Student Initiative and Questioning Strategies in Computer-Mediated Human Tutoring Dialogues

Pamela W. Jordan and Stephanie Siler

Learning Research and Development Center
University of Pittsburgh
Pittsburgh PA 15260 *
[pjordan,siler]@pitt.edu

Abstract

In this paper we explore student initiative in human tutoring dialogues and how it potentially influences tutor questioning strategies. We describe our annotation scheme for tutor questions and student responses in which student initiative is simply defined as any unsolicited response from the student. We examine what is categorized as student initiative and propose some further distinctions that may be useful for improving mixed-initiative interaction in tutoring systems.

Keywords: annotation, human dialogue, initiative, dialogue strategies

1 Introduction

In human tutoring dialogues, a tutor who ignores signs of student confusion in order to complete his own plan runs the risk of preventing learning (Chi, 1996). We also know that active learning, which can involve a student making inferences, elaborating, justifying, integrating, predicting, etc., can be beneficial for students (Chi et al., 2001). Ideally then one would expect that maximizing the amount of confusion addressed and maximizing active learning behaviors in students to be worthy tutorial goals. Although human tutors may find it difficult to refrain from taking on the active learning role themselves (e.g. giving long explanations) and may at times fail to recognize or choose to ignore signs of confusion or unprompted active learning on the part of their students, they are undoubtedly better at recognizing it and responding to it than our current intelligent tutoring systems are. Our hope is that we can find enough instances of this sort of behavior in human tutorial dialogues that we can learn to similarly adapt in our intelligent tutoring systems.

For the purpose of this paper we will call these unexpected student behaviors that we wish to recognize and to which we wish to react and to encourage, *StudentInitiative*, to distinguish it from other definitions of initiative in task-oriented dialogues. These behaviors are initiatives because they have introduced or started something new that was not part of the tutor's current plan; they are an interruption to the forward progress of the tutor's current plan relative to his last plan step and how he expects the student to contribute to this plan. When a tutor reacts to a *StudentInitiative*, it has the effect of changing the direction of the tutoring session and helps to customize it to meet the needs of the individual student. Our definition of *StudentInitiative* is similar to that of (Shah et al., 2002) where they define it as any contribution that is not an answer to a question asked by the tutor[1]. (Green and Carberry, 1999) have a related definition of initiative in dialogue; it is any contribution that exceeds the speaker's obligation.

For example, if the tutor is asking a series of focused, short-answer questions, then the tutor is not planning to elicit an elaboration or some other form of active learning from the student and is expecting to continue on with his line of reasoning with some slight adjustments given the kinds of errors he expects to see. So if the student decides to elaborate after answering one of these questions, it was the

*This work was funded by NSF grant 9720359 and ONR grant N00014-00-1-0600

[1] We first became aware of the definition used in their study following the annotation study we report here.

student's own initiative to do so and the tutor can either ignore this unexpected additional response and continue forward with his plan or he can react to the elaboration by either terminating or interrupting his current plan and addressing the student's unexpected contribution. If the unexpected student contribution shows promise, we might expect the tutor to switch away from a focused line of questioning to a more open-ended one in which he tries to encourage further active learning behavior. However, if the contribution reveals some misconceptions or errors the tutor may want to pursue this further by eliciting more active learning behavior (e.g. Why do you think this is true?) or initiating a different focused line of questioning.

In human-computer interactions, our goal is to move away from the one-sided interactions that are typical of our current software applications toward mixed-initiative interactions that better reflect the type of collaborative interactions we are accustomed to when interacting with each other. With one-sided initiative, either the human initiates every action or the computer does. When the computer is in control of the interaction, typically it is not looking for the human to initiate anything without invitation and when the human is in control, he rarely sees the computer initiate an uninvited action since it is so poor at recognizing when it is helpful and appropriate to interrupt the software user. With mixed-initiative interactions, either participant can initiate actions without invitation so a software agent needs to be able to recognize when it will be helpful to interrupt and when to allow itself to be interrupted.

There are many proposals and definitions as to what qualifies as initiative because what is of greatest importance varies by the type of interaction involved as well as what we are able to computationally model (see (Chu-Carroll and Brown, 1998) for a survey of these proposals). The work in (Whittaker and Stenton, 1988; Walker and Whittaker, 1990) is the basis of one often used definition of dialogue initiative. It is simple to implement and has been successful at modelling some aspects of advisory and directive dialogues. But the emphasis of this work was to model topic shifts and to recognize topic boundaries. It looked at when shifts can occur and how the shifts are negotiated using high-level, easily recognizable speech acts. Utterances are classified into 4 types: assertions, commands, questions and prompts and correspondences between topic boundaries and these speech acts were analyzed to arrive at rules for predicting when a topic shift occurs. The rules indicate that a speaker has control when his last utterance is 1) an assertion unless it is a response to a question, 2) a command or 3) a question unless it is in response to a question or command. With a prompt or an explicit repetition of what has already been stated so far, the speaker is signalling that he is relinquishing control. But if the speaker has not relinquished control, the hearer can initiate taking over control by introducing a new topic regardless of whether or not he first responds to the speaker's previous contribution. These rules are actually better for predicting topic shift initiative than topic control since it overlooks the fact that topic shift initiatives can fail (i.e. the hearer can reject the new topic). We will call this definition of initiative *TopicInitiative*.

But as pointed out by further examinations of collaborative dialogues (e.g. (Jordan and Di Eugenio, 1997; Chu-Carroll and Brown, 1998)), *TopicInitiative* alone is not enough to adequately inform mixed-initiative interactions and misses a sizeable number of initiatives (e.g. 24% according to (Chu-Carroll and Brown, 1998) for one type of advisory dialogue). To help correct these shortcomings, (Chu-Carroll and Brown, 1998) suggests that initiative be defined as actions that direct how the agents' task should be accomplished and actions that initiate the establishment of mutual beliefs. But we claim that this particular definition of initiative, while more comprehensive, needs some adjustment to be useful for modelling mixed-initiative in tutorial dialogue. We propose to give more emphasis to the initiation of active learning behaviors. Explanations and justifications are understandably excluded as initiatives by the above definitions (i.e. they support the establishment of mutual beliefs but do not initiate them) but as we pointed out earlier they are examples of important learning actions on the part of the student.

Our goal is to use a more comprehensive definition of initiative that includes active learning behaviors and relate it to dialogue strategies that could encourage the behaviors that positively influence learning. So as not to exclude any type of initiative a priori, we chose to call any unsolicited action an initiative in our initial study of tutorial dialogue. Although we are able to reliably annotate using this definition of *StudentInitiative* (K=.89), we expect that we will need to refine it in order for it to be useful within a tutoring system. Automatically classifying student responses relative to expectations is reasonably accurate but simply treating whatever is unclassifiable as unexpected may not be enough information to decide whether a change in dialogue strategies is needed. But before we try for better classification features, we first need to understand what kinds of initiative are going to be most useful in tutorial dialogue by learning about which kinds of initiative are most influential in dialogue strategies.

In this paper, we will describe how we characterized *StudentInitiative* and our annotation results. We will begin to explore this data by estimating how many instances of *StudentInitiative* can be subcategorized as *TopicInitiative*. Our estimate shows that 78% of all instances of *StudentInitiative* are not *TopicInitiative*. Finally we will make some proposals for how to better characterize *StudentInitiative* and how to identify strategies useful for mixed-initiative interactions. While *StudentInitiative* and its relationship to tutor responses has been explored (Shah et al., 2002), its relationship to dialogue strategies that encourage it have not.

2 Corpus Overview

Our corpus is a collection of computer-mediated human tutoring dialogues in which a tutor presents a student with a qualitative physics problem from a set of 30 such problems. We currently have collected 199 dialogues in which 5 tutors and 35 students interact. We have analyzed 15 of these dialogues for *StudentInitiative* by annotating whether the student is providing new unsolicited information. These 15 dialogues represent interactions between 1 tutor and 4 students working on a set of 11 problems.

For this particular corpus we did not measure student learning gains but it is still worthwhile to study it to characterize the dialogue behaviors and patterns under the assumption that any tutoring is expected to be effective. But to complete our study we will need to find out if any of the dialogue features we identify have a positive influence on learning. We are currently collecting a similar corpus for 10 training problems where we are measuring learning gains. For both corpora, all the students have recently taken a high-school or college-level introductory physics course and experienced physics professors are providing a majority of the tutoring. Although some of the tutors in our first corpus are physics graduate students, our analysis in this paper is of an experienced physics professor.

For each tutoring session the tutor and student are in separate rooms. All of the dialogues about qualitative physics are conducted via the computer interface. The student and the tutor both explicitly give up their dialogue turn by pressing a submit button and during a turn no interruptions are allowed. While restricted turn-taking alters the nature of the dialogues (Oviatt and Cohen, 1991; Whittaker, 1995; O'Conaill et al., 1993; Jordan, 2000), it is still a valid (Clark, 1996) and successful form of human communication (Jordan, 2000). Restricted turn-taking has the advantages of simplifying both the dialogue analysis and the implementation of a computer tutor by eliminating overlapping language and the need to determine the intended sequencing.

Both the student and the tutor's computer interfaces comprise an area in which the physics question is displayed, an essay entry window and a dialogue window. The tutor also has the ability to select which question to present to the student and can enable and disable the student's essay and dialogue windows in order to encourage the student to separate the task of essay writing from that of engaging in a dialogue.

After the tutor presents a qualitative physics question, the student enters his answer and explanation in the essay window, as with the corpus excerpt shown in (1), and then the student and tutor engage in a dialogue to correct and improve that response, as with the corresponding corpus excerpt shown in (2)[2]. The essay becomes part of the student and tutor's common ground for the dialogue so that the dialogue and essay are not independent. The tutor's highest level dialogue topics clearly depend upon the essay but once a topic is introduced the dependence upon the essay lessens.

(1) Q: When a car without headrests on the seats is struck from behind, the passengers often suffer neck injuries. Why?

R: Because of the force of the car from behind hitting them, their heads are knocked back. This is because, their heads want to travel at the same speed that their car is moving and when the car hits them from behind the force causes this motion to be stopped.

[2]The notations at the end of each turn will be explained and used in later sections. The dialogue turns appear verbatim as entered by the student and the tutor.

	Tutor Questions	Student Responses
All	.63 (N=29,p<.001)	.89 (N=29,p<.001)
Unambiguous	.91 (N=25,p<.00005)	.88 (N=25,p<.001)

Table 1: **Kappa values for Tutor Questions and Student Responses**

(2) T0: What is the direction of the force experienced by the car when it is hit from behind? [CL]

S0: the direction is towards the car that hit it. [NI]

T1: Do mean that the car which is hit from behind will experience a force directed backwards? [CL/OE]

S1: i guess it would be experience a force forwards [NI]

T2: What will this force cause the car to do? [CL]

S2: accelerate [NI]

T3: Now analyze what will happen to the head, the neck if the head rest was not there? [OE]

S3: ok if the headrest was not there, when the car from behind hit the car it would give it acceleration, the persons body would move at the speed of the car during the acceleration, but the head would not. [NI]

T4: Why should the neck get hurt? [OE]

⋮

T6: What is caused by force? [CL]

S6: acceleration, so I should mention force and acceleration instead of speed. [SI]

3 Annotating StudentInitiative

We define any utterance that provides new unsolicited information as a show of *StudentInitiative*. Since our definition of *StudentInitiative* depends upon the tutor's expectations, we first annotated tutor questions as to whether the expectation is for a small, finite set of subject relevant responses with no larger discourse relation other than a response to a question being involved, or whether there is an expectation for an informational relation (such as justification) in which the content that is to be provided by the student cannot be characterized as a small finite set. We called the first category of tutor questions closed questions and the second open-ended questions.

StudentInitiative then was annotated based on whether the student did more than simply respond in the expected way relative to the category assigned to the tutor question. As part of our annotation guidelines, we decided that we should never look ahead in the dialogue when determining the categories for tutor questions. We did not want the annotation of the tutor question to be influenced by how the student actually responded since it was to represent what we thought the tutor expected from the student. In the dialogue shown in (2), we illustrate our annotation categories using the following notation at the end of each turn:

[NI]: student response with no show of *StudentInitiative*

[SI]: student response with *StudentInitiative*

[OE]: open-ended question from the tutor

[CL]: closed question from the tutor

To assessed the intercoder reliability of our categorizations we computed the Kappa coefficient of agreement(Krippendorff, 1980; Carletta, 1996) between two annotators on 2 of the 15 dialogues (12%). The Kappa values for agreement on annotating tutor questions and student responses is shown in the first row of Table 1. A Kappa value is constrained to the interval $[0, 1]$ where K=0 means the agreement is no different than chance and K=1 means there is perfect agreement. Using Krippendorf's scale to assess the Kappa values, where K < .67 is discounted, K > .8 is conclusive and anything in between is tentative, we had good agreement on student responses but not on tutor questions.

Since the categorization of student responses is dependent upon the category of the tutor question, one may wonder how the intercoder reliability for student responses can be good while that for tutor questions is not. Consider that a student can fail to respond in the expected way to either category of question. If one annotator characterizes a question as open-ended, the student response is almost always going to be classified as not being an instance of *StudentInitiative*. However, the student may still fail to respond fully to this question with, for example, the justification the annotator may have thought the tutor was expecting and may just give an answer with no justification. If another annotator classifies the same question as a closed question, the student response will still be annotated as a non-initiative since the student did not give an unsolicited justification along with the response. This sort of annotation disagreement seems reasonable to expect in this case since the student may also have missed that the tutor was implicitly requesting a justification as with [T1] and [S1] in (2).

The second row of Table 1 shows the Kappa values when we remove the cases in which we claim the tutor's intent is potentially ambiguous to the student. To identify these cases of tutor ambiguity we looked for instances of disagreement on the question category followed by a response in which there was agreement that it was [NI] (e.g. as with [T1] and [S1] in (2)). But for this ambiguity argument to become convincing we need to subcategorize the student responses according to whether they meet the perceived tutor expectation or not. Although [T1] and [S1] in (2) fit the pattern for ambiguity, the student did not directly respond to the tutor's question so we cannot tell whether the student interpreted the question as an implicit why or a yes/no question.

Although we created a taxonomy of question types to characterize open-ended and closed questions as part of our annotation guidelines, we did not annotate the question subcategories and check the intercoder reliability for them. However, it is still informative to look at these subcategories since it better defines how we chose to classify tutor questions.

For closed questions the subcategories are:

- Tutor asks student for body which force acts on (limited by problem statement)
- Tutor asks student for force (limited number of forces, in part because limited number of bodies that can cause forces in problem statement)
- dichotomous questions (e.g. yes/no, either/or)
- relational questions (e.g. is the body1's velocity greater than or equal to body2's)
- directional questions (e.g. in which direction is this acceleration)
- Tutor requests terminology (e.g. "what will you say in scientific terms about the motion of the block?")

For open-ended questions the subcategories are:

- why or how types of questions (e.g. "ok, but why do you say so?")
- definitional questions (e.g. "what is force?")
- "what happens when..?" questions (e.g. "what will happen to the head and neck if the head rest was not there?")
- Tutor asks student for complex calculation or reasoning (e.g. "if the rock falls through this distance, what will be its final velocity?")

In developing these subcategories for our annotation guidelines, we did not resort to classifying tutor questions simply on the basis of syntactic form. For example, we could not declare that all syntactic yes/no questions were closed questions. This is because many of the syntactic yes/no questions were implicit why questions in the context of the dialogue as in (3).

(3) Are you suggesting that acceleration, height of fall and initial velocity can be different for the two balls and yet they will take the same time to hit the ground?

4 Proposals Regarding StudentInitiative and Closed vs. Open-ended Questions

Looking again at the question subcategories in the previous section, in the case of the closed question subcategories, we can generalize these as either requesting a low-effort inference, a clarification or a disambiguation. We can further generalize open-ended questions as requests for the student to engage in what (Chi et al., 2001) calls deeper forms of construction that lead to active learning. For example, why or how questions request that the student make inferences and integrate materials by asking for justifications, "what happens when.." questions request that the student make predictions, and asking for the value of a quantity requests that the student do some further problem solving to arrive at a value. Furthermore, our question categories bear some resemblance to the student question categories in (Graesser and Person, 1994) and it may be fruitful to use these in helping us better characterize the purpose of instances of *StudentInitiative*. So we propose that we should try to categorize unsolicited student contributions by which question subcategory they could hypothetically be responding to in addition to whether student turns adequately meet the tutor's expectations. We expect that this additional categorization would help us to better distinguish which kinds of initiative will be most valuable to recognize in tutorial dialogue.

So why should we be as interested in recognizing active learning initiative as we have been in *TopicInitiative* (e.g. responding to student questions)? In the case where the student is showing active learning initiative, we need to recognize it and encourage the student to continue with it by giving feedback. If a tutoring system ignores it, it may discourage the student from taking on the role of an active learner when they are ready for it and the system simply mis-predicted when it will be most successful at prompting active learning behaviors from the student. Furthermore, in those cases where the tutor is prompting for active learning behaviors, we can examine when it is that the tutor is going to be most successful at getting the student to engage in active learning behavior.

We propose that closed and open-ended questioning are two high-level strategies that a tutoring system may want to choose between and that *StudentInitiative* could be one important factor to consider in that decision process. The open-ended questioning prompts for active learning whereas the closed questioning gives more guidance and prepares the student for active learning behaviors. In the 15 dialogues that we've annotated for open-ended questions and closed questions, the tutor has an almost equal distribution of both types of questions across the dialogues (55% closed vs. 45% open-ended) so it seems that these may indeed be two important strategies that are involved in successful tutoring.

We've begun a subcategorization of the student responses relative to tutor question types but have not yet checked the intercoder reliability for the response subcategories. Table 2, shows the distributions of the student response subcategorizations relative to closed and open-ended questions according to one annotator. Anything other than cases of no initiative are cases of *StudentInitiative*. From this we can estimate how much more has been identified as initiative by expanding our definition of initiative from *TopicInitiative* to *StudentInitiative*.

First we estimate which cases of *StudentInitiative* can be subcategorized as *TopicInitiative* by grouping together student turns that are responses followed immediately by questions and student turns that are changes to previous responses[3]. We will declare all other *StudentInitiatives* to be subcategorized as something other than *TopicInitiative* and will conservatively call these *OtherInitiative*. We expect a large majority of this *OtherInitiative* subgroup to be instances of active learning initiatives (e.g. responses followed by unsolicited explanations or expansions) but this still has to be tested. We see overall that there are more instances of *OtherInitiative* than of *TopicInitiative* (78% vs. 22%).

It is also the case that there is much more *StudentInitiative* in response to closed questions than to the open-ended questions (25% vs. 9% $p(\chi^2) < .001$) and that 82% of the instances of *StudentInitiative* following closed questions are in the *OtherInitiative* subgroup. This further suggests that *OtherInitiative* more so than *TopicInitiative* may be a signal for the tutor to discontinue a closed questioning strategy.

What we haven't yet begun to address in our analysis is how the tutor reacted to shows of *Studen-*

[3] We emphasize that this is just an estimate of what would be annotated as *TopicInitiative* since we did not distinguish between whether or not changes to previous responses are in direct response to what the tutor said. We are conservatively calling these *TopicInitiative* to give the favored definition of initiative the advantage.

15 dialogue sessions - 1 annotator	Closed	Open-Ended
Response only (NI)	103	104
Response + explanation or expansion (SI, OI)	29	6
Response + question (SI, TI)	3	4
Change previous response (SI, TI)	3	0
Totals	138	114

Table 2: **Preliminary Subcategorization of Student Responses Relative to Tutor Question Types** - NI=non-initiative, SI=*StudentInitiative*, OI=*OtherInitiative*, TI=*TopicInitiative*

tInitiative[4] relative to changes in questioning strategies and the interactions between the closed and open-ended questioning strategies. Some potential hypotheses of interest for us to explore in human tutorial dialogue are:

> Hypothesis-1: the less disruptive a *StudentInitiative* is the more likely it is that the tutor responds to it.
>
> Hypothesis-2: when the current closed or open questioning strategy fails switch to the other.
>
> Hypothesis-3: when the student demonstrates active learning behavior switch to or continue a strategy that encourages it.

In conclusion, our goal is to use a more comprehensive definition of initiative that includes active learning initiatives and relate it to dialogue strategies that could encourage the initiatives that positively influence learning. Other research in initiative in tutoring dialogues either uses a definition of initiative that excludes active learning initiatives (Moore, 2002), or looks only at immediate responses to active learning initiatives (Shah et al., 2002).

Acknowledgments

We thank Kurt VanLehn for his conceptualization of and initial guidance on the annotation scheme. We also thank Antonio Roque, Diane Litman and the reviewers for interesting and helpful discussions and comments.

References

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Lingustics*, 22(2):249–254.

Michelene T. H. Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G. Hausmann. 2001. Learning from human tutoring. *Cognitive Science*, 25(4):471–533.

Michelene T. H. Chi. 1996. Learning processes in tutoring. *Applied Cognitive Psychology*, 10:S33–S49.

Jennifer Chu-Carroll and Michael K. Brown. 1998. An evidential model for tracking initiative in collaborative dialogue interactions. *User Modeling and User-Adapted Interaction*, 8(3-4):215–253.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

[4](Shah et al., 2002) shows the dependencies between the communicative goals of the tutor's response and their subcategorization of *StudentInitiative* but were not able to conclude anything about the interaction with what they call the tutor's *delivery mode*. They conjecture that this may have been due to too much overlap in their categories for *delivery mode*. While *delivery mode* is not the same as the questioning strategies we describe, the relationship between them deserves some consideration.

Arthur C. Graesser and Natalie K. Person. 1994. Question asking during tutoring. *American Educational Research Journal*, 31:104–137.

Nancy Green and Sandra Carberry. 1999. A computational mechanism for initiative in answer generation. *User Modeling and User-Adapted Interaction*, 9(1-2):93–132.

Pamela W. Jordan and Barbara Di Eugenio. 1997. Control and initiative in collaborative problem solving dialogues. In *Computational Models for Mixed Initiative Interaction. Papers from the 1997 AAAI Spring Symposium. Technical Report SS-97-04*, pages 81–84. The AAAI Press.

Pamela W. Jordan. 2000. *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study*. Ph.D. thesis, Intelligent Systems Program, University of Pittsburgh.

Klaus Krippendorff. 1980. *Content Analysis: an Introduction to its Methodology*. Beverly Hills: Sage Publications.

Johanna Moore. 2002. Managing natural and effective tutorial interactions. Presentation at the 2002 Symposium on Natural Language Tutoring, University of Pittsburgh.

Brid O'Conaill, Steve Whittaker, and Sylvia Wilbur. 1993. Conversations over video conferences: An analysis of the spoken aspects of video mediated communication. *To Appear in Human Computer Interaction*.

Sharon L. Oviatt and Philip R. Cohen. 1991. Discourse structure and performance: Efficiency in interactive and non-interactive spoken modalities. *Computer Speech and Language*, pages 297–326.

Farhana Shah, Martha Evens, Joel Michael, and Allen Rovick. 2002. Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions. *Discourse Processes*, 33(1).

Marilyn A. Walker and Steve Whittaker. 1990. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proc. 28th Annual Meeting of the ACL*, pages 70–79.

Steve Whittaker and Phil Stenton. 1988. Cues and control in expert client dialogues. In *Proc. 26th Annual Meeting of the ACL, Association of Computational Linguistics*, pages 123–130.

Steve Whittaker. 1995. Rethinking video as a technology for interpersonal communications: theory and design implications. *International Journal of Human-Computer Studies*, 42:501–529.