Minimal Feedback during Tutorial Dialogue[⋆]

Pamela Jordan and Diane Litman

Learning Research and Development Center
University of Pittsburgh, Pittsburgh PA 15260

Abstract. We analyzed unexpected student responses to a natural language (NL) ITS to determine if improved feedback could be beneficial. Our analysis of a corpus of ITS-student dialogues suggests that unexpected responses represent learning opportunities. We outline our plans for testing feedback appropriate to subclassifications of these responses.

1 Introduction

Some patterns found in human-human and human-computer interactions predict similar outcomes while some do not [1–3]. So it is valuable to look at both types of interactions when building ITS. One study of human-human interactions suggests that increased tutor feedback predicts a bad learning outcome and that student errors do not account for these negative correlations [1]. Our research focuses on student responses that an ITS categorizes as unrecognizable (or *default*). In such cases ITS typically provide only minimal feedback that indicates whether a student action is correct or not [4]. The variety of ways in which ITS handle the default category (e.g. always treat it as either correct or incorrect) [4] suggests that choosing one fixed way of handling these responses may not suffice in all situations.

Default student responses range from those that are not fully correct to those that are simply failures to respond [4]. We are exploring whether some default responses arise because the student does not recognize what the tutor wants (i.e. tutor's communicative intentions) since this could lead students to respond in a way that ITS builders did not anticipate. Generally during human-human dialogue a hearer is expected to show to what degree he has understood a speaker [5] so that repairs are better informed. If an ITS provided this type of feedback, the human dialogue partner would have more information with which to make a repair and could improve the efficiency of the interaction. For example, if the ITS says "You're close but you need to be more careful. Try it again." then the student knows he should focus on adding details. Otherwise the student doesn't know what sort of repair to attempt.

2 The NL ITS Testbed and Corpus

The corpus we are using to analyze default student responses comprises 448 typed dialogues between 64 students and a NL ITS that tutors seven qualita-

[⋆] This work was funded by NSF grant 0325054

	Example of a *no attempt* response
system:	Which variable quantity does the ’24.0 N’ represent?
student:	I don’t know. *(choices: applied force, attempted,* **no attempt***)*
system:	The ’24.0 N’ represents the man‘s applied force on the crate.
	Example of an attempted but wrong response
system:	Which variable quantity does the ’24.0 N’ represent?
student:	Force of friction *(choices: applied force,* **attempted***, no attempt)*
system:	No. The ’24.0 N’ represents the man‘s applied force on the crate.
	Example of a vague response
system:	What quantities influence the kinetic energy of the truck?
student:	The velocity of travel and the mass of the truck *(choices: mass and magnitude of velocity,* **attempted***, no attempt)*
system:	I disagree. From $KE = 1/2 * m * v^2$, we can infer that kinetic energy is influenced by the truck’s magnitude of velocity and its mass

Table 1. Corpus excerpts

tive physics problems in the work-energy domain [6]. Domain experts identified 30 knowledge components (KCs) that were needed to solve the problems and developed a 33 item test for pre-/post-testing. Every item tests one or more KCs. Learning gains were significant for a majority of the KCs, as were composite learning gains (e.g. the lowest composite gain was for KCs about net work F(3,61)=3.2, p<.03).

The NL ITS guided students through problem solving and asked for justifications for key KCs. Because the corpus was collected for the purpose of deriving dialogue strategies on when to elicit responses and when to request justifications, its fully automated NL understanding module was replaced with a human interpreter, called the *wizard*, to reduce the confounds of misrecognizing student explanations. The wizard's interface showed the dialogue history, the student's last response and a list of choices for classifying the student's response. Dialogue excerpts from the corpus are illustrated in Table 1, where the choices available are shown after the student response and the wizard's selection is in bold.

3 An Analysis of Students' Default Responses

Only two *default* subclasses were available for wizards to select; *attempted* and *no attempt.* For *no attempt* responses such as “I don't know.” the system gave no minimal feedback before it followed up and for *attempted* responses, it gave negative feedback since the response did not answer the intended question. We found that 21% of all NL responses from students were classified as *attempted* and 12% as *no attempt.* But only 2% of responses turned out to be *no attempt* because *attempted* responses were sometimes misclassified by wizards as *no attempt.* Possibly this was done to circumvent students receiving negative feedback. But even with these misclassifications, there were still significant correlations between classifications of responses and learning; there was a significant moderate positive correlation between post-test scores (after removing the effects of pre-test scores) and the percentage of students' non-default responses (R=.47,p=0) and there were significant weak negative correlations between post-test scores and responses that were classified as *no attempt* (R=-.30,p=.017) or *attempted* (R=-.32,p=.011). The non-default responses are step specific and are either correct or non-correct ones that warrant a specific follow-up. As in other studies substantive responses from students were predictive of learning regardless of correctness. Negative correlations with learning suggest that default responses are possible learning opportunities and thus warrant further analysis.

4 Discussion and On-going Work

We identified the following alternative subclasses for *default* responses in the corpus; *no attempt*, *wrong*, *vague* and *overly specific*. Intuition suggests that *vague* and *overly specific* responses indicate the student is close to having learned a KC while the remaining *default* responses indicate the opposite. These subclasses could reflect the student's progress on a KC. When the student receives a pointer on what kind of error to look for and fix, it is reasonable to expect that if he attempts a repair he will move closer to a fully correct response. Furthermore, the student may be more motivated to pay attention to any specific feedback that follows a retry. Thus students may be able to achieve *correct* contributions sooner during their tutoring when they receive generic feedback that indicates the type of error and have a chance to retry. To test this hypothesis, we will conduct a controlled experiment with two conditions; the treatment condition will receive appropriate minimal feedback for each subclass of a *default* response and the control condition will receive no minimal feedback for any *default* response. We will compare the learning gains, learning curves and correctness during dialogues for the two conditions.

References

1. Chi, M., Roy, M., Hausmann, R.: Learning from observing tutoring collaboratively: Insights about tutoring effectiveness from vicarious learning. Cognitive Science (in press)
2. Jackson, G., Person, N., Graesser, A.: Adaptive tutorial dialogue in autotutor. In: Proceedings of Workshop on dialogue-based intelligent tutoring systems at ITS 2004. (2004)
3. Litman, D., Forbes-Riley, K.: Correlations between dialogue acts and learning in spoken tutoring dialogues. Natural Language Engineering (2006)
4. VanLehn, K.: The behavior of tutoring systems. International Journal of Artificial Intelligence and Education **16** (2006)
5. Clark, H.H.: Using Language. Cambridge University Press (1996)
6. VanLehn, K., Jordan, P., Litman, D.: Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. In: Proceedings of SLaTE Workshop on Speech and Language Technology in Education. (2007)