an unknown target word by using clues from nearby words in the surrounding sentence. Students learned, and were tested on, a family of words known as Tier-2 words, which are words that are critical for understanding more advanced texts, but that are relatively rare in everyday use. These target words were expected to be difficult, but at least familiar or known to a small number of students. (In our first experiment, participants reported that they were Familiar with 26% of the Tier 2 target words, followed by 21% Known, and 53% Unknown (N=33) [13].)

2.1 Feature Extraction

In previous studies [13, 12], we analyzed students' responses in the pretest session and developed prediction models for off-task behaviors based on behavioral features extracted from log data. During sessions, DSCoVAR recorded how students interacted with the system by storing time-stamped event data and students' text responses. Based on the collected log data, we extracted two types of variables: response-time variables (RTV) and context-based variables (CTV). These variables contain more meaningful student behavior information than the raw log data, and are used as predictor variables in our off-task behavior classifiers.

RTVs collect information right after the student submits his or her response for each question, including time spent to initiate and finish typing a response, the number of spelling and response formatting errors, and orthographic and semantic similarity between the response and the target word. CTVs include history-based measures relating to how the student performed in previous trials (with different window sizes of 1, 3, 5, and 7), such as the average proportion of off-task responses in previous trials and average orthographic or semantic overlap between the current response and previous responses. Lastly, human raters created labels for off-task behaviors from log data. By using criteria based on Baker et al. [3], we obtain labels for certain types of off-task behavior, i.e. when responses seemed less serious and patterned, or when they involved repetitive errors.

2.2 Modeling Off-task Behaviors

With the RTVs and CTV features described above, we build off-task prediction models via mixed effect models and structure learning algorithms. Mixed effect models, such as the generalized linear mixed effect model (GLMM) or hierarchical Bayesian model, are suitable for analyzing the log data from ITS since they can account for variance across repeated measures like multiple responses from a single student or a particular target word.

Table 1 and 2 show the results of the GLMM model learned by the stepwise algorithm for predicting the off-task labels from RTV and CTV variables. GLMM includes random intercepts for target words and students, and the effect of random slopes for the student's prior familiarity level to the target word mentioned above 2 [13, 12]. The results show that RTV features like response length and orthographic similarity between the response and the target word are statistically significant for explaining the specific types of off-task behavior that we identified for the study. CTVs like average semantic similarity between the current response and previous three responses and orthographic similarities with previous seven responses were also significant. This model showed a better area under the curve statistic from ROC curve (0.970) than the RTV-only GLMM model (0.918).

Table 1: GLMM results for fixed effect variables (all predictors are statistically significant ($p < 0.001$))

Variables	Coeff	SE	z
(Intercept)	0.50	0.62	0.82
RTV: Response Length	-0.22	0.05	-4.10
RTV: Ort. Similarity	-5.98	1.79	-3.34
CTV: Sem. Similarity (prev. 3)	0.11	0.03	4.35
CTV: Ort. Similarity (prev. 7)	11.4	1.81	6.33

Table 2: GLMM results for random effect variables

Variables	Var.	Corr.
Target (Intercept)	1.05	
Target-Unknown:Known	2.47	-1.00
Target-Unknown:Familiar	23.0	-1.00
Subject (Intercept)	3.67	

Structure learning algorithms, such as the stepwise regression and the Hill-climbing algorithm, were used for automatically learning the model structure of off-task prediction models. The stepwise algorithm was useful in selecting which variables can bring the better fit to the regression model based on criteria like AIC or BIC. The Hill-climbing algorithm was helpful for identifying the complex interaction structures between variables based on conditional probabilities. By combining findings from different structure learning algorithms, we confirmed that adding interaction structures is helpful for prediction, especially with RTV-only models. An example of interaction structures learned from the Hill-climbing algorithm is shown in Figure 1.

3. PROPOSED CONTRIBUTIONS

First, the current work contributes to existing ITS studies by suggesting methods for extracting meaningful information from log data. For example, RTVs provided meaningful information to understand student performance on specific questions by using various language processing techniques, such as orthographic similarities measured using character trigrams, and semantic similarities measured using Markov Estimation of Semantic Association [9]. CTVs provided information on historical patterns of off-task behaviors. Combined with mixed effect models, our results suggest that traditional predictive features, such as time spent for initiating and finishing the response or number of error messages, can be substituted (when available) with features based on variance in repeated measures and contextual information.

Second, identifying off-task status at the item level can be a starting point for managing student engagement systematically, by letting the learning system know when to intervene in helping the student regain their engagement to the task. Off-task classifiers in the current studies provided examples of automatized models for checking student engagement in a vocabulary learning system.

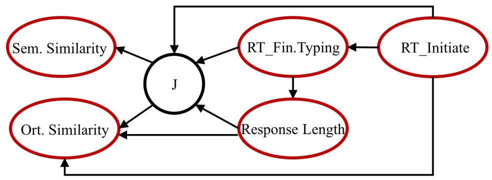

Figure 1: Interaction structure of RTVs learned by the Hill-climbing algorithm (Node J: Off-task label)

Third, this research can be helpful for achieving more accurate predictions on the student's vocabulary mastery level. For example, suggested classifiers provide item-level prediction for off-task behaviors based on previous responses. These results can be helpful for distinguishing between intentionally missed questions and accidentally erroneous responses, which in turn can be used to improve estimates provided by existing student learning prediction models, such as item response theory [5].

4. FUTURE DIRECTIONS

A key goal of this research is to build an adaptive vocabulary learning system. By using results from our current studies, we will implement an initial adaptive mechanism in DSCoVAR that personalizes the difficulty of training session's questions based on a student's estimated vocabulary mastery. This approach is expected to help retain student engagement with the system by providing the right level of 'desirable difficulty' while also making more efficient use of the student's learning time. However, it is unclear how features related to perceived question difficulty, such as amount of information given from feedback messages or size of spacing between questions that share the same target, could be used to model the overall student engagement with the question. Advice from experienced researchers on adaptively controlling task difficulty would help guide this research on personalized training to students.

Our current work depends on defining a specific type of off-task behavior, with labels generated from two human judges. While the inter-rater agreement was reasonable (Cohen's Kappa of 0.695) [12], it is an expensive process and the number of collectible judgments are limited. An alternative approach could be to use crowd-sourcing for labeling the log data. However, converting this expert labeling task into a fragmentary job for anonymous workers may require more carefully designed instructions and robust methods for validating the credibility of labels. Expert guidance on alternate definitions of off-task behavior, and improved approaches for gathering larger amounts of labeled data based on these definitions, would be helpful for expanding future studies.

5. ACKNOWLEDGMENTS

This research was supported in part by the Institute for Educational Sciences (IES R305A140647) through a grant to the University of Michigan. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. I also thank Dr. Kevyn Collins-Thompson and Dr. Gwen Frishkoff for their guidance and suggestions.

6. REFERENCES

[1] I. Arapakis, M. Lalmas, B. B. Cambazoglu, M.-C. Marcos, and J. M. Jose. User engagement in online news: Under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology*, 65(10):1988–2005, 2014.

[2] I. Arroyo and B. P. Woolf. Inferring learning and attitudes from a bayesian network of log file data. In *AIED*, pages 33–40, 2005.

[3] R. S. Baker, A. T. Corbett, and K. R. Koedinger. Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems*, pages 531–540. Springer, 2004.

[4] J. E. Beck. Engagement tracing: using response times to model student disengagement. *Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology*, 125:88, 2005.

[5] S. E. Embretson and S. P. Reise. *Item response theory for psychologists*. Psychology Press, 2000.

[6] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris. School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1):59–109, 2004.

[7] J. A. Fredricks and W. McColskey. The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In *Handbook of research on student engagement*, pages 763–782. Springer, 2012.

[8] G. Frishkoff, K. Collins-Thompson, and S. Nam. Dynamic support of contextual vocabulary acquisition for reading: An intelligent tutoring system for contextual word learning. In *Adaptive Educational Technologies for Literacy Instruction*. Taylor & Francis, Routledge:NY, In Press.

[9] G. A. Frishkoff, K. Collins-Thompson, C. A. Perfetti, and J. Callan. Measuring incremental changes in word knowledge: Experimental validation and implications for learning and assessment. *Behavior Research Methods*, 40(4):907–925, 2008.

[10] R. F. Kizilcec, C. Piech, and E. Schneider. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 170–179. ACM, 2013.

[11] K. R. Koedinger, E. Brunskill, R. S. Baker, E. A. McLaughlin, and J. Stamper. New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, 34(3):27–41, 2013.

[12] S. Nam, K. Collins-Thompson, and G. Frishkoff. Modeling real-time performance on a meaning-generation task. In *Annual Meeting of the American Educational Research Association*. AERA, 2016.

[13] S. Nam, K. Collins-Thompson, G. Frishkoff, and L. Hodges. Measuring real-time student engagement in contextual word learning. In *The 22nd Annual Meeting of the Society for the Scientific Study of Reading*. SSSR, 2015. https://goo.gl/CvTL1K.

[14] H. L. O'Brien and E. G. Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955, 2008.

[15] A. Ramesh, D. Goldwasser, B. Huang, H. Daume III, and L. Getoor. Learning latent engagement patterns of students in online courses. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[16] B. Ravindran, B. A. Greene, and T. K. Debacker. Predicting preservice teachers' cognitive engagement with goals and epistemological beliefs. *The Journal of Educational Research*, 98(4):222–233, 2005.

[17] J. P. Rowe, L. R. Shores, B. W. Mott, and J. C. Lester. Integrating learning and engagement in narrative-centered learning environments. In *Intelligent Tutoring Systems*, pages 166–177. Springer, 2010.

[18] G. M. Sinatra, B. C. Heddy, and D. Lombardi. The challenges of defining and measuring student engagement in science. *Educational Psychologist*, 50(1):1–13, 2015.