Using Written and Behavioral Data to Detect Evidence of Continuous Learning

H. Chad Lane[1], Mark Core[1], Dave Gomboc[1],
Mike Birch[1], John Hart[2], and Milton Rosenberg[1]

[1] University of Southern California, Institute for Creative Technologies
{lane, core, gomboc, mbirch, rosenberg}@ict.usc.edu

[2] U.S. Army Research, Development, and Engineering Command
Simulation and Training Technology Center
John.HartIII@us.army.mil

Abstract. We describe a lifelong learner modeling project that focuses on the use of written and behavioral data to detect patterns of learning over time. Related work in essay analysis and machine learning is discussed. Although primarily focused on isolated learning experiences, we argue there is promise for scaling these techniques up to the lifelong learner modeling problem.

Keywords: lifelong learner modeling, essay analysis, machine learning.

1 Introduction

To provide individualized content in computer-based learning environments, it is widely regarded as necessary to maintain an estimate of the learner's state (e.g., knowledge, emotions, interests, etc.). This estimate can be used in a variety of ways, including to organize learning materials, make pedagogical decisions, and visualize learning progress for the learner (i.e., open learner modeling). Although much research on learner modeling has focused on isolated learning episodes or contexts, such as tracking learning over a set of problems, researchers have recently begun exploring approaches to scale these techniques up so that learning may be modeled over longer periods of time, made available for inspection and made applicable across different systems, domains, and learning contexts.

Populating a lifelong learner model requires two broad categories of work on the part of a system: *accretion* and *resolution*. Accretion refers to the gathering of evidence, while resolution refers to the determination of the meaning of that data [9]. It is common for systems to provide an *open learner model* that allows learners to inspect visualizations of their understanding, which may support self-assessment and enhance learning [3]. In order for learners to "drill down" into estimates of their learning, open-learner models should also be *scrutable.* This means that users should be able to ask for explanations of changes in their learner model [9].

In this paper, we review some existing literature in areas related to these issues. Specifically, we focus on the use of automated essay analysis and machine learning to accrete evidence of learning. We also describe a work-in-progress, part of the

Technologies for Accelerated Continuous Learning (TACL) project at the University of Southern California. This system seeks to gather evidence from reflective essays and log files from simulations. The vision is to provide an open-learner modeling framework that learners might treat as an interactive journal/notebook that aids the learner in reflecting upon previous learning experiences and choosing new ones.

1.1 Lifelong learning support in the U.S. Army

If current lifelong learning efforts are to scale-up to this idea of a lifelong learning companion it will require both resources and experience in supporting long term learning. We look to the U.S. Army for guidance not only as our funding agency but also as a large, diverse organization committed to the career-long training of its members. In terms of numbers, the Army currently has about 500,000 Active and 200,000 Reserve Soldiers serving and more than 250,000 civilian employees [6]. Training requirements stretch across a wide array of job related activities. The Army defines lifelong learning as "the individual lifelong choice to actively pursue knowledge, the comprehension of ideas, and the expansion of depth in any area in order to progress beyond a known state of development and competency" [7].

Traditional approaches to implementing a wide ranging curriculum over a large population typically involve classroom instruction and opportunities for practice such as a recitation section. However, Soldiers are facing increasingly complex and dynamic operating environments, meaning that lectures and problem sets may be out of date. Other drawbacks can include not having enough classes and practice sessions to meet demand, requiring the physical presence of learners, requiring that groups train together, and giving the same/similar training to the entire group. Time is a critical resource and being able to provide tailored training to a Soldier at the right point in their career is very important. An example of an initiative that seeks to fill this gap is the Infantry School's Warrior University. This site provides a portal to multiple resources to facilitate lifelong learning and professional relationships. The purpose is to be "the center of gravity for Warrior Learning" and the executive agent to enhance resident training and meet the training needs of the units in the field.

A key challenge for such a portal is that the skills to be learned are complex and dynamic. Virtual practice opportunities are now commonplace, but instructors may not be available to provide support. Intelligent techniques are needed in cases such as these where the time required to build a specific ITS may be too great. In the rest of this paper, we discuss the possibilities of analyzing raw data from learners using simulations and the opportunities enabled by having learners write reflective essays.

1.2 Towards automatic detection of competence, learning, and growth

It is natural to expect that the ability of learners to demonstrate mastery of knowledge/skills in computer-based simulations, and express their understanding in written form will improve as they learn the knowledge/skill. A typical learner would begin with fragile knowledge, if any at all, about a specific domain, and over time, demonstrate patterns of competence in both how they talk about that domain and how they behave in problem solving situations. A general lack of evidence of learning from these two sources, especially when scores on standardized tests may indicate otherwise, is a sign that whatever instructional or experiential opportunities the learner is receiving may need to be reconsidered. Discrepancies between written and behavioral data can also be a sign that different sorts of pedagogical interventions would be useful. For example, a learner who succeeds with relative ease in a simulation but consistently receives lower scores on essays may have underdeveloped communication skills. Or, it could be that the learner may not be willing to invest the effort to craft high-quality essays. In the first case, the learner may need access to more resources to develop their writing skills. In the second, metacognitive training (e.g., to motivate the learner to reflect on experiences) may be beneficial.

We are engaged in an effort to build a lifelong learning system that uses written and behavioral data to detect patterns of learning over time, and provide visualizations of this growth for the learner to inspect. The vision is a system that automatically coordinates these two forms of evidence and integrates them into the learner model. For example, in a multi-player game for learning teamwork skills, if a learner writes about the importance of communication between teammates, the system should also seek corroborating evidence from the log files of that exercise, such as evidence that the player did indeed use the microphone or chat window to talk to teammates. In the sections that follow, we discuss text analysis and machine learning techniques that hold promise for the gathering of evidence of learning. We provide a description of our work-in-progress, and then finish with a discussion of future work.

2 Analysis of Essays and Reviews

Asking learners to reflect upon a learning experience via writing essays and reviewing essays of their peers has a number of advantages. One strong advantage is that it enables a lifelong learning system to deal with unstructured experiences such as a museum visit. The system does not need to know what happened at the museum but can instead give general instructions about writing and reviewing reflective essays on visiting a museum. However, for such a system to support a learner it must deal with these unconstrained text documents in some manner.

2.1 Previous Work

This is not the first system to encourage reflection (e.g., [18]) nor the first peer reviewing system (e.g., [5]); below we list lessons learned from previous efforts. The SWoRD project at the University of Pittsburgh [5] has shown that peer reviews can be used as a reliable grading system but recommend 4-6 reviewers and incentives for reviewers to do a good job. Peer critiques have also been used in collaborative student modeling systems, such as peerISM (e.g., [3]).In the short term, we treat reviews simply as an informal way for peers to share information. We will not have the resources to support appropriate quantity and quality of peer reviews.

The standard technique for automated processing of learner essays is Latent Semantic Analysis (LSA). Examples of LSA's use include [18] and [20], and LSA itself is described in detail here [12]. LSA is sometimes called a "bag of words" approach, because it considers the presence of words in the texts to be compared, but not the positions of these words. The technical term for such an approach is vector-based and LSA attempts to automatically derive semantically meaningful vector dimensions based on a training corpus. Alternatively, researchers can specify dimensions manually. [10] describes experiments with "term" vectors where terms are taken from textbook glossaries and manually extracted from a corpus.

LSA can support improving writing quality [20] and other approaches are also applicable (e.g., a parser could be used to search for syntax errors). Our initial goal is to focus on the content of the essay (i.e., are the important concepts mentioned) rather than how well the concepts are presented. There are a few applicable LSA-based approaches to analyzing essay content that we can use. We can create a *gold standard*, a list of the important concepts and representative texts for each, and use similarity comparisons to see which concepts are covered by a learner essay as discussed in [20]. Given a corpus of essays broken into suitable parts such as sentences you can also automatically extract clusters of related sentences without explicitly representing the topics in the corpus (as discussed in [20] and [18]).

2.2 Formative Evaluation

We have conducted a formative evaluation of our essay collection software using the unstructured experience of playing Team Fortress 2 (TF2), a multiplayer video game. TF2, a variation on the classic game of capture the flag, emphasizes teamwork. Like the example of a learner going to a museum, it was difficult to say ahead of time what the learner might take away from the experience.

In this collection of essays we saw some descriptions of teamwork especially with regard to the different character types that players can choose. For example, a medic is unlikely to capture the flag, but can accompany more powerful characters and heal them as needed. We collected representative texts for important domain concepts such as teamwork forming a gold standard that we will use to test a LSA-style approach to processing learner essays. In addition we have a keyword matcher to use as a baseline measure given the work of [19]. They compared the performance of keyword matching to LSA in matching human similarity ratings. Although LSA outperformed keyword matching, they note that keyword matching requires fewer computational resources and thus, in some cases may be preferable.

3 Analysis of behavioral and performance data

Essays provide important insights into the mind of a learner, but they reveal only part of the picture. Some learners may perform masterfully, yet fail to effectively describe their thought processes. Conversely, learners with advanced writing skills may appear to possess a deep understanding, but still lack the ability to perform specific domain tasks. Accordingly, it is also desirable to gather and analyze actual task data. In this section, we discuss possibilities to discern evidence of learning from log data produced by simulations. Although log structure can vary considerably between applications, so long as learner actions and their contexts are reasonably accessible, we can hope to find patterns similar to those expressed by experts and top learners, and perhaps infer that learning is occurring because of these similarities.

3.1 Previous Work

Machine learning techniques have been applied in a variety of ways to detect patterns of expertise and other behaviors of pedagogical interest. For example, statistical methods have been found appropriate for classifying learners according to ability. [16] employed unsupervised neural networks to cluster learners according to ability based solely on their click stream. Subsequently, [17] reports that classifying clicks as productive and non-productive could improve the learner classifications made. [15] describes "a combination of iterative nonlinear machine learning algorithms … to identify latent classes of student problem-solving strategies. The approach is used to predict students' future behaviors" (p. 218). Together, these approaches suggest that it is feasible to identify top performers automatically by analyzing their raw behavior data.

Researchers have also attempted to infer affective states from raw data. For example, [1] presents an analysis of student-tutor interactions and questionnaire responses in an attempt to discern motivation, learning, and help abuse. Inferences from their Bayesian network were reported to be accurate approximately 80% of the time. They report that inspection of conditional probability tables reveals interesting information, e.g., students that report desiring challenge and fearing to be wrong are more likely to have longer times between problem attempts. A combined HMM and IRT model is described in [8] that infers student motivation and gauges student proficiency simultaneously. Although their results were not statistically significant, they were suggestive that given more data the combined model would be more accurate than a model of proficiency alone. Related closely to affect, help abuse is a focus in [2] which describes a generalized detector of "gaming the system" behaviors.

Finally, machine learning has also been used to isolate specific patterns of expert performance and for knowledge acquisition. For example, [11] describes an approach for capturing production rules via programming-by-demonstration. Several studies are reported in [13] about applications of latent problem-solving analysis (LPSA) to dynamic tasks. The author states "simple ideas, such as similarity-based processing and pattern matching, could have a role even in cognitively complex tasks." Together, these approaches suggest basic building blocks for "competence detectors" that could be used alongside text analysis techniques for tracking learning over time.

3.2 Current Work

In our work thus far, we have been developing a server plug-in for the Source game engine used by Team Fortress 2 that logs player activity data deemed interesting based upon our analysis of our formative evaluation. We intend on applying a variety of machine learning algorithms to perceive commonalities and differences amongst learners, and how generated groupings correlate to general levels of competency. The general idea is to group actions and choices represented in the log files (to include specific world state features for context) and label them as coming from expert or novice players. This would provide training data to build a classifier which, in turn, will be used to analyze new game data from live sessions.

4 TACL reflective essay writing prototype

We have developed an initial prototype of the TACL reflective writing environment for the purpose of compiling a corpus of written and behavioral data. The prototype allows a small group of learners to enter reflective essays about their learning and engage in peer reviewing. The system is controlled by a server that manages the peer reviewing process and maintains individual user information. Learners begin by entering essays that describe take-aways from some learning experience. The system sorts the essays and anonymously redistributes them so that learners may read and review each others' essays. Learners write critiques describing the strengths and areas for improvement for the essays. Additionally, they rate the essays according to a 5 point Likert scale based on the quality of information in the essay. Once complete, the reviews are displayed so that learners can revise their essays based on the feedback.

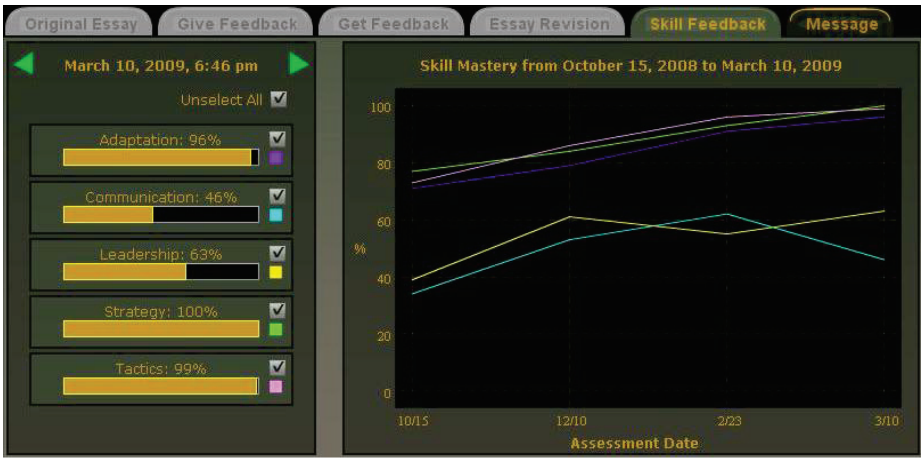

Figure 1. A topic-based learner model with a snapshot (left) and a visualization of growth over time (right). Both use mocked-up data.

The system is modular such that different methods of feedback can be integrated later, such as automated essay feedback instead of (or in addition to) peer feedback. We have also implemented a preliminary open-learner modeling visualization (figure 1) that displays learning progress along domain concepts both at a given date and over several snapshots (these are standard open learner modeling visualizations, e.g., [3,9]). The user in figure 1 has mastered three domain topics, but falls short in two others. We envision this kind of interface being useful for a learner who wants to scan their history, check for improvement in their weaker areas, or be alerted of possible decay in their skills. Also, with respect to analyzing trends, we are working on providing explanations for changes in the graph. For example, the user will be able to hover their mouse pointer over downward trending estimates and receive a message indicating an explanation (e.g., "the last two times you played TF2 you didn't communicate with your teammates very much.").

5 Future work

Besides plans to integrate automated text and log file analysis into the system, we are actively pursuing the adaptation of training system scenarios in concert with our topic-based lifelong learner model. For example, if an essay omits an important concept that is expected (based on top-rated essays from other learners), that model could be handed back to the simulation so that extra attention could be given to the weak concept [14]. This work forces consideration of complex issues related to interoperability and distributed student modeling discussed in [9].

A crucial evaluation that will test the system is in consideration of its context as a lifelong learning companion. The open learner model gives the user a set of evaluation measures to consider in addition to what the learning experience provides and what peers provide. This may promote increased learning over a condition where the learner model is not available for inspection. We can also evaluate the active use of the learner model in guiding the student. One option is that a learning companion in the role of guide selects a relevant educational resource to recommend to the learner along with explanations for why it believes such recommendations are useful (related to the notion of scrutability). This could simply be a relevant essay to read or a complex resource such as an online course or a serious game. Like the results of a search engine, these recommendations can be graded as to their relevancy.

In sum, we seek to synthesize behavioral and textual data to provide estimates of a learner's growth over time. Currently we are integrating these techniques into an open-learner modeling framework, and plan to use it to provide general guidance and support via a lifelong learning companion.

Acknowledgments. The project described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Arroyo, I., Woolf, B.: Inferring learning and attitudes from a Bayesian network of log file data. In: Looie, C.K., McCalla, G., Bredeweg, B., Breuker, J. (eds.) Proc of AIED2005, pp. 33-40. IOS Press, Amsterdam (2005)

2. Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R.: Developing a generalizable detector of when students game the system. User Modeling and User-Adapted Interaction, 18, 3, 287-314 (2008).
3. Bull, S., Brna, P., Critchley, S., Davie, K., & Holzherr, C.: The missing peer, artificial peers and the enhancement of human-human collaborative student modeling. In S.P. Lajoie & M. Vivet (Eds.), Proceedings of the International Conference on Artificial Intelligence in Education, IOS Press, Amsterdam, 269-276 (1999).
4. Bull, S. & Pain, H.: 'Did I say what I think I said, and do you agree with me?': Inspecting and Questioning the Student Model, in J. Greer (ed), Proceedings of World Conference on Artificial Intelligence in Education, Association for the Advancement of Computing in Education (AACE), Charlottesville, VA, 501-508 (1995).
5. Cho, K., Schunn, C. D., & Wilson, R.: Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. Journal of Educational Psychology, 98(4), 891–901 (2006)
6. Department of the Army, United States: About the Army, retrieved March 25, 2009 http://www.goarmy.com/about/personnel.jsp
7. Department of the Army, United States: Army Leadership (FM 6-22), http://usacac.army.mil/CAC/CAL/FM6_22.pdf (2006)
8. Johns, J., Woolf, B: A Dynamic Mixture Model to Detect Student Motivation and Proficiency. In: Proc of AAAI06, pp. 163-168. AAAI Press, Menlo Park (2006)
9. Kay, J.: Lifelong learner modeling for lifelong personalized pervasive learning. IEEE Trans. On Learning Technologies. 1(4), 1-14 (2008)
10. Kim, J., Shaw, E., Ravi, S., Tavano, E., Arromratana, A., Sarda, P.: Scaffolding On-line Discussions with Past Discussions: An Analysis and Pilot Study of PedaBot. In: Proc. Of ITS08, pp. 343-352 (2008)
11. Matsuda, N., Cohen, W.W., Koedinger, K.R.: Building cognitive tutors with programming by demonstration. In: Kramer, S., Pfahringer, B. (eds.) Proc of the Int Conf on Inductive Logic Programming, pp. 41-46.
12. Martin, D. I., Berry, M. W.: Mathematical Foundations Behind Latent Semantic Analysis. In: Landauer, T. K., McNamara, D. S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum, Mahwah, NJ (2007)
13. Quesada, J.: Spaces for Problem Solving. In: Landauer, T. K., McNamara, D. S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum, Mahwah, NJ (2007)
14. Riedl, M.O., Stern, A., Dini, D., Alderman, J.: Dynamic experience management in virtual worlds for entertainment, education, and training. In Int. Tran. on Systems Science and Applications, Sp. Iss. on Agent Based Systems for Human Learning, 4(2), 23-42 (2008)
15. Soller, A., Stevens, R.: Applications of Stochastic Analyses for Collaborative Learning and Cognitive Assessment. In: G. R. Hancock, K. M. Samuelson (eds.), Advances in Latent Variable Mixture Models, pp. 217-253 (2007).
16. Stevens, R., Ikeda, J., Casillas, A., Palacio-Cayetano, J., Clyman, S.: Artificial Neural Network-Based Performance Assessments. Comp. in Human Behavior, 15, 295-313 (1999)
17. Vendlinski, T., Stevens, R.: The Use of Artificial Neural Nets (ANN) to Help Evaluate Student Problem Solving Strategies. In: Fishman, B., O'Connor-Divelbiss, S. (eds.), Proc. of the 4th Int. Conf. of the Learning Sciences, Erlbaum, pp. 108-114. Mahwah, NJ (2000).
18. Villalón, J., Kearney, P., Calvo, R.A., & Reimann, P.: Glosser: Enhanced Feedback for Student Writing Tasks. In Proc of Eighth IEEE International Conference on Advanced Learning Technologies, Santander, Spain, (2008)
19. Wiemer-Hastings, P. Wiemer-Hastings, K., & Graesser, A.: How Latent is Latent Semantic Analysis? In: Proc. of the 16th Int. Jt Congress on Artificial Intelligence, pp. 932-937 (1999).
20. Wiemer-Hastings, P. & Graesser, A.: Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. Int Learning Env., 8(2), 149-169 (2000).