Does Size Matter? Investigating User Input at a Larger Bandwidth

Laura K. Varner, G. Tanner Jackson, Erica L. Snow, and Danielle S. McNamara

Department of Psychology, Learning Sciences Institute, Arizona State University, Tempe, AZ 85287
{Laura.Varner, Tanner.Jackson, Erica.L.Snow, dsmcnama}@asu.edu

Publish Date: May, 2013

Abstract

This study expands upon an existing model of students' reading comprehension ability within an intelligent tutoring system. The current system evaluates students' natural language input using a local student model. We examine the potential to expand this model by assessing the linguistic features of self-explanations aggregated across entire passages. We assessed the relationship between 126 students' reading comprehension ability and the cohesion of their aggregated self-explanations with three linguistic features. Results indicated that the three cohesion indices accounted for variance in reading ability over and above the features used in the current algorithm. These results demonstrate that broadening the window of NLP analyses can strengthen student models within ITSs.

Introduction

Developers of intelligent tutoring systems (ITSs) for ill-defined domains increasingly incorporate natural language processing (NLP) as a method of interacting with students via dialogue (VanLehn et al., 2007) and modeling student knowledge (McNamara et al., 2007; Jackson et al., 2003). Many systems use NLP techniques to model student abilities, adapt pedagogy, and personalize instruction. Additionally, natural language evaluations have allowed systems to track development at a fine-grained level and serve as an added source of data for student outcomes.

An example of one ITS that has incorporated NLP techniques is iSTART (McNamara, Levinstein, & Boonthum, 2004). iSTART provides students with training to use self-explanation strategies to better understand challenging text. In this system, students type self-explanations of target sentences and receive feedback based on natural language assessments. The resulting self-explanation scores provided by the algorithm provide useful models of student ability and allow the system to provide adaptive formative feedback across a wide variety of texts (Jackson, Guess, & McNamara, 2010).

The algorithm used in iSTART primarily considers the relationship between the self-explanation and the target text, but does not use information about the linguistic features of the self-explanation, such as its lexical properties, syntactic complexity, or cohesion. Previous efforts have investigated the linguistic features of students' individual self-explanations and how they might enhance the current student model within iSTART. Many of these studies have been internal to the iSTART lab, and because they were unsuccessful, were not published, with one exception (Jackson & McNamara, 2012). These studies have found that linguistic measures at the sentence level provide no increase in accuracy to the iSTART algorithm. In fact, Jackson and McNamara (2012) demonstrated that the addition of linguistic features resulted in an overall reduction in the accuracy of the student model.

Notably, all of these studies have investigated student contributions at the level of individual self-explanations. In the current study, we broaden the scope of the assessment by examining aggregated self-explanations (across entire passages). We hypothesized that a global assessment of student performance at a larger, more context-sensitive bandwidth would provide a more accurate running model of students' self-explanation and comprehension abilities.

iSTART

The iSTART tutor was developed to train students on reading comprehension strategies (McNamara, Levinstein, & Boonthum, 2004). The system focuses on the concept of self-explanation, which benefits students on numerous complex tasks (Chi et al., 1989; McNamara, 2004). iSTART training is divided into three modules: introduction, demonstration, and practice. The modules tutor students through the use of didactic instruction, demonstrations of self-explanations, and practice applying the strategies to texts. iSTART practice contains two sections: initial practice (housed within the two-hour training portion of iSTART) and extended practice, where students can self-explain texts over a period of weeks or months. In the extended practice module, teachers can input and assign new texts for students to practice. Therefore, the assessment algorithm must be flexible enough to evaluate self-explanations produced for any text.

iSTART Evaluation Algorithm

To provide feedback, iSTART uses a localized assessment algorithm that evaluates the quality of individual self-explanations (McNamara et al., 2007). The current assessment method focuses on individual self-explanations using a local window. The algorithm evaluates a student's self-explanation quality with several word-based measures as well as latent semantic analysis (LSA; Landauer et al., 2007). The word-based measures provide lower-level information about the self-explanation, such as length and overlap of content words. This level of assessment primarily identifies self-explanations that are too short, too close to the target sentence, or off-topic. LSA is combined with the word measures to assess the quality of the self-explanation and the degree to which it includes information from the surrounding text and outside relevant information.

Based on this combination of the word-based and LSA-based measures, iSTART assigns self-explanation scores on a scale of 0-3. A score of "0" is assigned to self-explanations that are either too short to assess or that are comprised of irrelevant information. A score of "1" is associated with a self-explanation that mainly relates only to the target sentence (sentence-based). A score of "2" implies that the self-explanation integrates aspects of the text beyond the target sentence (text-based). Finally, a score of "3" suggests that the self-explanation incorporates outside information at a global level. In previous tests, the performance of the iSTART algorithm was comparable to humans (Jackson, Guess, & McNamara, 2010).

iSTART-ME

Although the initial and extended practice modules in iSTART have been shown to increase students' self-explanation and comprehension abilities, these modules tended to become repetitive and lead some students to disengage from training. iSTART-ME (Motivationally Enhanced; Jackson, Dempsey, & McNamara, 2010) was created to to increase students' engagement during extended practice. This system contains the three training modules from iSTART however, the extended practice module includes a selection menu that allows users to interact with game-based features. iSTART-ME contains three forms of generative practice (two of the three are game-based), all which utilize the same assessment algorithm from iSTART.

Current Study

The iSTART algorithm builds student models at a local level (i.e., each self-explanation is evaluated separately for a specified target sentence). This approach provides little information about how students process information across entire texts. Therefore, the current work investigates the potential for modeling students' abilities at a more global level. We examined aggregated self-explanations for each text self-explained by a student. These aggregated self-explanations typically consisted of eight or more individual self-explanations (one for each target sentence in the text) and represented the students' contributions at a larger grain size than do individual self-explanations. Coh-Metrix was used to calculate three indices of cohesion for each of the aggregated self-explanations (McNamara & Graesser, 2012). These measures were averaged across texts to provide each student with global cohesion scores. We investigated the relation between students' cohesion scores and their reading abilities. We hypothesized that these global level cohesion features would be positively related to reading ability and provide added predictive power over and above the current localized NLP algorithm.

Procedure

High-school students (n=126) completed an 11-sessoin experiment with a pretest, 8 training sessions, a posttest, and a delayed retention test. Of these students, 65 interacted with the original iSTART system and 61 interacted with iSTART-ME. As these students completed the same tasks and were assessed via the same algorithm, the current analyses are collapsed across both conditions.

Measures of Student Performance

At pretest, students were given the Gates-MacGinitie Reading Test as a measure of their general reading comprehension ability. During training, self-explanation quality was assessed using the iSTART algorithm. The scores for each individual self-explanation across all training texts were averaged to provide a *training score*.

Coh-Metrix Analysis

The aggregated self-explanations were analyzed using Coh-Metrix, a computational tool that provides lexical, syntactic, cohesion, and other linguistic measures for given texts (McNamara & Graesser, 2012). We focused on three cohesion indices that are theoretically related to text comprehension at the word (lexical diversity), sentence (minimal edit distance), and global (latent semantic analysis cosines between paragraphs) levels. These indices were chosen based on models of text comprehension that emphasize the importance of the levels of knowledge used to process and comprehend texts (Kintsch, 1998).

Students' self-explanations were combined for each text read and self-explained during training. For a target text with p paragraphs and n target sentences, the resulting aggregated self-explanation file would contain p paragraphs and n self-explanations corresponding to the relative position of the target sentence.

Lexical Diversity. The lexical diversity of a given text measures the range of words used in a text (McCarthy & Jarvis, 2007). High lexical diversity is generally assumed to reflect an individual's linguistic skills or competence as a speaker or writer. Within Coh-Metrix, lexical diversity is assessed using multiple formulas, such as *M*, *D,* and *MTLD*. Prior research has found strong relations between the above formulas and overall text quality. For the purposes of this analysis, the *D* measure was used.

Syntactic Cohesion. Coh-Metrix evaluates syntactic cohesion using Minimal Edit Distance (MED). MED measures differences in the sentence positioning of content words, parts of speech (POS), or lemmas (i.e., the base forms of words). MED values are the inverse direction to measures of referential overlap because high MED values indicate the degree to which word locations vary across sentences in a text (McCarthy, Guess, & McNamara, 2009). MED correlates with measures of both syntactic complexity and referential cohesion.

Global Semantic Cohesion. Latent Semantic Analysis (LSA) is a statistical and mathematical representation of word and text meaning (Landauer et al., 2007). In this study, we assessed students' aggregated self-explanation files using the LSA paragraph-to-paragraph measure (LSAppa). This value indicates the semantic similarity of the concepts included in the self-explanations across the paragraphs of the passage. Hence, it provides a global measure of semantic cohesion.

Averaging of Aggregated Self-Explanation Values. The three cohesion variables were calculated for each of the aggregated self-explanation files. For each student, this Coh-Metrix output was averaged across texts to create an average score on each of the three linguistic measures. Each student has an average score for lexical diversity, MED (all words), and LSA paragraph-to-paragraph. These averaged scores provide a measure of students' aggregated self-explanations at three distinct linguistic levels.

Results

We examined the relation between the average linguistic scores on students' aggregated self-explanations, their prior reading comprehension ability, and training scores. Pearson correlations were conducted and followed with linear and hierarchical multiple regression analyses.

Correlation Analyses

Participants' reading comprehension and training scores were significantly correlated with the linguistic features of students' aggregated self-explanations (see Table 1).

Table 1: Correlations between Cohesion Indices and Students' Performance at Pretest and Training

Cohesion Indices	Reading Ability	Training Scores
Lexical Diversity	.452**	.349**
MED (all words)	.419**	.389**
LSA Paragraph	.456**	.538**

** *p* < .01

The linguistic features of the aggregated self-explanations were significantly related to students' reading ability. Students with higher reading scores exhibited diverse words and sentence structures and maintained consistent themes across multiple self-explanations for a given text, as evidenced by the LSAppa cohesion measure. The results also indicate that the average cohesion scores were related to the training scores.

Regression Analyses

Stepwise Regression. A stepwise regression analysis was conducted to examine the percent of variance in reading scores accounted for by the three indices, and to confirm that each of the indices accounted for unique variance. The three cohesion indices were regressed onto reading comprehension scores yielding a significant model, F(3,125) = 19.26, p < .001; R^2 = .32. Unique variance in the reading comprehension scores was predicted by LSAppa [β =.27, t(3,122)=3.09, p = .002, $R^{2\ change}$ = .208], lexical diversity [β =.27, t(3,122)=3.18, p = .002, $R^{2\ change}$ = .088], and MED [β =.19, t(3,122)=2.12, p = .036, $R^{2\ change}$ = .025]. Thus, students with higher reading scores exhibited increased semantic similarity and structural and lexical diversity in their aggregated self-explanations.

Hierarchical Multiple Regression. Our second question regarded the ability of the cohesion indices to predict students' reading ability over and above the scores provided by the iSTART algorithm. This analysis broadens the scope of the analysis conducted by Jackson and McNamara (2012) and examines whether the *linguistic* features of aggregated self-explanations were predictive of students' reading ability over the current model (i.e., the students' training scores). We conducted a hierarchical multiple regression analysis to predict students' reading comprehension ability. Training scores represent the current method of student evaluation and were input as the first predictor block in the model, with the second block comprising the three linguistic measures (see Table 2).

Table 2: Hierarchical Multiple Regression Analysis for Linguistic Variables Predicting Students' Reading Ability

Variable	*B*	SE *B*	β	ΔR^2
Model 1				.33**
Training Scores	.26	.03	.57**	
Model 2				.10**
Training Scores	.18	.04	.39**	
Lexical Diversity	.00	.00	.22**	
MED (all words)	.11	.07	.13	
LSA Paragraph	.20	.17	.10	

** *p* < .01

Model 1 provides confirmation that the current iSTART algorithm significantly contributes to a model of students' reading comprehension ability, accounting for 33 percent of the variance. Model 2 indicates that three linguistic measures account for unique variance in students' reading comprehension ability over the current algorithm. Thus, by broadening the window of our NLP analyses, we are able to use linguistic measures of cohesion to improve the iSTART model of students' comprehension abilities.

Conclusions

NLP has proven to be a valuable addition to ITSs, as it provides students the opportunity to participate in tutorial interactions that adapt to their unique responses. Accordingly, ITSs can leverage natural language input to develop sophisticated models of students' knowledge and abilities. Although the assessment algorithm housed within iSTART is accurate and reliable, it currently evaluates student input solely at a local level. Here, we examined whether evaluations of student performance at a larger bandwidth would provide an improved student model.

Results indicate that the linguistic features of aggregated self-explanations were positively related to training and reading comprehension scores. The training scores accounted for a significant proportion of the variance in students' reading scores. This result reconfirms the efficacy of the current NLP algorithm to model students' comprehension abilities. However, the results showed that an additional 10 percent of the variance in reading scores was accounted for by three linguistic features of the aggregated self-explanations. These findings indicate that there are benefits to be gained from analyses of natural language input at a more global, context-sensitive level.

Though many studies have investigated the efficacy of integrating NLP into ITSs, relatively less focus has been placed varying the bandwidth of these NLP analyses. Prior research on the iSTART algorithm found no additional benefits of linguistic measures on assessments of self-explanations. Our results provide evidence for the contrary, supporting the idea that linguistic measures can enhance the accuracy of NLP algorithms. This method of contextualized analysis may apply to other ITSs across a number of domains. These analyses could offer complementary information to support student models that typically focus on local natural language contributions.

The current study, along with results from future research, can offer significant benefits to student models in ITSs. In the future, we intend to examine the advantage of calculating a running model of student ability using iterative calculations of linguistic scores for aggregated self-explanations. These scores can be used to update student models and inform responses on ensuing texts. More broadly, for any NLP-based ITS, higher-bandwidth NLP methods can help to expand the window of student models to more accurately represent students' abilities.

Acknowledgments

This research was supported in part by the Institute for Educational Sciences (IES R305G020018-02; R305G040046, R305A080589) and National Science Foundation (NSF REC0241144; IIS-0735682). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES or NSF.

References

Chi, M. T. H.; Bassok, M.; Lewis, M.; Reimann, P.; and Glaser, R. 1989. Self-explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cognitive Science* 13: 145-182.

Jackson, G. T.; Dempsey, K. B.; and McNamara, D. S. 2010. The Evolution of an Automated Reading Strategy Tutor: From Classroom to a Game-enhanced Automated System. In *Science of Learning: Cognition, Computers and Collaboration in Education,* 283-306). New York, NY: Springer.

Jackson, G. T.; Guess, R. H.; and McNamara, D. S. 2010. Assessing Cognitively Complex Strategy Use in an Untrained Domain. *Topics in Cognitive Science* 2: 127-137.

Jackson, G. T.; Mathews, E. C.; Lin, D.; and Graesser, A. C. 2003. Modeling Student Performance to Enhance the Pedagogy of AutoTutor. In *Lecture Notes in Artificial Intelligence: 2702,* 368-372. New York: Springer.

Jackson, G. T.; and McNamara, D. S. 2012. Applying NLP Metrics to Students' Self-explanations. In *Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution,* 261-275. Hershey, PA: IGI Global.

Kintsch, W. 1998. *Comprehension: A Paradigm for Cognition*. New York: Cambridge University Press.

Landauer, T., McNamara, D. S., Dennis, S., and Kintsch, W. eds. 2007. *Handbook of Latent Semantic Analysis.* Mahwah, NJ: Erlbaum.

Magliano, J. P.; Todaro, S.; Millis, K.; Wiemer-Hastings, K.; Kim, H. J.; and McNamara, D. S. 2005. Changes in Reading Strategies as a Function of Reading Training: A Comparison of Live and Computerized Training. *Journal of Educational Computing Research* 32: 185-208.

McCarthy, P. M.; Guess, R. H.; and McNamara, D. S. 2009. The Components of Paraphrase Evaluations. *Behavioral Research Methods* 41: 682-690.

McCarthy, P.; and Jarvis, S. 2007. VOCD: A theoretical and empirical evaluation. *Language Testing* 24: 459-488.

McNamara, D. S.; Boonthum, C.; Levinstein, I. B.; and Millis, K. 2007. Evaluating Self-explanations in iSTART: Comparing Word-based and LSA Algorithms. In *Handbook of Latent Semantic Analysis,* 227-241. Mahwah, NJ: Erlbaum.

McNamara, D. S.; and Graesser, A. 2012. Coh-Metrix: An Automated Tool for Theoretical and Applied Natural Language Processing. In *Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution,* 188-205. Hershey, PA: IGI Global.

McNamara, D. S.; Levinstein, I. B.; and Boonthum, C. 2004. iSTART: Interactive Strategy Trainer for Active Reading and Thinking. *Behavioral Research Methods, Instruments and Computers* 36: 222-233.

VanLehn, K.; Graesser, A. C.; Jackson, G. T.; Jordan, P., Olney, A.; and Rose, C. P. 2007. When are Tutorial Dialogues More Effective than Reading? *Cognitive Science* 30: 3-62.