Assessing Cognitively Complex Strategy Use in an Untrained Domain

G. Tanner Jackson (gtjacksn@memphis.edu)
Institute for Intelligent Systems, University of Memphis
Memphis, TN 38152 USA

Rebekah H. Guess (rguess@memphis.edu)
Department of English, University of Memphis
Memphis, TN 38152 USA

Danielle S. McNamara (d.mcnamara@mail.psyc.memphis.edu)
Department of Psychology, University of Memphis
Memphis, TN 38152 USA

Abstract

Researchers of advanced technologies are constantly seeking new ways of measuring and adapting to user performance. Appropriately adapting system feedback requires accurate assessments of user performance. Unfortunately, many assessment algorithms must be trained on and use pre-prepared data sets or corpora in order to provide a sufficiently accurate portrayal of user knowledge and behavior. However, if the targeted content of the tutoring system changes depending on the situation, the assessment algorithms must be sufficiently independent to apply to untrained content. Such is the case for iSTART, an intelligent tutoring system that assesses the cognitive complexity of strategy use while a reader self-explains a text. iSTART is designed so that teachers and researchers may add their own (new) texts into the system. The current paper explores student self-explanations from newly added texts (which iSTART hadn't been trained on) and focuses on evaluating the iSTART assessment algorithm by comparing it to human ratings of the students' self-explanations.

Keywords: Automatic assessment; reading strategies.

Introduction

Modeling student performance has become a somewhat ubiquitous aspect of computer systems, particularly intellignet tutoring systems. Student models range from simple (age, gender) to very complex (series of actions triggering dynamic representations of knowledge structures). The simpler models can be easily transferred across systems and into new domains; however they lack the detail and sophistication to personalize system behavior. More complex models can provide flexibility and robustness within the system; however they usually require specific task attributes or training that may limit their adaptability to new domains.

Situated in this struggle is any number of systems that model cognitive processes, drive interactive android behaviors, or simulate human tutors. This study focuses on an evaluation of an algorithm designed to assess the cognitive complexity of student generated self-explanations within an intelligent tutoring system, iSTART.

iSTART

Interactive Strategy Training for Active Reading and Thinking (iSTART) is a web-based tutoring system designed to improve students reading comprehension by teaching self-explanation strategies. The iSTART system was originally modeled after a human-based intervention called Self-Explanation Reading Training, or SERT (McNamara, 2004; McNamara & Scott, 1999; O'Reilly, Best, & McNamara, 2004). The automated iSTART system has consistently produced gains equivalent to the human-based SERT program (Magliano et al., 2004; O'Reilly, Sinclair, & McNamara, 2004; O'Reilly, Best, & McNamara, 2004). Unlike SERT, iSTART is web-based, and can potentially provide training to any school or individual with internet access. Furthermore, because it is automated, it can work with students on an individual level and provide self-paced instruction. iSTART also maintains a record of student performance and can use this information to adapt its feedback and instruction for each student. Lastly, the iSTART system combines pedagogical agents and automated linguistic analysis to engage the student in an interactive dialog and create an active learning environment (e.g., Bransford, Brown, & Cocking, 2000; Graesser, Hu, & Person, 2001; Graesser, Hu, & McNamara, in press; Louwerse, Graesser, & Olney, 2002).

iSTART Modules

iSTART incorporates pedagogical agents that engage users with the system and tutor them on how to correctly apply various reading strategies. The agents were designed to introduce students to the concept of self-explanation and to demonstrate specific strategies that could potentially enhance their reading comprehension. The iSTART program consists of three system modules that implement the pedagogical principle of modeling-scaffolding-fading: introduction, demonstration, and practice.

The introduction module uses a classroom-like discussion format between three animated agents (a teacher and two student agents) to present the relevant reading strategies within iSTART. These agents interact with each other,

providing students with information, posing questions to each other, and giving example explanations to illustrate appropriate strategy use (including counterexamples). These interactions exemplify the active processing that students should use when providing their own self-explanations. After each strategy is introduced and the agents have concluded their interaction, the students are asked to complete a set of multiple-choice questions that gauge their understanding of the recently covered concepts.

After all strategies are introduced, students progress to the demonstration module. In the demonstration module, new animated characters interact (Merlin & Genie) and guide the students as they attempt to analyze example explanations provided by the Genie agent. In this capacity, Genie acts as another example student, reads text aloud, and provides a self-explanation for each sentence. Meanwhile, Merlin instructs the learner to identify the strategies used within each of Genie's explanations. Merlin provides feedback to Genie on his explanations and to the students on the accuracy of their strategy identifications. For example, Merlin will tell Genie that his explanation is too short and ask him to add information, or he will applaud when the student makes a correct identification. The feedback provided to Genie is similar to the feedback that Merlin will give to the students when they finish that section and move on to the practice module.

Once the students are in the practice module, Merlin serves as their self-explanation coach. He provides feedback on their explanations and prompts them to generate new explanations using their newly acquired repertoire of strategies. The main focus of this module is to provide students with an opportunity to apply the reading strategies to new texts and to integrate their knowledge from different sources in order to understand a challenging text. Their explanation may include knowledge from prior text, or come from world and domain knowledge. Merlin provides feedback for each explanation generated by the student. For example, he may prompt them to expand the explanation, ask the students to incorporate more information, or suggest that they link the explanation back to other parts of the text. Merlin sometimes takes the practice one step further and has students identify which strategies they used and where they were used. Throughout this interaction, Merlin's responses are adapted to the quality of each student's explanation. For example, longer and more relevant explanations are given more enthusiastic expressions, while short and irrelevant explanations prompt Merlin to provide more scaffolding and support.

There are two types of practice modules. The first practice module is situated within the context of the initial 2 hour training. That is, initially, the student goes through the introduction, demonstration, and practice in about 2 hours. The initial practice includes practice with two texts, on which the iSTART algorithms were trained (McNamara, Boonthum, Levinstein, & Millis, 2007).

The second phase of practice, extended practice, begins subsequently. Extended practice can be used in situations where the classroom or a student has committed to using the system over time, such as over the course of a year. During this practice phase, the student is assigned to read texts that are usually chosen by the teacher. These are texts that may be entered into the system with little notice. Because of the need to provide texts *on the fly*, the iSTART feedback algorithms must provide appropriate feedback, not only for the texts during initial practice (for which the iSTART algorithms are highly tuned), but also for new texts. For this reason, the iSTART evaluation algorithms must be highly flexible and must be able to generalize to virtually any text.

iSTART Evaluation Algorithm

Determining the appropriate feedback for each explanation is dependent on the evaluation algorithm implemented within iSTART. Obviously the feedback has the potential to be more appropriate when the evaluation algorithm more accurately depicts explanation quality and related characteristics. In order to accomplish this task and interact with students in a meaningful way, the system must be able to adequately interpret natural language text explanations.

Several versions of the iSTART evaluation algorithm have been tested and validated with human performance (McNamara et al., 2007). The resulting algorithm utilizes a combination of both word-based approaches and latent semantic analysis (LSA; Landauer et al., 2007). The word-based approaches provide a more accurate picture of the lower level explanations (ones that are irrelevant, or simply repeat the target sentence). They are able to provide a finer distinction between these groups than LSA. In contrast, LSA provides a more informative measure for the higher level and more complex explanations. Therefore, a combination of these approaches is used to calculate the final system evaluation.

The word based approach originally required a significant amount of hand-coded data, but now uses automatic methods when new texts are added. The original measure required experts to create a list of "important" words for each text and then also list of associated words for each "important" word. This methodology was replaced, and now the word-based component relies on a list of "content" words (nouns, verbs, adjectives, adverbs) that are automatically pulled from the text via Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004; McCarthy, Guess, McNamara, in press; McNamara, Louwerse, & Graesser, 2002). The word-based assessment also includes a length criterion where the student's explanation must exceed a certain number of words (calculated by multiplying the number of words in the target sentence by a prespecified coefficient for each assessment category).

The LSA based approach uses a set of benchmarks to compare student explanations to various text features. These LSA benchmarks include 1) the title of the passage, 2) the words in the target sentence, and 3) the words in the previous two sentences. The third benchmark originally involved only words from causally related sentences, but this required more hand-coding, and thus was replaced by

the words from recent sentences. Within the science genre, this replacement was expected to do well, because of the linear argumentation most often employed in science textbooks. However, it is unclear how well these assessment metrics will apply to new texts or domains.

The uncertainty of generalization raises an additional constraint for the current system. Namely, the strategies within iSTART, though taught via science texts, actually apply to a wide range of text genres and the algorithm should therefore be able to accommodate a comparable range of explanations. This evaluation system was designed to be automated so that the text repository could be expanded without consequence to system performance (i.e., so that the evaluation algorithm should perform equally well across newly encountered texts). This need motivated the current study.

Experiment

Participants included 549 students recruited from science classes in Mid-Southern High Schools. Throughout the course of a semester, students spent time each week interacting with iSTART. Their teachers added and assigned new science texts for the students to use within iSTART. The long-term interaction between students and iSTART provided a large repertoire of student explanations to analyze (around 40,000). The current analyses represent a subset of these that were coded by human raters, consisting of 5,400 student generated self-explanations.

Procedure

All students interacted with the same version of iSTART (as described above). After training was complete (introduction module, demonstration module, and practice module), students used an iSTART extended practice module to self-explain texts assigned by their teachers. This extended practice module functioned just like the practice module within iSTART (including feedback). The only difference between the practice module and the extended practice module was the text being used. The practice module uses the same set of texts for every student, whereas the texts in the extended practice module were added by the teachers themselves and were not included in any training or validation by the iSTART algorithm.

Assessments

As students interacted with the texts in extended practice their self-explanations were assessed by iSTART, feedback was provided, and the explanations were logged for future use.

iSTART Algorithm The iSTART assessment algorithm evaluated every student self-explanation. This assessment was coded as a 0, 1, 2, or 3. An assessment of "0" relates to explanations that are either too short or contain mostly irrelevant information. An iSTART score of "1" is associated with an explanation that primarily relates only to the target sentence itself (sentence-based). A "2" means that the student's explanation incorporated some aspect of the text beyond the target sentence (text-based). If an explanation earns a "3" from the iSTART evaluation then the explanation incorporates information from a global level, and may include outside information or refer to an overall theme across the whole text (global-based).

Table 1: Examples of Self-Explanation Categories

iSTART Category	Example
Target Sentence	"Energy-storing molecules are produced on the inner folds.
Irrelevant	"Hello, I am a taco."
Sentence-based	"the molecules holding on to the energy are created on the inner folds."
Text-based	"These sentences say that the mitochondria's inner membrane produces energy storing molecules."
Global-based	"The inner folds develop energy-storing molecules that help store more energy for the plant and help it grow, survive, and reproduce."

Human Ratings After all the student self-explanations were collected from extended practice (approximately 5400 explanations), three human experts were asked to provide independent ratings. These human experts have little or no knowledge in how LSA works and they were extensively trained on self-explanation strategies, and the ratings were provided independently of the iSTART algorithm scores (i.e., raters never saw the output from iSTART). The human raters provided scores for each self-explanation on seven categories: garbage, vague/irrelevant, repetition, paraphrase, local bridging, elaboration, and global bridging. An explanation contained garbage if any words or entries were not coherent. Vague and irrelevant explanations consisted of information out of context that does not pertain to and would not contribute to comprehension of the text. The student explanations that simply repeated the words from the target sentence were classified as repetition. When students used their own words to restate the ideas from the target sentence an explanation was categorized as a paraphrase. Local bridging occurred when students self-explained by using information from previous sentences within the text. An elaboration required that students expand on the information from the text and incorporate some of their personal knowledge. Lastly, the raters considered global bridging to be present when a student explanation attempted to draw together a main idea or theme from the

text (this could possibly be a combination of both local bridging and elaboration together).

Ratings were provided on all seven categories for each student self-explanation. The ratings ranged from 1 (definitely not present) to 6 (definitely present). Raters were told to consider the difference between each rating level to be equal, such that the distance between 1 and 2 was equal to the distance between 2 and 3, and so forth. Interrater reliability on a training set of data (including all 3 raters) resulted in an average correlation of .70 across all seven categories. The ratings for some of these categories were combined to form composite scores that represent scores analogous to those provided by the iSTART algorithm. In essence, the human ratings were combined to provide a similar score of 0, 1, 2, or 3 that represented nonsense/irrelevant explanations, sentence-based explanations, text-based explanations, and global-based explanations, respectively. The original seven categories provide a fine grained account of the student explanations, but these composite scores provide a more continuous measure of the cognitive processing that is most likely to contribute to each self-explanation.

The composite ratings provide a general indicator for the amount of cognitive complexity involved in generating each self-explanation. The nonsense/irrelevant explanations obviously lack appropriate effort and/or focus and require no processing of the text. The nonsense/irrelevant explanations receive a score of "0" because they represent the least amount of cognitive effort (which is none at all). Generating a sentence-based explanation requires only minimal processing of the target text, and does not demonstrate any inference making or knowledge activation. These explanations are a step up from the previous category and involve minimal processing, so they receive a score of "1". Creating text-based explanations requires integration between the target sentence and prior sentences, at least to the extent of connecting two ideas explicitly present in different parts of the text. Text-based explanations receive a score of "2" because drawing a connection between the current sentence and a sentence/idea previously covered is more complex than addressing a single sentence alone. Going beyond the text-based local connections, a global-based explanation requires the activation of outside knowledge and/or making generalizations across multiple points within the text. The global-based explanations include the activation and integration of knowledge, as well as using it appropriately (i.e., not including irrelevant information), and therefore represent the highest category of explanation here (a score of "3"). Together these scores provide a direct analog between the human scores and the iSTART scores.

Results and Discussion

Analyses were conducted to assess the relation between iSTART's evaluation algorithm and the assessments made by expert human raters. Note that the current analyses all involve iSTART assessments on student explanations for new texts on which the iSTART algorithm was never trained.

Table 2: Correlations between iSTART and human ratings.

Human Composite Ratings	iSTART algorithm
Nonsense/Irrelevant Factor	-.362**
Sentence-Based Factor	-.127**
Text-Based Factor	.637**
Global-Based Factor	.593**
Human Algorithm Analog	.661**

** indicates $p<.001$

The correlations presented in Table 2 demonstrate a reliable relation between the human ratings and those provided by the iSTART algorithm. The first four correlations illustrate the relation between the iSTART algorithm and each of the individual explanation levels. We would expect a significant negative correlation between the nonsense factor and the iSTART algorithm. This negative relation indicates that iSTART successfully attributed lower scores to explanations when they contained a higher amount of nonsense and irrelevant information. Thus iSTART was able to assess when students were off track or lost focus during their self-explanations. In a similar line of reasoning, we would expect a positive correlation between the global factor and the iSTART evaluation. This positive correlation illustrates that when students tended to include a more global focus in their self-explanation the iSTART algorithm evaluated this information appropriately. The last significant correlation between the human algorithm analog and the iSTART algorithm reveals that the iSTART algorithm is indeed extendable to new texts and is significantly similar to human performance.

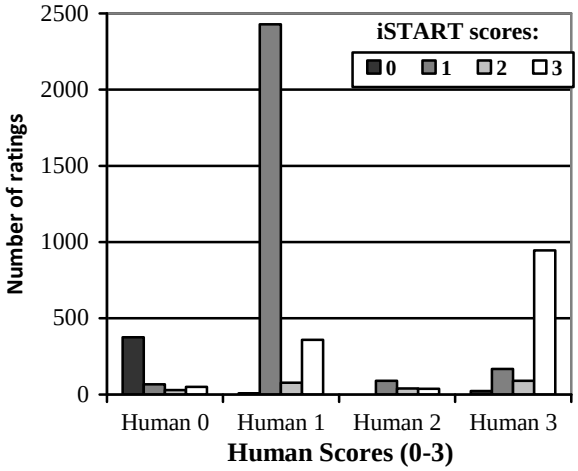

Figure 1: Comparing ratings between iSTART and Humans.

Figure 1 provides a general illustration of how well the iSTART algorithm ratings coincide with human ratings. It is evident that humans and iSTART mostly agree on

explanations that are nonsensical or irrelevant (both rate as a score of 0), sentence-based (both rate as a score of 1), and global-based (both rate as a score of 3). It appears that the text-based explanations are more difficult to determine. However, both humans and iSTART did agree that the incidence of text-based explanations (n_{human}=167, n_{iSTART}=237) was much lower than either sentence-based (n_{human}=2872, n_{iSTART}=2754) or global explanations (n_{human}=1228, n_{iSTART}=1393). The relative difficulty of classifying text-based explanations has been demonstrated in previous iSTART analyses (McNamara et al., 2007).

Further analyses were conducted to examine the specific agreement between the iSTART algorithm and the human algorithm analog. A linear-weighted kappa yielded a significant level of agreement between the human and iSTART scores, kappa = .646, p<.001. This result means that the iSTART evaluation algorithm and human experts significantly agree on their assessments for student explanations. The result is encouraging, considering that these assessments were made on untrained texts and across an entire semester.

Conclusions

The results from this analysis suggest that the iSTART algorithm has the ability to adapt to new texts and information in an appropriate and informative manner. Though similar analyses have been performed with the iSTART algorithm, none of previous evaluations have included extraneous texts. The significant results indicate that iSTART's evaluations are sufficiently accurate for learning purposes (as additional studies have demonstrated consistent learning), and can reliably predict the amount of cognitive processing required to generate self-explanations.

The agreement between iSTART and human raters means that the system can accurately represent the students' explanations and therefore provides the system an opportunity to give accurate and meaningful feedback. This increases the validity of the training, and provides greater assurance of a consistent application of appropriate pedagogy.

The results here are particularly interesting for the cognitive science community because they validate a tool on a completely untrained dataset as it accurately assesses the amount of cognitive processing required to produce natural language explanations. As mentioned earlier, each self-explanation category (nonsense, sentence-based, text-based, or global-based) was designed to assume an incremental step up in the level of cognitive processing. The garbage and irrelevant explanations would require no amount of processing on the part of the student. These explanations could include typing in the alphabet, inputting a single character, or making disparaging remarks. The sentence-based explanations obviously require the student to at least read the target sentence. This category involves information only related to one specific sentence. So the student essentially paraphrases content already present on the screen, which does not require activation of prior knowledge or any inference generation. A text-based explanation may include some paraphrasing, but it also requires the student to make some connection within the text. Creating a successful text-based explanation means that the students must actively connect current information, with information previously covered. While this process involves a small level of local inference generation, it does not require any integration of existing knowledge from the text or outside information. A global-based explanation, on the other hand, requires integration of knowledge, either from the text itself or from outside information. This global focus requires the student to go beyond the text itself and to actively process the material and how it relates to various sources. The fact that both iSTART and humans can agree on these classifications demonstrates that the algorithm can reliably distinguish between various levels of cognitive complexity required to generate a self-explanation.

Although these data are encouraging with respect to the iSTART algorithm, a couple of aspects may limit its' overall generalizability. The first limiting aspect is the fact that the iSTART system was originally designed to cover science texts, and during this study all newly added texts were also within the science genre. While these new texts were within the same genre, they could cover vastly different topics and share almost no explicit information with the original texts.

The second limiting aspect regards the identification of text-based explanations. As indicated by Figure 1, the text-based explanations are difficult to identify and seem to occur less often (both human experts and iSTART showed a low density of text-based explanations). The difficulty of identifying text-based explanations could be due to the lack of sufficient numbers, but more likely this can be attributed to the nature of the explanations themselves. It seems that students tend to one of two things: 1) do very little processing (nothing more than paraphrasing), or 2) they become engaged and actively process the information. Obtaining a more homogenous distribution of strategies would help investigate the current claim that each category requires an incremental step in cognitive processing.

The data presented here support the effort to improve automatic assessment techniques that relate to human cognitive processing. Though the algorithm structure itself doesn't correlate to any specific processes, it can use natural language to reliably predict the amount of processing required for a student contribution. This computational challenge has taken considerable effort, and now allows for the system to be extended into areas not previously considered. Future research will continue to evolve the current algorithm and will incorporate new features into the system.

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. Thanks also to Dr. Phil McCarthy for his contribution to the paper.

References

Bransford, J., Brown, A., & Cocking, R., Eds. (2000). *How people learn: Brain, mind, experience, and school.* Washington, D.C.: National Academy Press. Online at: http://www.nap.edu/html/howpeople1/

Graesser, A. C., Hu, X., & McNamara, D. S. (2005). Computerized learning environments that incorporate research in discourse psychology, cognitive science, and computational linguistics. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer* (pp. 183-194). Washington, D.C.: American Psychological Association

Graesser, A. C., Hu, X., & Person, N. (2001). Teaching with the help of talking heads. In T. Okamoto, R. Hartley, Kinshuk, J. P. Klus (Eds.), *Proceedings IEEE International Conference on Advanced Learning Technology: Issues, Achievements and Challenges* (460-461).

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, *36*, 193-202.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

Louwerse, M. M., Graesser, A. C., Olney, A., & the Tutoring Research Group. (2002). Good computational manners: Mixed-initiative dialog in conversational agents. In C. Miller (Ed.), *Etiquette for Human-Computer Work, Papers from the 2002 Fall Symposium, Technical Report FS-02-02*, 71-76.

Magliano, J. P., Todaro, S., Millis, K. K., Wiemer-Hastings, K., Kim, H. J., & McNamara, D. S. (2004). *Changes in reading strategies as a function of reading training: A comparison of live and computerized training.* Submitted for publication.

McCarthy, P. M., Guess, R.H, McNamara, D. S. (in press). The components of paraphrase. *To appear in Behavior Research Methods.*

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, 38, 1-30.

McNamara, D.S., Boonthum, C., Levinstein, I.B., & Millis, K. (2007). Evaluating self-explanations in iSTART: comparing word-based and LSA algorithms. In T. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.),*Handbook of Latent Semantic Analysis* (pp. 227-241). Mahwah, NJ: Erlbaum.

McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2002). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension*. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.

McNamara, D. S., & Scott, J. L. (1999). Training reading strategies. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the Twenty-first Annual Meeting of the Cognitive Science Society* (pp. 387-392). Hillsdale, NJ: Erlbaum.

O'Reilly, T., Best, R., & McNamara, D. S. (2004). Self-Explanation reading training: Effects for low-knowledge readers. In K. Forbus, D. Gentner, T. Regier (Eds.), *Proceedings of the Twenty-sixth Annual Meeting of the Cognitive Science Society* (pp. 1053-1058). Mahwah, NJ: Erlbaum.

O'Reilly, T., Sinclair, G. P., & McNamara, D. S. (2004). Reading strategy training: Automated verses live. In K. Forbus, D. Gentner, T. Regier (Eds.), *Proceedings of the Twenty-sixth Annual Meeting of the Cognitive Science Society* (pp. 1059-1064). Mahwah, NJ: Erlbaum.