When Does Disengagement Correlate with Learning in Spoken Dialog Computer Tutoring?

Kate Forbes-Riley and Diane Litman

Learning R&D Ctr, University of Pittsburgh, Pittsburgh, PA 15260
forbesk,litman@cs.pitt.edu

Abstract. We investigate whether an overall student disengagement label and six different labels of disengagement type are predictive of learning in a spoken dialog computer tutoring corpus. Our results show first that although students' percentage of overall disengaged turns negatively correlates with the amount they learn, the individual types of disengagement correlate differently with learning: some negatively correlate with learning, while others don't correlate with learning at all. Second, we show that these relationships change somewhat depending on student prerequisite knowledge level. Third, we show that using multiple disengagement types to predict learning improves predictive power. Overall, our results suggest that although adapting to disengagement should improve learning, maximizing learning requires different system interventions depending on disengagement type.

Keywords: types of disengagement, learning, correlations, spoken dialog computer tutors, manual annotation, natural language processing

1 Introduction

The last decade has seen a significant increase in computer tutoring research aimed at improving student learning (and other performance metrics) by tailoring system responses to changing student affect and attitudes, over and above correctness. Student (dis)engagement behaviors have been of particular interest in this research, including displays of gaming, boredom, indifference, (lack of) interest, (low) motivation, curiosity, and flow (e.g. [7, 3, 9, 11, 12]). Correlational analyses of student (dis)engagement behaviors in tutoring system corpora have indicated that these behaviors are predictive of learning. For example, gaming [3, 1] and boredom [9] have been associated with decreased learning during computer tutoring, while flow [9] and engagement [4] have been associated with increased learning. In addition, a number of automatic gaming detectors have been implemented and evaluated in computer tutors, with results indicating that gaming behaviors can be reliably detected in real-time using features of the tutoring interaction (cf. [3]). Moreover, controlled experiments using gaming-adaptive computer tutors - i.e., tutors enhanced with interventions that target student gaming - have shown that adapting to gaming can improve student learning [2, 3] or other performance metrics (such as reducing gaming) [13, 1].

Our own research builds on this prior work, with the larger goal of enhancing our spoken dialog computer tutor to automatically detect and respond to student disengagement over and above correctness and uncertainty,[1] and thereby improve learning and other metrics. However, in contrast to prior work, which has focused on detecting and adapting to only one or two disengagement behaviors (typically gaming), our goal is to detect and respond differently to a wider range of student disengagement, with the system interventions differing depending on the *type* of disengagement. Our work is also novel in that it focuses on spoken language-based disengagement displays. Working towards our end goal, in prior work we developed and evaluated an annotation scheme for manually labeling an overall measure of disengagement, as well as different types of disengagement, in our spoken dialog computer tutoring corpora (Section 2).

In this paper, we extend the results of others' prior work correlating disengagement behaviors and learning (Section 3). First, we show that although our overall measure of disengagement is predictive of decreased student learning in our spoken dialog computer tutoring corpus, different types of disengagement correlate *differently* with learning: some negatively correlate, while others don't correlate at all. Furthermore, the amount of prerequisite knowledge a student has changes these relationships somewhat. Finally, we show that using multiple disengagement types to predict learning improves predictive power. Importantly, our results suggest that while adapting to an overall measure of disengagement can improve student learning, maximally improving learning requires different system interventions depending on the type of disengagement.

2 Computer Tutoring Disengagement Data

Our research is performed on a corpus of spoken dialogs from a controlled experiment evaluating an uncertainty-adaptive version of our tutoring system, ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialog system), which is a speech-enhanced and otherwise modified version of the Why2-Atlas qualitative physics tutor (cf. [6]). The experimental procedure was as follows: college students with no college-level physics (1) read a short physics text, (2) took a multiple choice pretest, (3) worked with ITSPOKE, (4) took a survey, and (5) took an isomorphic posttest. The resulting corpus contains 360 spoken dialogs (5 per student) from 72 students (6044 student turns). Figure 1 shows a corpus example.

Briefly, ITSPOKE tutors 5 physics problems (one per dialog), using a Tutor Question - Student Answer - Tutor Response format. After each tutor question, the student speech is sent to the Sphinx2 recognizer, which yields an automatic transcript. This answer's (in)correctness is then automatically classified based on this transcript, using the TuTalk semantic analyzer [8], and the answer's (un)certainty is automatically classified by inputting features of the speech signal, the automatic transcript, and the dialog context into a logistic regression

[1] As discussed further elsewhere, our current system already adapts to student uncertainty over and above correctness; our goal is thus to enhance this system to adapt to multiple affective states (disengagement and uncertainty) [7].

model. The appropriate tutor response is determined based on the answer's (in)correctness and (un)certainty and then sent to the Cepstral text-to-speech system, whose audio output is played through the student headphones and is also displayed on a web-based interface. See [6] for details.

Our disengagement annotation scheme is empirically derived from observations in our data but draws on prior work, including appraisal theory-based emotion models, which also distinguish emotional behaviors from their underlying causes (e.g.,[5])[2], as well as prior approaches to manually annotating disengagement or related states in tutoring corpora [9, 11, 12]). Our inter-annotator reliability evaluation on a corpus subset showed that our overall disengagement label (0.55 Kappa) and disengagement type labels (0.43 Kappa) can be annotated with moderate reliability on par with prior emotion annotation work [7]. For the current analysis, all student turns in the corpus were manually annotated as summarized below. See [7] for full details of the annotation scheme:

An **overall Disengagement label (DISE)** was used for all turns expressing moderate to strong disengagement in the tutoring process, i.e., answers given without much effort or without caring about correctness. Answers might also be accompanied by signs of inattention, boredom, or irritation. Clear examples include answers spoken quickly in leaden monotone or with sarcastic or playful tones, or with off-task sounds such as rhythmic tapping or electronics usage.[3]

One of the six **Disengagement Type labels** summarized below accompanied each DISE label. These labels represent the (inferred) underlying causes of disengagement *as well as* the behavior and context evidencing them. In particular, they distinguish different student reactions to the system's limited natural language processing abilities (NLP-Distracted/NLP-Gaming), different student perceptions of the tutoring material (Easy/Hard/Presentation), and a "catch-all" category for other student reactions as the session progresses (Done).

NLP-Distracted: Student became distracted and hyperarticulated[4] this answer because the system misunderstood an immediately prior answer due to its limited natural language processing capabilities.

Hard: Student lost interest because this tutor question was too hard (e.g., presupposes too much prior knowledge).

NLP-Gaming: Student didn't try to work out the answer to this tutor question; s/he instead deliberately gave a vague or incorrect answer or a guess to try and fool the system's limited natural language processing capabilities.

---

[2] Appraisal theories argue that one's appraisal of a situation causes emotion; i.e., emotions result from (and don't occur without) an evaluation of a context (e.g.,[5]).

[3] Affective systems research has found that total disengagement is rare in laboratory settings (e.g., [7, 9]). As in that research, we thus equate the "disengagement" label with either no or low engagement. Since total disengagement is common in real-world unobserved human-computer interactions (e.g., deleting unsatisfactory software), it remains an open question as to how well laboratory-based findings generalize.

[4] That is, gives the answer with unnatural pitch, cadence, stress, or loudness in an attempt to make the computer better understand him/her. This label was renamed from "Language" in our prior work [7] for clarity.

Presentation: Student didn't pay attention to this tutor question because the presentation was too long or complex; his/her answer reflects unawareness of the fact that the tutor turn strongly hinted at the correct answer.

Easy: Student lost interest because this tutor question was too easy (e.g., a similar question was asked and answered earlier in the session).

Done: Student just wants the interaction to be over (typically later in the dialogs) - s/he is bored, tired, and/or not interested in continuing at this moment (or no other label fits).

This scheme should generalize to other learning environments, including analogs of NLP-Gaming and NLP-Distraction, since these two types represent two disengagement behaviors stemming from a system's inherent interaction processing inflexibility, which exists regardless of the communication medium.[5]

$\mathbf{T}_9$: What's the numerical value of the man's acceleration? Please specify the units too.

$\mathbf{S}_9$: The speed of the elevator. Meters per second. **(DISE: NLP-Gaming)**

...

$\mathbf{T}_{15}$: What is the definition of Newton's Second Law?

$\mathbf{S}_{15}$: I have no idea <*sigh*>. **(DISE: Hard)**

...

$\mathbf{T}_{21}$: Based on our discussion, we conclude that the keys will remain in front of the man's face during the entire fall. [...] Would you like to do another problem?

$\mathbf{S}_{21}$: No <*laugh*>. **(DISE: Done)**

Fig. 1. Corpus Example Illustrating Disengagement Annotation Scheme

Figure 1 illustrates the scheme. $\mathbf{S}_9$ is labeled DISE with the NLP-Gaming Type because the student avoided giving a specific numerical value, offering instead a vague (and incorrect) answer. $\mathbf{S}_{15}$ is labeled DISE with the Hard Type because the student gave up immediately and with irritation when too much prior knowledge was required. $\mathbf{S}_{21}$ is labeled DISE with the Done Type because the student answered 'No' semi-jokingly in regards to continuing the experiment.

Note that our NLP-Gaming label represents a subset of the gaming behaviors addressed in prior work (Section 1), which focuses on hint abuse and systematic guessing.[6] ITSPOKE does not provide hints upon request, and the dialog is the only recorded behavior, thus all detectable gaming behavior is linguistic. Altogether our Disengagement types label a range of behaviors associated with disengagement, including off-task, bored, or low-motivated actions that don't attempt to exploit the system. Moreover, our labels capture the fact that these behaviors can be associated with different underlying causes. E.g., a student who disengages because a question is too hard may exhibit any of these behaviors.

Note finally that this turn-level annotation scheme captures both fleeting disengagement states as well as long-term disengagement escalation across turns.

[5] NLP-Distraction differs from the other types in that although students do lose the tutoring flow, this is not of their own (un)conscious volition.

[6] This prior work defines gaming as attempting to succeed by exploiting the system rather than learning the material and using that knowledge to answer correctly [3].

3 Prediction Results

To investigate whether our overall disengagement (DISE) and disengagement type labels are predictive of learning in our corpus, we computed the percentage of each label's occurrence for each student, and the partial Pearson's correlation between the percentage and posttest score, controlling for pretest to account for learning gain. Table 1 shows first the mean percentage (Mn%) and its standard deviation (sd) over all students, the Pearson's Correlation coefficient (R) and significance (p) with significant results bolded ($p \leq 0.05$), and the total number of occurrences (Tot) for each label in the entire dataset. These statistics are then provided for students with low and high pretest scores (see below). The last two rows show test scores for each group (Mn% and sd).

Table 1. Correlation Results between Disengagement or Disengagement Types and Learning in the ITSPOKE Corpus (N=72; Low Pretests N=40; High Pretests N=32)

	All Students			Low Pretests			High Pretests		
Measure	Mn%(sd)	R(p)	Tot	Mn%(sd)	R(p)	Tot	Mn%(sd)	R(p)	Tot
DISE	14.5(8.2)	**-.33(.01)**	886	16.2(8.3)	**-.37(.02)**	555	12.2(7.7)	-.26(.15)	331
NLPDistract	0.4(1.4)	-.03(.78)	28	0.6(1.8)	-.07(.68)	22	0.2(0.8)	.04(.81)	6
Hard	2.8(2.9)	**-.36(.01)**	172	3.6(3.3)	**-.35(.03)**	124	1.7(2.0)	**-.46(.01)**	48
NLPGame	3.0(3.0)	**-.34(.01)**	186	3.2(2.8)	**-.31(.05)**	108	2.9(3.2)	**-.39(.03)**	78
Easy	1.4(2.6)	.12(.33)	83	1.1(2.0)	-.02(.92)	36	1.8(3.2)	.30(.11)	47
Present	3.0(2.2)	**-.27(.02)**	182	3.6(2.1)	-.22(.17)	124	2.1(2.0)	**-.35(.05)**	58
Done	3.9(3.2)	-.08(.52)	235	4.2(3.2)	-.11(.53)	141	3.5(3.3)	-.04(.85)	94
Pretest	51.0(14.5)			40.5(7.8)			64.1(9.2)		
Posttest	73.1(13.8)			66.9(12.8)			80.8(10.9)		

Considering the results over all students, comparison of means shows that of the 14.5% overall disengaged turns on average per student, Done is the most frequent type of disengagement, followed by NLP-Gaming and Presentation, Hard, Easy, and NLP-Distracted. Since Done is defined as a "catch-all" category, it is not surprising that it is the most frequent; that it occurs only slightly more than three of the other types suggests that our six categories are sufficiently representative of the range of disengagement behaviors (and underlying causes) in our data. The high standard deviations suggest that the amount of overall DISE, and the disengagement types, are highly student-dependent.

The correlation results over all students show that overall DISE is significantly correlated with decreased learning. This supports prior work (Section 1) showing negative relationships between learning and boredom or gaming. Our results also show significant negative correlations between learning and the Hard, NLP-Gaming, and Presentation Types. This suggests that the negative DISE correlation is primarily due to these three types. Prior work suggests that gaming behaviors associated with poorer learning often occur when students lack the knowledge to answer the question [3, 2].[7] Similarly, we hypothesize that students often exhibited linguistic (NLP) gaming in our corpus because the system's

[7] Other suggested reasons for gaming in this prior work include a performance-based mentality (as opposed to learning-based) and low motivation to learn.

limited natural language processing abilities prevented them from eliciting information they needed to answer the question. Together, the results for the NLP-Gaming and Hard Types suggest that if not remediated, disengagement can negatively impact learning when it is caused by questions presupposing knowledge the student doesn't have. Relatedly, the negative Presentation correlation suggests that if not remediated, disengagement can also negatively impact learning when it is caused by the inflexibility of the system's half of the dialog.

There are no significant correlations over all students for the NLP-Distracted, Easy, or Done Types. This indicates that student disengagement during tutoring is not always negatively related to learning. In particular, although some students may get distracted and irritated by system misunderstandings, this (NLP-Distracted) is not associated with decreased learning. Of course, the NLP-Distracted Type was very rare in our corpus; more frequent occurrences may impact learning. In addition, although some students may temporarily lose interest when a tutor question is too easy, this (Easy) is not associated with decreased learning. This result supports prior work suggesting that disengagement behaviors in highly knowledgeable students may have little relation to learning, while the same behavior in students with low prerequisite knowledge is associated with poorer learning [3]. Of course, our subjects were all novices; a very high proportion of easy questions is more likely to be associated with poor learning. Interestingly, the lack of a negative Done correlation suggests that temporary losses of student interest that occur as the tutoring dialog or session nears its end (or for other unclear reasons) are also not related to poorer learning.

To further investigate how students' prerequisite knowledge level impacts the relationship between disengagement behavior and learning in our data, we split students into high (N=32) and low (N=40) groups based on their mean pretest score,[8] and then reran the correlations on each group individually.

Comparison of means in Table 1 shows similar relative frequencies of the types across both groups: Done, Presentation and NLP-Gaming occur most often, and NLP-Distracted least often. However, the relative frequencies of Hard and Easy differ depending on knowledge level. Comparing absolute frequencies, one-way ANOVAs showed that only DISE, Hard, and Presentation differed significantly across the two groups ($p<.05$), occurring more for low pretesters.

Regarding the correlations, neither group patterned identically to the combined group. The low pretest group did not show the negative correlation between learning and Presentation, while the high pretest group did not show it for overall DISE. It may be that students with high prerequisite knowledge are most sensitive to the way the system presents the material, i.e., are more likely to disengage and stop learning if they have difficulty immediately understanding the presentation. Although not quite a trend, the positive Easy correlation appears

[8] We didn't use a median split because it placed the same score in both groups. A T-test showed the two groups represent different populations ($p<.001$). Also note that while a repeated test-measure ANOVA has indicated that all students learned during the tutoring ($F(1,69) = 225.688$, $p<0.001$) [6], a one-way ANOVA showed no difference in normalized learning gain between the high and low pretest groups.

to counterbalance the negative correlations in the high pretest group, perhaps explaining the lack of an overall DISE correlation. Interestingly, and in contrast to prior work, our results suggest that NLP-Gaming negatively impacts learning regardless of prerequisite knowledge. This may be because prior work focused on hint abuse and systematic guessing, which are gaming methods targeted at manipulating the system into giving the correct answer. In contrast, students don't know whether NLP-Gaming will result in the correct answer.

Finally, after examining how each disengagement metric predicts learning in isolation, we investigated their relative usefulness in a more complex learning model. We used stepwise linear regression to predict posttest, allowing the model to select its inputs from pretest and our seven disengagement metrics. The following model yielded the best significant training fit to our data (R^2=.49, p<.001). As shown, two disengagement types were incorporated along with pretest. The (standardized) feature weights indicate relative predictive power in accounting for posttest variance. As shown, the Hard Type (p<.01) is more predictive of decreased posttest than the Presentation Type (p=.03), but both work together to significantly increase the model's predictive power over pretest alone.

Posttest = .41***Pretest** - .28***%Hard** - .21***%Presentation**

4 Current Directions

We extended prior research by investigating how overall disengagement (DISE) and its subtypes relate to learning in spoken dialog computer tutoring. We showed that overall DISE negatively correlates with learning, as do the Hard, Presentation, and NLP-Gaming Types, but the NLP-Distracted, Easy and Done Types do not. We showed that prerequisite knowledge level impacts these relationships: only high pretesters exhibit the Presentation correlation, while only low pretesters exhibit the DISE correlation. We also showed that using both the Hard and Presentation Types to model learning improves predictive power.

These results are now impacting our next step: enhancing ITSPOKE to adapt to disengagement. They suggest that maximizing learning requires different adaptations depending on DISE type. Thus we are now using machine learning to automatically recognize DISE types, based on linguistic features (e.g., acoustic-prosodic, lexical and dialog) previously used to predict affect in speech (cf. [6]), and system-specific features (e.g., correctness, timing, knowledge level, and question difficulty) previously used to predict gaming (e.g., [3, 2, 13, 4]).

Our adaptations assume that identifying the causes of affect can help determine how to best respond (cf. [5]). They build on our current results and on prior evaluations of gaming adaptations in computer tutors that involved preventing gaming (e.g., [13, 4, 10, 1]), metacognitive feedback about better ways to learn [2, 13, 1], easier exercises focusing on the gamed material [3], and performance feedback reminding students of task value [2, 13]. Our current results suggest that the Easy, NLP-Distracted and Done Types should receive minimal, non-invasive interventions; they don't impact learning (at least at current levels), thus their adaptation should aim to reduce disengagement without reducing learning (e.g., via metacognitive and performance feedback). Since the Hard, NLP-Gaming,

and Presentation Types negatively correlate with learning and involve a lack of understanding of the tutor question, they require more substantial interventions (e.g., feedback to promote re-engagement and an easier version of the question).

Acknowledgments

This work is funded by National Science Foundation (NSF) award #0914615 and #0631930. We thank Art Ward and our anonymous reviewers for comments.

References

1. Aleven, V., McLaren, B., Roll, I., Koedinger, K.: Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. In: Proc. 7th International Intelligent Tutoring Systems Conference (ITS). pp. 227–239. Maceio, Brazil (2004)
2. Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Merheranian, H., Fisher, D., Barto, A., Mahadevan, S., Woolf, B.: Repairing disengagement with non-invasive interventions. In: Proc. Artificial Intelligence in Education (AIED). pp. 195–202 (2007)
3. Baker, R.S., Corbett, A., Roll, I., Koedinger, K.: Developing a generalizable detector of when students game the system. User Modeling and User-Adapted Interaction (UMUAI) 18(3), 287–314 (2008)
4. Beck, J.: Engagement tracking: using response times to model student disengagement. In: Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED). pp. 88–95. Amsterdam (2005)
5. Conati, C., Maclaren, H.: Empirically building and evaluating a probabilistic model of user affect. User Modeling and User-Adapted Interaction 19(3), 267–303 (2009)
6. Forbes-Riley, K., Litman, D.: Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. Speech Communication (2011), In Press
7. Forbes-Riley, K., Litman, D.: Annotating disengagement for spoken dialogue computer tutoring. In: D'Mello, S., Calvo, R. (eds.) Affect and Learning Technologies. Springer (2011), To Appear
8. Jordan, P., Hall, B., Ringenberg, M., Cui, Y., Rose, C.: Tools for authoring a dialogue agent that participates in learning studies. In: Proc. Artificial Intelligence in Education (2007)
9. Lehman, B., Matthews, M., D'Mello, S., Person, N.: What are you feeling? Investigating student affective states during expert human tutoring sessions. In: Intelligent Tutoring Systems Conference (ITS). pp. 50–59. Montreal, Canada (June 2008)
10. Murray, R.C., vanLehn, K.: Effects of dissuading unnecessary help requests while providing proactive help. In: Proc. of the International Conference on Artificial Intelligence in Education. pp. 887–889 (2005)
11. Porayska-Pomsta, K., Mavrikis, M., Pain, H.: Diagnosing and acting on student affect: the tutor's perspective. User Modeling and User-Adapted Interaction: The Journal of Personalization Research 18, 125–173 (2008)
12. de Vicente, A., Pain, H.: Informing the detection of the students' motivational state: An empirical study. In: Proceedings of the Intelligent Tutoring Systems Conference (ITS). pp. 933–943 (2002)
13. Walonoski, J., Heffernan, N.: Prevention of off-task gaming behavior in intelligent tutoring systems. In: Proc. Intelligent Tutoring Systems (ITS). pp. 722–724 (2006)