

# Initiative Management for Tutorial Dialogue

Mark G. Core and Johanna D. Moore and Claus W. Zinn  
Division of Informatics, 2 Buccleuch Place, University of Edinburgh  
Edinburgh EH8 9LW, UK  
markc,jmoore,zinn@cogsci.ed.ac.uk

## Abstract

Tutors must maintain a delicate balance allowing students to do as much of the work as possible and to maintain a feeling of control, while providing students with enough guidance to keep them from becoming too frustrated or confused. In this paper, we describe our work in progress on defining a model of initiative and automatically detecting where initiative should shift in tutorial dialogue.

## 1 Motivation

Previous work on student learning has shown that one-on-one human tutoring is more effective than other modes of instruction. Tutoring raises students' performance as measured by pre- and post-tests by 0.40 standard deviations with peer tutors (Cohen et al., 1982) and by 2.0 standard deviations with experienced tutors (Bloom, 1984). What is it about human tutoring that facilitates this learning? Many researchers argue that it is the collaborative dialogue between student and tutor that promotes the learning (Merrill et al., 1992a; Fox, 1993; Graesser et al., 1995). Through collaborative dialogue, tutors can intervene to ensure that errors are detected and repaired and that students can work around impasses (Merrill et al., 1992b). Previous research has also shown that students must be allowed to construct knowledge themselves to learn most effectively (Chi et al., 1989; Chi et al., 1994). The consensus from these studies is that experienced human tutors maintain a delicate balance allowing students to do as much of the work as possible and to maintain a feeling of control, while providing students with enough guidance to keep them from becoming too frustrated or confused. Thus, we believe that correctly managing this control which we refer to as *initiative*, is critical to developing a successful intelligent tutoring system.

An initiative model indicates who has initiative and **when** the tutor should take initiative from the student. Timing is important; the tutor should allow students time to correct errors and misconceptions on their own, but take the initiative and intervene before the student becomes frustrated or confused.

The tutor should also allow the student to take the initiative to construct his<sup>1</sup> own knowledge by asking questions and proposing solutions. However, sometimes the tutor should take control from the student if the dialogue is becoming too unfocused or no progress is being made.

For an intelligent tutoring system to decide when it should take the initiative, it must know when it does not already have the initiative. Thus, the intelligent tutoring system needs to recognize when the student takes the initiative; linguistic cues (*e.g.*, "But" in example 1<sup>2</sup>) can help.

```
(1) T: If your formula is correct then IR=IV
      Do you still believe this?
      S: I suppose not.
      But doesn't I=V/R?
      ...
```

We are also interested in **how** the tutor should signal taking of initiative. Tutors will likely signal taking of initiative in many of the same ways as students. However, tutors may need to be more polite. In example 2, the student does not address the tutor's question.

```
(2) T: Can you tell me what is the
      voltage in the circuit now?
      S: How do I know resistance from
      looking at the circuit?
      T: First answer my question.
      Then I'll answer yours. :)
      ...
```

If the roles were reversed, it would be impolite of the tutor to ignore the student's question; the student might be discouraged from asking questions later and be less motivated to consult the tutor. We would like to explore the differences in how tutors and students mark the taking of initiative.

<sup>1</sup>In this paper, we refer to tutors and generic dialogue participants using female gender and refer to students using the masculine gender.

<sup>2</sup>The example comes from our corpus of human-human tutorial dialogues about basic electricity and electronics. See [http://www.cogsci.ed.ac.uk/~jmoore/tutoring/BEE\\_corpus.html](http://www.cogsci.ed.ac.uk/~jmoore/tutoring/BEE_corpus.html) for more details and to access the corpus.

Thus, our goals are (1) to build an initiative model that recognizes who has initiative and when the tutor should take initiative, and (2) to determine how tutors should signal taking of initiative (and how tutor signals differ from student signals). Our approach is to annotate human-human tutorial dialogues for initiative and cues for initiative shifts. We can use this data to train a model of initiative as well as recognize how the taking of initiative is signaled.

We first discuss the problem of defining initiative (section 2). We then review initiative management in existing dialogue systems (section 3). We focus specially on the initiative model of Chu-Carroll and Brown (1998) in section 4 and discuss how to extend it to tutorial dialogue in section 5. We then discuss our plans to annotate corpora (section 6) to train this model. We end the paper with discussion (section 7). Since our project is in the early stages, the goal of this paper is to discuss our plans rather than suggest we have results.

## 2 Defining Initiative

The intelligent tutoring systems literature does not precisely define initiative and we must look to the computational linguistics literature. Chu-Carroll and Brown (1998) point out that initiative as discussed in this literature can be broken into *task-initiative* and *dialogue-initiative*, although there is no general agreement on their exact meaning.

Chu-Carroll and Brown stipulate that a dialogue participant (DP) has the dialogue initiative if she takes the lead in determining the current discourse focus; a DP has the task initiative if she is taking the lead in the development of the participants' domain plan. Jordan and Di Eugenio (1997) state that DPs take task initiative when they perform a task-level action such as relaxing problem solving constraints, proposing a solution, or reconstructing a proposal. Jordan and Di Eugenio follow Walker and Whittaker (1990) and refer to dialogue initiative as control of the conversation. Green and Carberry (1999) note that initiative is taken when a DP contributes more information than was her obligation in a particular discourse turn (thus controlling the flow of information). According to Chu-Carroll and Brown's definitions, the type of initiative taken in this case depends on the meaning of the contribution (*e.g.*, contributing to the problem solution would mean taking both the dialogue and task initiative). Guinn (1996) says that a DP has the task initiative if she controls how mutual goals will be solved by the collaborators. Novick and Sutton (1997) propose a three factor model of initiative consisting of (i) the choice of task (determining what the conversation is about); (ii) the choice of speaker (governing turn taking); and (iii) the choice of outcome (*e.g.*, identifying the actions necessary to achieve the task). Novick and

Sutton's notion of turn taking as a third type of initiative did not catch on; most researchers prefer to separate turn taking from initiative.

## 3 Managing Initiative

We first look at initiative management in dialogue systems for non-tutorial tasks and then look at tutorial dialogue systems. Horvitz (1999) describes a dialogue system linking an email reader to a calendar program. Based on the content of the email message currently being read (*e.g.*, the presence of dates and times), the system must decide whether it should take the initiative by either (1) asking the user if she wants to use the calendar, or (2) bringing up the calendar interface without asking. The system faces the problem of when it should take initiative and how (ask a question or bring up the calendar).

Horvitz makes these decisions based on utility. If the system is very sure that displaying the calendar is useful, then the calendar is displayed. If the system is only somewhat sure that displaying the calendar is useful, then the system asks the user. In addition to content, other parameters could affect the estimated utility of bringing up the calendar: screen real estate (would bringing up the calendar overlap other windows?), user workload (is the user performing another task such as programming while reading her email?), and system success (is the user spending more time refusing unwanted help than it would take her to activate the calendar herself?). In determining when to take initiative, the system takes into account the estimated time needed to read the email and will not interrupt the reader during that time. If the user does not answer a system question (*e.g.*, "do you want to schedule an appointment?"), then after a certain length of time, the system assumes that the user is taking back the initiative.

(Litman and Pan, 2000) presents a dialogue system for database retrieval that uses an adaptive model of initiative. The system has three initiative management policies: (1) users can take the initiative at any time, (2) users can take the initiative when permitted by the system, (3) users can never take the initiative. Note, it is not clear how the system decides to give away initiative in policy 2. The system uses speech recognition confidence scores in deciding when to switch policies. The system will take the initiative when it cannot understand the user. The user is not allowed to direct the conversation when the system is having trouble understanding her. Note, in general, we can gauge understanding using confidence scores from natural language understanding components as well as speech recognizers (Walker et al., 2000).

Although tutoring systems must manage initiative (*e.g.*, deciding when to intervene in student problem solving, deciding whether to answer student ques-

tions), usually this is done through a fixed policy (*e.g.*, immediately flag student errors, always try to answer student questions). The EDGE system (Cawsey, 1989) has a more flexible approach to dealing with student questions. If the answer to the student's question is on EDGE's agenda of tutoring goals, EDGE will attempt to take back the initiative by asking the student if he will wait and see if his question is later answered. Otherwise, EDGE will answer the student's question.

The Duke Programming Tutor (Keim et al., 1997) has an innovative approach to topic selection (*i.e.*, choice of task, according Novick and Sutton's (1997) definition of initiative). The Duke Programming Tutor uses a *temperature-based* student model. The student model is a semantic network containing the concepts to be taught and the relations between them. Each node has a series of numeric features corresponding to: belief that the student understands the concept, importance of concept, distance from concept in focus, how many times this concept has been discussed, and student interest in this node. Temperature is a weighted sum of these features. If a student indicates he would like to discuss a particular topic, the student interest feature of the associated node in the model is increased. The temperatures are re-calculated and the tutor picks the node with the highest temperature as the next topic. Feature values propagate from adjacent nodes after the temperature is calculated. This propagation captures intuitions such as "if a student understands a topic then they are likely to understand related concepts" and "if a student is interested in a concept then they are likely to be interested in related concepts".

#### 4 Chu-Carroll and Brown's Model

The initiative models in the previous section were all designed to answer specific questions: When should the system bring up the calendar? When should it insist that a database query is filled out in a certain order? When should the computer-based tutor postpone a student question? When should it pick the topic? We turn to the work of Chu-Carroll and Brown (1998) to answer the more general question: when should a computer-based tutor take initiative? Chu-Carroll and Brown break cues to potential initiative shifts into three classes: *analytical*, *explicit*, and *discourse*. Analytical cues are based on the meaning of the utterances just spoken by the current dialogue participant (DP). In the case of tutoring, these cues would correspond to student utterances that the tutor must correct or clarify because they are incorrect (*e.g.*, "current and voltage are the same") or vague (*e.g.*, "the battery makes the circuit go"). The cues used in Keim *et al.*'s temperature model (1997) are analytical cues since they

depend on the meaning of previous utterances in the dialogue.

Explicit cues to initiative shifts are explicit requests for the other participant to take the initiative (*e.g.*, "Could you tell me how to connect the leads?") or explicit notifications that the speaker is taking initiative (*e.g.*, "Let me show you how leads are connected.").

Discourse cues to initiative shifts are the actions of asking questions, fulfilling obligations, and providing (no) new information. Discourse cues also include the kind of acoustic cues to initiative shifts used in (Litman and Pan, 2000). Any type of question directs the dialogue toward the topic of that question (*e.g.*, "what type of component is a battery?") giving initiative to the speaker of the question. After a speaker meets an obligation, initiative may shift. Meeting an obligation can involve performing a complex action or can be as simple as answering a question. In the constructed example below, after the student answer is accepted the initiative is up for grabs.

- (3) T: How do you connect the leads?  
 S: The red lead goes on tab 5 and  
     the black lead goes on tab 6  
 T: Right

Although Chu-Carroll and Brown do not equate turn-taking with initiative-taking, sometimes the giving away of a turn (by providing no new information) can indicate the giving away of initiative. The current speaker may pause to give the listener the opportunity to have a turn. The listener may abdicate by staying silent, uttering a prompt (acknowledgments without propositional content such as "yeah" or "uh-huh"), or repeating previously conveyed information.

Conversely, a speaker can take the initiative by providing more information than asked for as shown in the constructed example below. Here the tutor is only asking about the red lead but the student also discusses the black lead, taking the dialogue and task initiative. Even though the student's utterance may be wrong, he is at least attempting to further the problem solving process by discussing where the black lead should be attached.

- (4) T: Where would you connect the red lead?  
 S: I would connect the red lead to tab 5  
     and the black lead to tab 6

For each turn in a dialogue, Chu-Carroll and Brown use the Dempster-Shafer theory of evidence to predict who has the dialogue initiative and who has the task initiative. Chu-Carroll and Brown's model consists of three sets of bpa (basic probability assignment) functions: (1)

$m_{init,cue}(\{DP\})$ , (2)  $m_{init,curr-turn}(\{DP\})$ , and (3)  $m_{init,next-turn}(\{DP\})$ . *init* is either dialogue or task initiative, and *cue* is a particular cue to initiative shift such as 'question'. The first set of bpa functions encode the evidence given by *cue* that *DP* has *init* in the next turn. For example,  $m_{dialogue-init,question}(\{speaker\}) = 0.4$  means that questions give evidence of strength 0.4 out of 1.0 that the speaker of the current turn has dialogue-initiative in the next turn. The second set of bpa functions encode the overall evidence that *DP* has *init* in the current turn, and the third set of bpa functions encode the overall evidence that *DP* has *init* in the next turn. For example,  $m_{dialogue-init,curr-turn}(\{speaker\}) = 0.5$  means the strength of the evidence in support of the speaker having dialogue initiative for the current turn is 0.5 out of 1.0. Dempster-Shafer theory specifies how to combine the bpa functions of the current turn (sets 1 and 2) to calculate the bpa functions of set 3.

Chu-Carroll and Brown train the bpa functions of set 1 from a corpus labeled with cues for possible initiative shifts as well as who actually has dialogue and task initiative. The bpa functions are initialized such that cues have no impact on the likelihood that initiative shifts. As the training corpus is processed, the bpa functions are adjusted so that the resulting predictions of the model are correct. The model predicts that the speaker of the current turn has initiative, *init*, in the next turn if  $m_{init,next-turn}(\{speaker\}) \geq m_{init,next-turn}(\{listener\})$ . The predictive power of this framework was evaluated in four task-oriented domains and achieved, on average, 97% and 88% accuracy for task and dialogue initiative respectively.

That evaluation used 16 different cues that were annotated by humans. Rather than discuss these cues in detail, we focus on the cues used in MIMIC, Chu-Carroll's (2000) dialogue system for database retrieval. MIMIC's initiative model does not use explicit cues to initiative shifts (which might be difficult to recognize automatically) and does not use the discourse cues of questions (users do not generally ask questions) and obligations fulfilled (which might be difficult to recognize automatically). Chu-Carroll uses the discourse cues: TakeOverTask (the user provides more information than the system requested) and NoNewInfo (the recognized meanings of two consecutive user turns are identical), and the analytical cues of InvalidAction (the user query returns no results from the database), InvalidActionResolved, AmbiguousAction (a mandatory attribute is missing from a query or more than one value is specified for an attribute), and AmbiguousActionResolved. The bpa functions for these cues were trained on a labeled corpus in this database retrieval domain. The model predicts whether the system should have di-

alogue/task initiative in the next turn. Despite the limited evidence available to this model, Chu-Carroll and Nickerson (2000) showed that MIMIC was more successful than versions of the system that always took initiative or never took initiative.

## 5 Extending Chu-Carroll and Brown's Model

Chu-Carroll and Brown (1998) have differentiated dialogue and task initiative, noting that participants can change the topic of conversation without helping along the problem-solving process. In applying this work to tutorial dialogue, we have hypothesized the existence of *pedagogical initiative*. If a dialogue participant (DP) takes pedagogical initiative, she is taking control of the learning by changing the current set of learning goals or how those learning goals are being addressed. In tutorial dialogue, tasks to be performed are in service of learning goals rather than being ends in themselves. DPs can act on three different levels: dialogue, task, and pedagogy. By separating task and pedagogy, we can account for cases where the tutor takes the initiative to correct a student misconception unrelated to the task being performed, and cases where the tutor takes the initiative to address a learning goal even though the problem solving task has been completed.

We expect that Chu-Carroll and Brown's cues to initiative shifts can function to give away or take pedagogical initiative. A DP can explicitly ask to have pedagogical initiative (e.g., "S: I think I understand. Can I try again?"). There are also analytical cues to pedagogical initiative shifts. Student errors may indicate that the tutor should take pedagogical initiative and interrupt the student's problem solving efforts. Discourse cues can also signal shifts in pedagogical initiative; silence may indicate that the student is stuck and the tutor should intervene in the problem solving process. Fulfilling of a learning goal may mean that the pedagogical initiative is up for grabs.

A DP can take pedagogical initiative by asking a question; we repeat example (2) in (5) to illustrate this point. Here the tutor initially has dialogue, task, and pedagogical initiative. The task in this dialogue is to compute power. The tutor is initially controlling the topic of the conversation, how the problem is being solved, and how the associated learning goal is being addressed. The student takes dialogue, task, and pedagogical initiative in utterance 2; he wants to address resistance first not voltage. In utterances 3 and 4, the tutor takes back all initiative.

- (5) 1 T: Can you tell me what is the voltage in the circuit now?  
 2 S: How do I know resistance from looking at the circuit?

3 T: First answer my question.  
4 Then I'll answer yours. :)  
...

We hypothesize that the tutor deemed the student's question less relevant than her original question. If the tutor thought that the student would not know the voltage without knowing the resistance then the tutor would likely answer the question. However, in this case, the student should know the voltage without knowing the resistance. Thus, relevance impacts how initiative shifts.

Grice (1975) argues that dialogue participants' contributions should be relevant to the current topic. However, in tutorial dialogue, learning goals can be introduced at any time and students typically have a deficient model of the domain (of what is relevant). Thus, tutors may have to gently refuse to pursue irrelevant student-initiated tangents. The difficulty in investigating relevance is defining it precisely enough to test our hypothesis empirically. For now, we will leave relevance out of our initiative model.

In this section, we have been discussing cues signaling that initiative may shift in the next turn. A more global cue is student performance. We would like to test the hypothesis stated in (Sanders, 1995) that tutors are more likely to answer the tangential questions of good students. Student performance can be derived from the student model and dialogue history.

Given Chu-Carroll and Brown's (1998) success in predicting when initiative shifts, we plan to extend their model to our domain. Recall that this model used three sets of basic probability assignment functions:  $m_{init,cue}(\{DP\})$ ,  $m_{init,curr-turn}(\{DP\})$ , and  $m_{init,next-turn}(\{DP\})$ . We can add pedagogical initiative to the model by simply allowing *init* to range over dialogue, task, and pedagogical initiative. To modify the cues used by the model, we merely change the values that *cue* can range over.

Since we also want to see how initiative is taken, we need to test for correlations between initiative shifts and discourse markers such as "so", uncertainty markers such as "maybe", and words indicating the speaker is taking control ("Let's focus on the original question"). If correlations are found, these can be used as additional cues in our model as well as informing our natural language understanding and generation.

## 6 Annotating Initiative

We need to train the basic probability assignment functions discussed in the previous section to build an initiative model that recognizes who has initiative and when the tutor should take initiative. We also need to determine how the tutor should signal its taking of initiative. To meet these goals, we need

dialogues labeled with initiative and cues for initiative shifts.

Chu-Carroll and Brown (1998) annotated initiative and cues to initiative shifts. For the annotation of initiative, Chu-Carroll and Brown measured inter-annotator reliability using the kappa statistic (Siegel and Castellan Jr., 1988). Inter-annotator reliability was 0.57 for task initiative and 0.69 for dialogue initiative. Carletta (1996) defines a reliability threshold at 0.67; kappas above that value can be used to draw tentative conclusions and kappas below that value are unreliable.

Given Chu-Carroll and Brown's success in annotating dialogue initiative, we are optimistic that we also will be able to annotate dialogue initiative reliably. For task and pedagogical initiative, it is not clear whether these are fundamentally imprecise notions or whether they can be defined in an annotation manual such that annotators agree when task and pedagogical initiative shift. Chu-Carroll and Brown's annotation was informal and made use of spoken instructions to their annotators rather than an annotation manual. We will take the next step and write an annotation manual; have naive annotators label a corpus; and if necessary categorize the disagreements made by the annotators and repeat the process. A reliable annotation scheme is necessary if general claims are to be made about how initiative is marked. However, Chu-Carroll and Nickerson (2000) showed that even informal annotation is sufficient to improve system performance in the database retrieval domain.

Rather than attempt to write from scratch an annotation manual for cues to initiative shifts, we are taking elements from three currently existing annotation schemes: the DAMSL annotation scheme<sup>3</sup>, dialogue games (Mann, 1988), and the CIRCSIM annotation scheme (Kim, 1999). DAMSL annotates communicative function and is composed of forward and backward looking functions. Forward looking functions consist of statements, questions, and requests. Backward looking functions consist of responses and feedback, and include answer, accept, reject, and request-clarification. Dialogue games complement this annotation by labeling when a question is finally answered or agreement finally reached about the truth of a statement or commitment of one of the speakers. We take from the CIRCSIM annotation scheme labels of student correctness.

We believe these annotations will capture many of the cues for initiative shifts discussed in (Chu-Carroll and Brown, 1998). Ana-

---

<sup>3</sup>The DAMSL annotation scheme for communicative functions was developed by the Discourse Resource Initiative. For more information, see <http://www.cs.rochester.edu/research/cisd/resources/damsl/>

lytical cues are captured by student correctness and DAMSL tags. We anticipate breaking correctness into the following cues: incorrect, partly-correct, correct but irrelevant, correct with minor errors, and student-does-not-know. DAMSL tags `signal-non-understanding` and `request-clarification` mean that the speaker is asking for clarification of an ambiguous or incomprehensible utterance. DAMSL tags `reject` and `mixed-response` mean that the speaker is judging the previous utterance negatively. The DAMSL tag `hold` means the speaker is requesting more details before judging an utterance.

DAMSL tags will also identify discourse cues. Prompts (*e.g.*, “yeah”, “okay”) will be captured by `acknowledge` and `accept` tags. Questions are explicitly labeled in DAMSL. Cases where a responder does not answer a question but instead asks a question of their own can be identified as patterns of two questions in a row.

Dialogue games assign a default initiative. If a tutor asks a question (starting dialogue game A) and gets a wrong answer she may ask a followup question (starting the nested dialogue game A1). If the student gets this question right then game A1 ends. However if dialogue game A is still open (the question associated with A has not been answered) then the tutor still has initiative by default. Of course the student may override this default by asking a question of their own.

An advantage to labeling communicative function (DAMSL) and dialogue games is that we can determine the relationship between initiative, communicative function, and low-level discourse structure. For example, we can see if unanswered questions always mean initiative has shifted. We can see how initiative shifts over the course of a dialogue game. Since dialogues games can be nested, we can ask how embedded dialogues games effect the flow of initiative. Are embedded dialogue game boundaries likely places for initiative to change? The level of embedding may also have an impact on how initiative shifts. The tutor may try to prevent the dialogue from becoming too nested fearing that the student will get confused.

Below we summarize the cues to initiative shift to be used in our model and in investigating linguistic signals to initiative shifts. Note, we plan to derive silence automatically. We plan to estimate student performance by counting the number of wrong and partly-correct utterances made by the student.

```
silence
student performance
incorrect utterance
partly-correct utterance
correct but irrelevant utterance
utterance with minor errors
```

```
student says ‘‘I don’t know’’
signal of non-understanding
request for clarification
rejection
mixed-response
hold
acknowledgment
acceptance
question
question followed by a question
dialogue game assignment of initiative
```

We are optimistic that annotations produced by this hybrid scheme will be reliable. DAMSL has been shown to produce reliable annotations (Stent, 2000). Assuming annotators are domain experts, correctness should be unambiguous. Annotation of dialogue games will be more difficult. (Carletta et al., 1997) describes a promising but not totally reliable attempt to annotate dialogue games. The primary problem was unreliability about determining when there was a nested game. Intra-annotator reliability was tested by having an annotator recode a dialogue two months after she first annotated it, having worked with many other dialogues in between. The results show that she could replicate her decisions with high reliability indicating that well-written instructions should allow games to be labeled reliably. Proving so would be a significant result of our work.

We plan to annotate four sources of human-human dialogue data: (1) typed basic electricity and electronics (BEE) dialogues collected previously in our project, (2) spoken dialogues of experts explaining circuits to novices, used in building the EDGE system (Cawsey, 1989), (3) typed CIRCSIM dialogues where students and tutors discuss the circulatory system (Khuwaja et al., 1994), and (4) spoken direction giving dialogues from the Map Task Corpus (Anderson et al., 1991). The BEE dialogues are valuable because our goal is to produce a computer tutor replacing the human tutor in this domain. One drawback of these dialogues is that only one tutor was used and the corpus was collected using paid subjects who did not necessarily have any incentive to learn the material. The EDGE dialogues have four different explainers, and the CIRCSIM dialogues involve two tutors. In the CIRCSIM dialogues, the subjects are medical students and have an interest in learning the material presented by the tutor. Using these three corpora, we can balance our needs for a relevant corpus, exposure to different teaching styles, and a motivated set of subjects.

We will examine the Map Task dialogues to determine the differences in how initiative shifts (and how these shifts are marked) in task oriented and tutorial dialogues. The Map Task dialogues will allow us to compare spoken interaction to the typed interaction

of the BEE and CIRCSIM dialogues.<sup>4</sup> Map Task dialogues are already marked with a variety of annotations (*e.g.*, dialogue games) so this comparison should be relatively easy and may help us with our goal of applying previous research on task-oriented dialogues to tutorial dialogue systems.

## 7 Discussion

In this paper, we have outlined our plans (1) to build a model to recognize who has initiative and predict when the tutor should take initiative, and (2) investigate how initiative shifts are marked. It will also be necessary to evaluate and possibly retrain the tutorial system embodying these results.

We anticipate that all the cues discussed above can be recognized by the tutorial dialogue system, BEETLE<sup>5</sup>, that we are developing (Core et al., 2000). We hope that a robust parsing approach to natural language understanding will provide a level of understanding sufficient to determine DAMSL functions and allow the domain reasoner to gauge correctness despite the fact that student input is often ungrammatical. We anticipate that a simple set of rules should suffice to recognize DAMSL functions from parser output. The form of the utterance (interrogative, declarative, imperative) is one clue to its function. What is being asked for (action, permission, confirmation, clarification) further defines the functions of interrogative utterances. Lexical information is also important: *e.g.*, “no” -> reject, “okay” -> accept.

However, we have to face the fact that BEETLE’s natural language understanding skills will definitely be far below that of a human. The initiative management strategy learned during training may be perfect for the human tutor but not so good when confronted with natural language understanding errors. Such problems can be addressed in the evaluation stage.

During evaluation, users of BEETLE will take pre- and post-tests to gauge their learning gain, and answer questionnaires measuring perceived initiative and evaluating BEETLE as a tutor and dialogue system. Other measures can be automatically recorded whenever the system is used: dialogue efficiency (time, length of utterances/turns), intelligibility (number of clarification requests), and flow of initiative as perceived by the system. We hope to learn relationships such as “answering student questions leads to more clarification requests” and “many clarification requests harm learning gain”. Our goal is to develop an initiative model that can be retrained based on this data.

<sup>4</sup>The EDGE dialogues were also spoken but we only have transcripts for them and not the original speech.

<sup>5</sup>BEETLE stands for Basic Electricity and Electronics Tutorial Learning Environment.

Another goal of the evaluation is performing a first test of whether an adaptive model of initiative is really more effective than a fixed initiative tutor. To perform this test, we will use three versions of BEETLE. BEETLE-MI will be a “mixed-initiative” tutor incorporating the initiative model described in this paper. BEETLE-TI will always have the task, dialogue, and pedagogical initiative just like the CIRCSIM tutor (Khuwaja et al., 1994), and BEETLE-SI just like the Andes tutor (Schulze et al., 2000) will never take any initiative. Although the test will be preliminary we hope that the advantages of an adaptive model of initiative will be immediately obvious.

## References

- Anne H. Anderson, Miles Bader, Ellen G. Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson, and Regina Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34(4):351–366.
- B. S. Bloom. 1984. The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. In *Educational Researcher*, volume 13, pages 4–16.
- Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.
- Jean Carletta. 1996. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(2):249–254.
- Alison Cawsey. 1989. Explanatory dialogues. *Interacting with Computers*, 1(1):69–92.
- Micheline T. H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2):145–182.
- Micheline T. H. Chi, Nicholas de Leeuw, Mei-Hung Chiu, and Christian Lavancher. 1994. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477.
- Jennifer Chu-Carroll and Michael K. Brown. 1998. An evidential model for tracking initiative in collaborative dialogue interactions. *User Modeling and User-Adapted Interaction*, 8:215–253. Special issue on Computational Models of Mixed Initiative Interaction.
- Jennifer Chu-Carroll and Jill S. Nickerson. 2000. Evaluating automatic dialogue strategy adaptation for a spoken dialogue system. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, pages 202–209.
- Jennifer Chu-Carroll. 2000. Mimic: An adaptive

- mixed initiative spoken dialogue system for information queries. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 97–104.
- Peter A. Cohen, James A. Kulik, and Chen-Lin C. Kulik. 1982. Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19:237–248.
- Mark G. Core, Johanna D. Moore, and Claus Zinn. 2000. Supporting constructive learning with a feedback planner. Technical Report FS-00-01, American Association for Artificial Intelligence, 445 Burgess Drive, Menlo Park CA 94025. Papers from Symposium on Building Dialogue Systems for Tutorial Applications.
- Barbara A. Fox. 1993. *The Human Tutorial Dialogue Project: Issues in the design of instructional systems*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Arthur C. Graesser, Natalie K. Person, and Joseph P. Magliano. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9:495–522.
- Nancy Green and Sandra Carberry. 1999. A computational mechanism for initiative in answer generation. *User Modeling and User-Adapted Interaction*, 9:93–132.
- H. Paul Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics. Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.
- Curry I. Guinn. 1996. An analysis of initiative selection in collaborative task-oriented discourse. In *Abridged version in: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 278–285.
- Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of CHI'99, ACM SIGCHI Conference on Human Factors in Computing Systems, Pittsburgh, PA, May*.
- Pamela W. Jordan and Barbara Di Eugenio. 1997. Control and initiative in collaborative problem solving dialogues. In *AAAI 1997 Spring Symposium on Computational Models for Mixed Initiative Interactions*, Stanford, CA.
- Greg A. Keim, Michael S. Fulkerson, and Alan W. Biermann. 1997. Initiative in tutorial dialogue systems. Technical Report SS-97-04, American Association for Artificial Intelligence, 445 Burgess Drive, Menlo Park CA 94025. Papers from Symposium on Computational Models for Mixed Initiative Interactions.
- Ramzan A. Khuwaja, Martha W. Evens, Joel A. Michael, and Allen A. Rovick. 1994. Architecture of CIRCSIM-tutor (v.3): A smart cardiovascular physiology tutor. In *Proceedings of the 7th Annual IEEE Computer-Based Medical Systems Symposium, Winston-Salem, NC*, pages 158–163. IEEE Computer Society Press.
- Jung Hee Kim. 1999. A manual for SGML mark up in tutoring transcripts. From the CIRCSIM-Tutor Project, March.
- Diane J. Litman and Shimei Pan. 2000. Predicting and adapting to poor speech recognition in a spoken dialogue system. In *Proceedings of the National Conference on Artificial Intelligence (AAAI '00)*.
- William C. Mann. 1988. Dialogue games: Conventions of human interaction. *Argumentation*, 2:511–532.
- Douglas C. Merrill, Brian J. Reiser, and S. Landes. 1992a. Human tutoring: Pedagogical strategies and learning outcomes. Paper presented at the annual meeting of the American Educational Research Association.
- Douglas C. Merrill, Brian J. Reiser, Michael Ranney, and J. Gregory Trafton. 1992b. Effective tutoring techniques: Comparison of human tutors and intelligent tutoring systems. *Journal of the Learning Sciences*, 2(3):277–305.
- David G. Novick and Stephen Sutton. 1997. What is mixed-initiative interaction? In *AAAI 1997 Spring Symposium on Computational Models for Mixed Initiative Interactions*, Stanford, CA.
- Gregory A. Sanders. 1995. *Generation of Explanations and Multi-Turn Discourse Structures in Tutorial Dialog, Based on Transcript Analysis*. Ph.D. thesis, Illinois Institute of Technology.
- K. G. Schulze, R. N. Shelby, D.J. Treacy, and M. C. Wintersgill. 2000. Andes: A coached learning environment for classical newtonian physics. In *Proceedings of the 11th International Conference on College Teaching and Learning*, Jacksonville, FL, April.
- Sidney Siegel and N. John Castellan Jr. 1988. *Non-parametric Statistics for the Behavioral Sciences*. McGraw-Hill, second edition.
- Amanda J. Stent. 2000. The monroe corpus. Technical report, Department of Computer Science, University of Rochester, Rochester, NY 14627-0226.
- Marilyn A. Walker and Steve Whittaker. 1990. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 70–79.
- Marilyn A. Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane J. Litman. 2000. Learning to predict problematic situations in a spoken dialogue system: Experiments with how may i help you? In *Proc. of the 1<sup>st</sup> Annual Meeting of the North American Chapter of the ACL (NAACL-00)*, Seattle.