Disengagement Detection Within an Intelligent Tutoring System

Su Chen[1,2], Anne Lippert[1,2], Genghu Shi[1,2],Ying Fang[1, 2] and Arthur C. Graesser[1, 2]

[1] University of Memphis, Memphis TN 38111, USA
[2] Institute for Intelligent Systems, Memphis, TN
schen4@memphis.edu

Abstract. This paper describes a novel automated disengagement tracing system (DTS) that detects mind wandering in students using AutoTutor, an Intelligent Tutoring System (ITS) with conversational agents. DTS is based on an unsupervised learning method and thus does not rely on any self-reports of disengagement. We analyzed the reading time and response accuracy of 52 low literacy adults who interacted with AutoTutor to learn reading comprehension strategies. Our results show that students completing a lesson with 20 questions tend to start mind wandering at the 11th ~15th question. Question chunks with mind-wandering have an accuracy of 20%, in contrast to 70% in accuracy for non-mind wandering.

Keywords: CSAL AutoTutor, Mind Wandering, Disengagement.

1 Introduction

In many respects, intelligent tutoring systems (ITS) live up to their reputation as "next generation" learning environments. Well-designed ITSs are technology driven, automated, and offer a personalized and adaptive instruction that is difficult, if not impossible to implement in a traditional classroom setting. In other respects, ITSs are no more advanced than human instructors when it comes to challenges in student learning. For instance, both ITS designers and human teachers struggle with how best to keep learners focused, interested, and stimulated by material. Regardless of whether they learn from an ITS or in a classroom, students are likely to become disengaged due to various reasons such as fatigue, distraction by environment, loss of interest or falling behind in a course. Though there have been efforts by ITS designers and developers to make systems more generally attractive and interactive to users (Graesser, Cai, Morgan, & Wang, 2017), it is likely that effective interventions will need to be personalized. Studies have only recently been conducted with personalized interventions to prevent or interrupt disengagement activities and guide an individual learner back on track (D'Mello & Graesser, 2012). A critical component of such an intervention is an ITS built-in disengagement tracing algorithm which can capture "mind-wandering"(MW) promptly and accurately.

We define MW to be the disengagement of attention from an assigned task, which is largely involuntary and related to "off-track" behaviors such as boredom and distraction. Besides leading to low performance, MW can present a problem for researchers because it may contaminate the actual reading time (or time spent on one question) and thus confound the true signal/pattern in the data. MW students usually take too long (thinking about something irrelevant to the reading task) or too short (quickly finish the session without comprehension) on one question chunk (i.e. a chunk that a student spends on one question). A disengaged reader is extremely slow or fast with low performance, depending on how readers handle the frustration of underperforming. Data analyzed without addressing the abnormal reading time due to MW may lead to unreliable and misleading results. It is well established that MW is negatively related to reading comprehension (Mills, Graesser, Risko&D'Mello, 2017).

Existing MW detection methods applied supervised learning approaches to train models using self-reported MW (Mills, Graesser, Risko&D'Mello, 2017). The participants are probed during reading with a stimulus signal, upon which they report whether or not they are MW. Self-reported MW is not always available for a concurrent disengagement monitoring system; such self-reports are collected at the end of training sessions and used for post-hoc research. However, these judgments may have a response bias to the extent that disengaged students may feel guilty and prefer not to admit that they have been MW. Beck (2005) proposed an approach using item response theory to detect whether a student is engaged in answering questions. The estimated probability of disengagement depended on the response time and accuracy of the responses. However, Beck's method requires a reasonably large sample size to build a model that accounts for inter-student and -question type variability since a large number of parameters were introduced. Apparently, the required size is difficult to obtain, even for Beck, who was unable to test the approach due to insufficient data.

In this paper, we propose an unsupervised self-learning algorithm to monitor whether a student is engaged in answering questions within AutoTutor lessons. Disengagement is measured in terms of the time that a student spends on a question, as well as his or her relative short-term performance. Disengaged students tend to spend too long or short time on a particular question and thereby perform poorly on the question. The algorithm utilizes the first 3 to 5 well-performed questions to learn a student's pace in a specific lesson and then tracks his/her learning process for questions for which they exhibit disengagement.

2 Description of CSAL Auto Tutor

CSAL AutoTutor is a derivative of AutoTutor developed to help adult learners with low literacy skills improve reading comprehension as part of an intervention led by the Center for the Study of Adult Literacy (CSAL, http://csal.gsu.edu). AutoTutor teaches comprehension strategies by holding conversations called "trialogues" between two computer agents (a tutor and peer) and the human student (Graesser, Li, & Forsyth, 2014; Lehman & Graesser, 2016). The 35 lessons of AutoTutor focus on one or more specific theoretical levels of reading comprehension. The lessons are adaptive in the

sense that they present reading material of varying difficulty depending on the student's performance. Typically, the system will first present students a medium level text and ask 8-12 questions about the text. Depending on students' performance on the questions, they will subsequently get a hard (if above a threshold) or easy (if below a threshold) level text and assessment (Graesser, Feng, and Cai, 2017). Some lessons only provide one medium level text followed by up to 30 questions.

3 Method

3.1 Participants and Design

Participants were 52 adult students from literacy classes in Atlanta and Toronto. They completed a 100-hour intervention over four months. Their ages ranged from 16–69 years (M = 40, SD = 14.97) and 73.1% were female. All participants read at 3.0–7.9 grade levels. On average, the 52 participants completed 23 lessons (ranging from 2 to 29 lessons1), and each lesson contained 14.6 questions (medium level) ranging from 6 to 30 questions. The lessons were scaled on different levels of text and discourse analysis. Specifically, Graesser and McNamara's multilevel theoretical framework of comprehension specifies six theoretical levels: word (W), syntax (Syn), the explicit textbase (TB), the referential situation model (SM), the genre/rhetorical structure (RS), and the pragmatic communication level. AutoTutor taps all of these levels except for syntax and pragmatic communication. The 29 lessons were assigned a primary level (but typically had a secondary or even tertiary level, but these were not considered in this paper). The word level addresses topics such as word meaning clues, learning new words, and multiple meaning words. TB lessons focus on pronouns, punctuation, and main ideas. The SM lessons concern connecting ideas and making inferences from text, whereas RS lessons cover the structure of different genres, such as steps in procedures and problems and solutions. Of the 29 lessons, only 12 provide a single medium level text assessed by 15 to 30 questions. The other 17 lessons start with a medium level text (~15 questions) and then branch to an easy/hard level text according to a student's performance on the first text. The counts of lessons from each theoretical level and branch status are provided in Table 1.

3.2 Disengagement Tracing Algorithm

A disengagement tracing system (DTS) in AutoTutor is expected to automatically learn a student's reading ability and set it as a reference of the participant for disengagement detection. Capturing behavior that as "off-track" will allow us to identify whether a student is "mind-wandering"(MW) on a specific question. The amount of time a student takes to respond to a question, namely "response time" (RT) can be used to determine when a student is off-track. MW students will involuntarily shift

[1] 6 of the 35 AutoTutor lessons were not in the scope of the intervention curriculum so students did not receive these lessons.

Table 1. Distribution of Theoretical Levels Across the 29 lessons (Number of lessons)

Theoretical Level	W	TB	SM	RS
One Text	1	1	6	4
Two Texts (Branch to easy/hard)	3	4	5	5

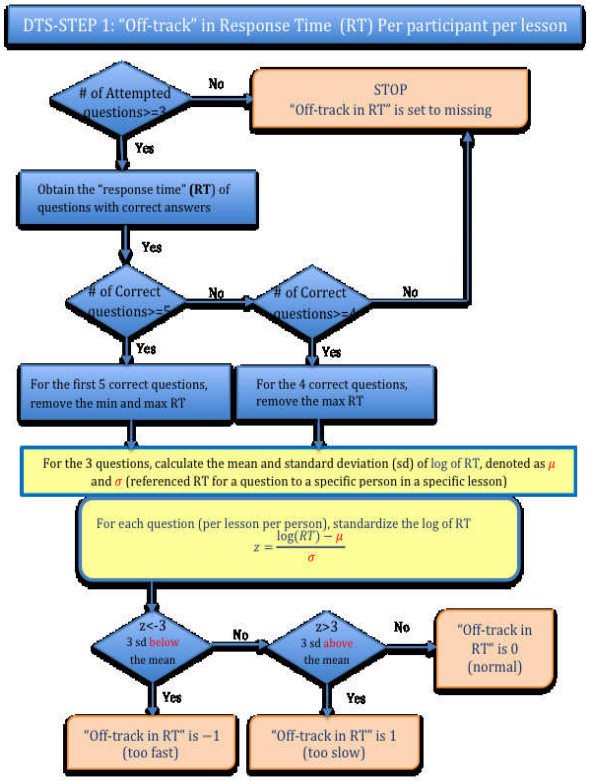

Fig. 1. DTS-Step 1: "off-track" in response time (RT)

attention away from the targeted task towards task-unrelated thoughts and comprehension is likely to suffer. When we assume students' reading abilities are unlikely to improve or worsen in a short time period, one indication that students are off-track is if they try to compensate for a lack of comprehension by answering "too fast" or "too slow" (relative to their personalized "normal" RT) on a question. **The DTS algorithm consists of two steps:** the first step (illustrated in Fig. 1) in detecting disengagement on a question is to identify off-track in RTs. To this end, we make two assumptions: (1) students tend to be engaged at the beginning of a lesson when answering the first few questions and (2) if a student correctly answered a question, he/she most probably was engaged. This means the average time a student spends on the first few correctly answered questions of a lesson reflect RTs while the student is "on-track" or engaged. However, what is on-track for one student may be off-track for another, Furthermore, an individual's reading ability may vary depending on the characteristics of the texts (e.g. difficulty, type) included in each lesson. Because of these sources of variation, it

is necessary to establish multiple baselines or reference behaviors for each learner in each lesson. Therefore, we extract a set of references, or a "reference library" of RTs for each participant for each lesson from the log files of AutoTutor, which contain this information. Specifically, we computed the mean and standard deviation of the log of RT of the first m correctly answered questions ($m = 5$ in this study) and treated it as a "reference RT" for a certain individual at a specific lesson. It is possible that one question is answered correctly by accident. To take this into consideration, we dropped the highest (and lowest) reading time before calculating the benchmark statistics. If the student has less than 3 (correctly answered) questions, the algorithm lacks information to learn for the reference library and will return "missing" until the student answers enough questions correctly. Response time is naturally right-skewed. Our numerical study shows that a log transformation takes the data to a normal distribution. Given the normal distribution, the "3-standard deviation rule" applies. Once the reference library is created, we can say 'a student is off-track on a question (too fast or too slow)' if the log of reading time is below or above 3 standard deviations from the reference engaged data sample.

Disengagement detection only based on response time would lead to a large number of "false positive". Some lessons start with a very easy or "confidence-boosting" question, which means learners will respond more quickly to this question than others with high accuracy. Disengaged students usually perform poorly since they are not focusing on the question. However, a student with an overall accuracy of 80% for a lesson may still answer 3 questions incorrectly in a sequence and take more time than usual to do so. This indicates a high chance that this student is off-track while working on these 3 questions. Some questions in a lesson are very straightforward (or complicated). Students may take significantly less (or longer) time than their reference engaged time. Our target MW questions are those with off-track response times, poor local performance, but possibly adequate overall performance. Overall performance of a lesson per participant is measured by the overall correct proportion for the lesson. Local performance of a question per participant is given by moving average of correctness proportion. The k^{th} order moving average of t^{th} question is given by $\frac{\sum_{i=t-k}^{t+k} X_i}{2k+1}$, where X_i is 1 if the i^{th} question is correctly answered and 0 otherwise. In this study, we take $k = 1$. Step 2 of DTS refines results from Step 1 by filtering out well-performed questions for students who spent too long (or short) time on a questions.

4 Results

We applied the proposed DTS algorithm to the data extracted from AutoTutor (18,863 question-chunks, 52 participants) and identified 900 mind-wandering question-chunks from 51 participants. We were interested in, first, which "questionID"s (questions are answered sequentially) in a lesson are most likely to lead to disengagement? Second, do the patterns of MW differ across the four theoretical levels? We plotted the proportions of MW by each "questionID" for lessons in each of the four theoretical levels

(Fig. 2). The number of question chunks is different for each "questionID". For example, there are more observed question chunks in Question#1 than Question#12 due to the facts that (a) some lessons have less questions than others or (b) some students did not complete all the questions in a lesson. DTS algorithm assumes that the response time of questions within one lesson is from the same distribution. We are mainly interested in differences of MW pattern between theoretical levels although response time may vary between lessons within a theoretical category. In Fig. 2, we also plotted the frequency of question chunks for each "questionID". Fig. 2 suggests different trends in MW for the different theoretical levels. In general, an increasing number of MW is observed as "questionID" goes from 1 ~ up to 30.

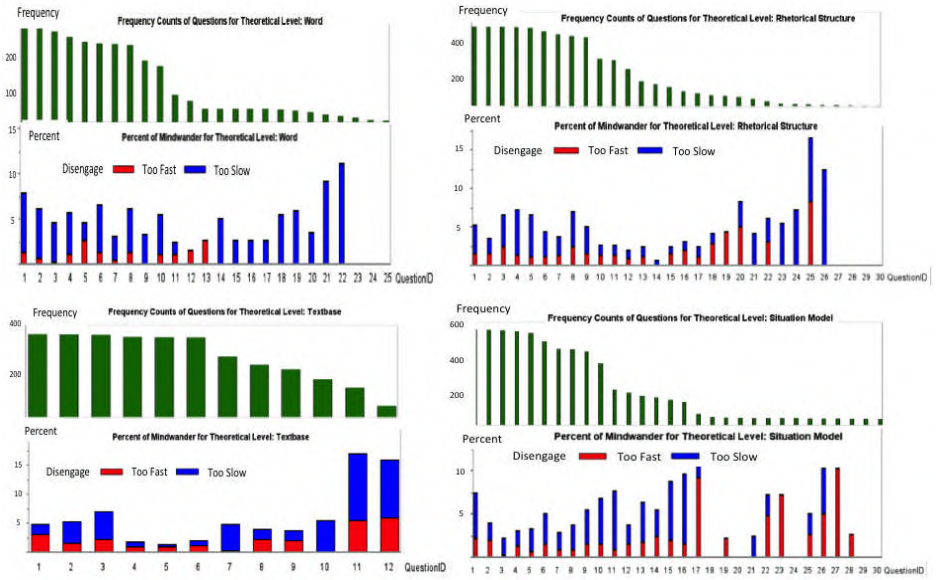

Fig. 2. Disengaged proportion versus question ID at four theoretical levels

Higher proportions of question chunks are identified as "disengaged" in terms of response time and performance for larger "questionID", which coincide with common sense. Students may get tired. Surprisingly, there is a small peak (in MW rate) at the first question of lessons. However, we note the first question is a special case, which may not truly reflect disengagement. For instance, participants may require additional time to adjust to the text/ lesson or they may encounter confusion in using the technology. This appears to be the case for the TB level where the first question has a high rate of participants answering "too fast" followed by a slight drop in the disengagement rate. We also see that some theoretical levels show an increase in disengagement between Question 2 to 4, which may indicate that students were learning skills or getting familiar with the content. For lessons at the SM level, students required a longer learning period (too slow rate increased until Question 6). For all levels it appears that after five to seven questions students gradually gained necessary skills for the lesson, and disen-

gagement decreased until 11th ~15th question, after which it increased again. The question chunks with high MW rate before 11th question should not be considered as "true" disengagement. Students tended to start MW at the 11th ~15th question, after which engagement rates steadily increased. Participants may have felt fatigue or got bored, which could slow their speed in problem solving or induce quick answers without deep thinking. To our surprise, the disengagement rate of the **last** 1~3 questions suddenly dropped to zero (except TB), which contradicts common sense. We checked the frequency counts of questions for each "questionID" and found that the total counts of these last 1~3 questions are very small (nearly zero). Thus, we would not be able to observe disengaged question-chunks with such a small sample size. Furthermore, we found fewer disengaged chunks after question 11 because some of the lessons in our sample contained less than 12 questions. Naturally, any contribution from these lessons to the frequency of MW becomes zero after question 11. Another explanation is that some of the students did not complete all the questions and quit in the middle of the lesson. We can see how this explanation makes sense particularly for lessons containing only one text since they may ask students up to 30 questions. It may be that giving students > 11 questions leads to boredom/fatigue or frustration (if questions are too difficult) and so they voluntarily disengage from the tutoring system. For lessons with two unique texts and ~ 12 questions per text, the story is likely to be different. When a student is presented with a second text, he or she spends extra time constructing a new mental model to make sense of this new information- similar to what occurs at the beginning of a lesson when material is first presented. We are likely to see this additional time show up as increased response times and increased MW for the first few questions pertaining to the second text. After this, the mental model is somewhat stable and response times should level out.

To determine the effectiveness of the DTS proposed in Section 3.2, we compared the accuracy of the responses given while MW versus not MW. Out of the 900 MW question-chunks, 178 (20%) questions were correctly answered. In contrast, 12,657 (70%) of the 17,867 non-MW question-chunks were correctly answered. The accuracy of non-MW question-chunks is 70%, which is significantly higher than the 20% for the MW group ($\chi^2 = 40.6, \mathrm{p} < .001$). To better illustrate the power of the proposed DTS algorithm, we predicted the off-track reading time (Step 1 of DTS) by classical outlier detection method, i.e. 3 IQR (Interquartile range) rule. An extreme outlier is detected when the data is below Q1 (first quartile) $-3 * IQR$ or above Q3 (third quartile)$+3 *$IQR. To fairly compare the proposed DTS algorithm, we filtered out the poorly performed questions identified in Step 2 of DTS from the questions with extreme outliers in reading time. The accuracy of non-MW versus MW questions was 69% and 55% respectively, indicating our DTS algorithm performs better in predicting MW question-chunks.

5 Discussion and Summary

This paper provides an intelligent self-learning algorithm to monitor student engagement during instruction. The algorithm learns a student's baseline reading ability from

his/her first 3~5 well performed questions in a specific lesson and then creates a personalized reference RT. An off-track question chunk is identified if abnormal deviation from the reference is found. The proposed method does not require any self-reported MW evaluation from the participants and can provide disengagement feedback promptly during the lesson. Furthermore, the proposed algorithm is simple and fast, which makes it amenable for use on projects with massive data. The DTS algorithm assumes that questions in a lesson are similar/exchangeable in terms of difficulty and context. Additional adjustments are needed if questions in a lesson are designed to be in different levels. In addition, DTS may report "false disengagement" in the first 10 questions. Users should be cautious in interpreting the early signal of "disengagement" by DTS algorithm.

Disengagement/MW detection and monitoring is critical in improving the efficiency of intelligent tutoring systems. Feedback from the proposed disengagement monitoring system can elucidate factors that lead to distractions. Accordingly, effective interventions can help engage the off-track learner at the right time. For example, once the disengagement is identified, a pop-up window with a kind reminder like "It seems like that you are mind wandering. Do you need a break? Or would you like to read more details about XX?" Or we could have the agents say something shocking when mind wandering is detected. Then users will turn their attention back to the lesson. These types of human-like interactions can be integrated into ITS to grasp the user's attention. The DTS technique "cares" about the student in that it looks for situations when the student is bored or frustrated and can adapt material or prompts to the student.

References

1. Beck, J.E. (2005). Engagement tracing: using response times to model student disengagement. In Proceedings of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology. IOS Press, Amsterdam, The Netherlands, The Netherlands, 88-95.
2. D'Mello, S. K. & Graesser, A. (2012). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems*, 2(4), 23: 1-38.
3. Graesser, A.C., Feng, S., & Cai, Z. (2017). Two technologies to help adults with reading difficulties improve their comprehension. In E. Segers and P. Van den Broek (Eds.), Developmental perspectives in written language and literacy. In honor of Ludo Verhoeven (pp. 295-313). John Benjamin Publishing Company.
4. Graesser, A.C., Li, H. & Forsyth, C. (2014) Learning by Communicating in Natural Language with Conversational Agents. Curr Dir Psychol Sci 23:374–380
5. Graesser, A.C., Cai, Z., Morgan, B., & Wang, L. (2017). Assessment with computer agents that engage in conversational dialogues and trialogues with learners. *Computers in Human Behavior*, 76: 607-616.
6. Lehman, B., & Graesser, A.C. (2016). Arguing your way out of confusion. In F. Paglieri (Ed.), The Psychology of argument: Cognitive approaches to argumentation and persuasion. London: College Publications.
7. Mills, C., Graesser, A., Risko, E. F., & D'Mello, S. K. (2017). Cognitive coupling during reading. *Journal of Experimental Psychology: General*, 146(6), 872.