An Evidential Model for Tracking Initiative in Collaborative Dialogue Interactions

Jennifer Chu-Carroll and Michael K. Brown
Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974, U.S.A.; E-mail: {jencc,mkb}@bell-labs.com.

Abstract.
In this paper, we argue for the need to distinguish between *task initiative* and *dialogue initiative*, and present an evidential model for tracking shifts in both types of initiatives in collaborative dialogue interactions. Our model predicts the task and dialogue initiative holders for the next dialogue turn based on the current initiative holders and the effect that observed cues have on changing them. Our evaluation across various corpora shows that the use of cues consistently provides significant improvement in the system's prediction of task and dialogue initiative holders. Finally, we show how this initiative tracking model may be employed by a dialogue system to enable the system to tailor its responses to user utterances based on application domain, system's role in the domain, dialogue history, and user characteristics.

Key words: initiative, control, dialogue systems, collaborative interactions

This paper has not been submitted elsewhere in identical or similar form, nor will it be during the first three months after its submission to UMUAI.

1. Introduction

Naturally-occurring collaborative dialogues are very rarely, if ever, one sided. Instead, initiative of the interaction shifts among dialogue participants in a primarily principled fashion, signaled by features such as linguistic cues, prosodic cues, and in face-to-face interactions, eye gaze and gestures. Furthermore, patterns of initiative shifts between participants may differ depending on the task the participants are attempting to accomplish, on the roles they each play in this task, and on their experience in interacting with each other in current and previous dialogues. Thus, in order for a dialogue system to successfully collaborate with its user on their interaction, it must be able to dynamically track initiative shifts during their interaction by recognizing the user's cues for initiative shifts, and by providing appropriate cues, when necessary, in its responses to user utterances.

To illustrate merely one of the many decisions that a dialogue system must make when generating responses to user utterances, consider the following dialogue segment between a bank teller and a customer where we show three possible responses that the teller may provide in response to the customer's question:

(1) C: I need some money.

(2) *How much do I have in my 6-month CD?*

(3a) T: *You have $5,000 in that CD.*

(3b) T: *You have $5,000 in that CD, but that CD will not mature for another 3 months.*

(3c) T: *You have $5,000 in that CD, but that CD will not mature for another 3 months. However, you have $3,000 in another CD that will be available next week.*

In response (3a), the teller directly answers the customer's question. In (3b), the teller conveys her belief in the invalidity of the customer's proposed plan, while in (3c), she conveys the invalidity of the proposal and proposes an alternative solution. Given that all three alternative responses are reasonable continuations of the dialogue, the question that comes to mind is *what criteria should a dialogue system adopt to select one response over the others given a particular circumstance?* Existing cooperative response generation systems (e.g. (van Beek, 1987; Pollack, 1990; Chu-Carroll and Carberry, 1994)) are able to select response (3a) versus (3b)/(3c) based on whether or not the customer's proposal is in fact valid; however, to our knowledge, no existing model is able to distinguish between responses (3b) and (3c), and to determine the circumstance under which each response is appropriate. We propose to approach this decision-making process from an initiative point of view, i.e., by viewing the difference between the alternative responses as a difference between the levels of initiative exhibited by each dialogue participant. This model then allows a dialogue system to select an appropriate response based on how initiative is distributed among the participants.

In the rest of this paper, we argue that most existing models of initiative conflate two types of initiatives, which we call *task initiative* and *dialogue initiative*, and show how distinguishing between these two types of initiatives accounts for phenomena in collaborative dialogues that previous models were unable to explain. In particular, we show that the distinction among responses (3a)-(3c) in the above dialogue can be modeled by the distribution of these two types of initiatives between the dialogue participants, i.e., the teller having neither task nor dialogue initiative in (3a), having dialogue but not task initiative in (3b), and having both task and dialogue initiatives in (3c). We discuss our evidential model for tracking initiative shifts between dialogue participants during their interaction using a set of cues that can be recognized based on linguistic and domain knowledge alone, i.e., not considering physical cues such as gesture and eye gaze. We demonstrate 1) the performance of our model by showing that it can correctly predict the task and dialogue initiative holders in 99.1% and 87.8% of the dialogue turns (compared against manually labeled initiative shifts), respectively, in the TRAINS domain in which the model is trained, and 2) the generality of the model by showing that its application in various other collaborative domains consistently increases the accuracies in the prediction of task and dialogue initiative holders by 2-4 and 8-13 absolute percentage points, respectively, compared to a simple prediction method without the use of cues. Finally, we illustrate the usefulness of such an initiative tracking model by showing how it may be incorporated into a dialogue system to allow the dialogue system to vary its responses to user utterances based on the factors that affect patterns of initiative shifts, such as the application domain, the system's role in the dialogue, the dialogue history, and individual user characteristics.

2. Related Work

2.1. Views of Initiative

Previous work on mixed-initiative dialogues focused mainly on tracking and allocating a single thread of control among dialogue participants (Novick, 1988; Whittaker and Stenton, 1988; Walker and Whittaker, 1990; Kitano and Van Ess-Dykema, 1991; Smith and Hipp, 1994; Guinn, 1998; Lester et al., 1998). However, these researchers differ in terms of what they consider to be *initiative*, and the circumstances under which an agent is considered to have the initiative in a dialogue. Novick considered a dialogue participant to have the initiative if he controls the flow and structure of the interaction (Novick, 1988). Whittaker, Stenton, and Walker (Whittaker and Stenton, 1988; Walker and Whittaker, 1990) equated initiative with control, and argued that as *initiative* passes back and forth between the discourse participants, *control* over the conversation gets transferred from one participant to another. Kitano and Van Ess-Dykema (1991) considered an agent to have control of the conversational initiative if the agent makes an utterance that instantiates a discourse plan based on her domain plan, i.e., if the agent makes a task-related proposal. Smith and Hipp considered initiative in dialogue to be a representation of the control of the task (whose goals currently have priority), and argued that the level of initiative in the dialogue should mirror the level of initiative in the task (Smith and Hipp, 1994). Guinn considers an agent to have initiative over a mutual (task) goal when the agent controls how that goal will be solved by the collaborators (Guinn, 1996; Guinn, 1998).[1] Finally, in their work on mixed initiative problem solving, Lester et al. (1998) equate initiative with the control on problem solving.

More recently, researchers have come to the realization that simply tracking one thread of initiative does not adequately model human-human dialogue

[1] Guinn provides definitions for both *task initiative* and *dialogue initiative*, but argues that when an agent holds the task initiative, he must also hold the dialogue initiative.

interactions. Novick and Sutton (1997) analyzed the earlier views of initiative and proposed a multi-factor model of initiative, including *choice of task*, *choice of speaker*, and *choice of outcome*. Choice of task determines what the conversation is about, choice of speaker models turn-taking among dialogue participants, and finally, choice of outcome allocates the decision or action necessary to achieve the task. Jordan and Di Eugenio (1997), on the other hand, argued in contrary to Whittaker, Stenton, and Walker's view that *control* and *initiative* are equivalent. Instead, they proposed that *control* should be used to describe a dialogue level phenomenon, and that *initiative* should refer to the participants' problem solving goals. Cohen et al. (1998) presented alternative theories of initiative. In one of their theories, they argued that theories of initiative that model merely the flow of the conversation or the task-level actions in collaborative problem solving are problematic because of their inappropriate mixing of *initiative* and *conversational control*. As a result, they proposed a theory that follows Jordan and Di Eugenio's distinction between initiative and control. Rich and Sidner (1998) distinguish between *global initiative*, which is concerned with whether or not a dialogue participant has something to say, and *local initiative*, which addresses problems local to a discourse segment such as turn-taking and grounding. In their interface agent, however, they focus mainly on modeling global initiative. In Section 3, we motivate our view of initiative, which consists of modeling *task initiative* and *dialogue initiative* (Chu-Carroll and Brown, 1997a). Note that although our distinction between task and dialogue initiatives is similar to Novick and Sutton's choice of outcome and choice of task, as well as to Jordan and Di Eugenio's initiative and control, these ideas were conceived independently.

2.2. Analysis and Use of Initiative

Previous work on mixed-initiative interaction can loosely be grouped into three classes. First, some researchers have developed models that capture mixed-initiative behavior in dialogues. Novick developed a computational model that utilizes *meta-locutionary acts*, such as *give-turn*, *clarify*, and *confirm-mutual*, to explain issues such as turn-taking, negotiation of reference, and confirmation of the mutuality of knowledge in mixed-initiative dialogue interaction (Novick, 1988). In addition, Kitano and Van Ess-Dykema extended Litman and Allen's plan recognition model (Litman and Allen, 1987) to explicitly track the conversational initiative based on the domain and discourse plans behind the utterances (Kitano and Van Ess-Dykema, 1991).

Second, some researchers have investigated the causes of initiative shifts in mixed-initiative dialogue interaction and their effect on the structure of discourse. Whittaker and Stenton (1988) devised rules for allocating dialogue control based on utterance types, which included *assertions*, *commands*, *questions*, and *prompts*. They then analyzed patterns of control shifts by applying their rules to a set of expert-client dialogues on resolving software problems. They noted that the the majority of control shifts are signaled by *prompts*, *repetitions*, or *summaries*, while in the remainder of the cases, control shifts as a result of *interruptions*. Walker and Whittaker (1990) subsequently utilized the control allocation rules devised by Whittaker and Stenton to perform 1) a comparative study of the cues used to signal control shifts in different types of dialogues (advice-giving dialogues and task-oriented dialogues), and 2) an analytical study of control segmentation and the structure of discourse by analyzing the distribution of anaphora with respect to control segments in advice-giving dialogues.

Third, some researchers have taken into account the notion of initiative in dialogue interactions and have developed dialogue systems that vary their responses to user utterances based on their models of initiative. Smith and Hipp (1994) developed a dialogue system that varies its responses to user utterances based on four dialogue modes, *directive*, *suggestive*, *declarative*, and *passive*. These dialogue modes characterize the level of initiative that the system has in a dialogue, and affect the topic selection in the system's response generation process. For instance, in the *directive* dialogue mode, the system has complete dialogue control and will not allow interruptions from the user to other subdialogues. Thus, during topic selection, the system simply pursues its current goal without regard for the user's focus. However, in their system, the dialogue mode is determined at the outset and cannot be changed during the dialogue. Guinn (1996; 1998) subsequently developed a system that allows change in the level of initiative based on each agent's competency in completing the current subtask. Guinn employs a probabilistic model for evaluating competency based on the likelihood of each agent's path for solving the goal being successful. The initiative setting of the dialogue is then based on the result of this competency evaluation.

Our work overlaps with work in classes two and three above. We investigated potential cues that trigger initiative shifts, and developed a model that tracks the distribution of task and dialogue initiatives between participants during the course of a dialogue based on the combined effect of a set of observed cues. The long term goal of this work is to incorporate our initiative tracking model into a dialogue system in order to allow the system to tailor its responses to user utterances under different circumstances.

3. Task Initiative vs. Dialogue Initiative

3.1. Motivation

As discussed in the previous section, most existing models of initiative focus on tracking a single thread of initiative, often considered to be the *conversational lead*, among dialogue participants during their interaction. However, we argue that merely maintaining the conversational lead is insufficient for modeling complex behavior that affects a dialogue system's decision making process during response generation. We illustrate this argument by analyzing in further detail the dialogue segment shown in Section 1 using a model of initiative that tracks only the conversational lead, such as that of Whittaker and Stenton (1988).

In utterances (1) and (2), C states her goal and requests for information as part of formulating a plan to achieve this goal; thus C has the conversational lead at this point in the dialogue. In utterance (3a), T provides a direct response to C's question; hence the conversational lead remains with C. On the other hand, in utterances (3b) and (3c), instead of merely answering C's question, T takes control of the dialogue by further initiating a subdialogue to correct C's invalid plan of attempting to withdraw money from an immature CD. However, existing models for initiative, when adopted by a dialogue system, cannot distinguish between situations when it may be more appropriate to provide a response such as (3b) and those when a response such as (3c) may be more desirable. A comparison between (3b) and (3c) shows that the two responses differ in the level of involvement that T has in the agents' planning process. More specifically, in (3b), T merely conveys the invalidity of C's proposal, while in (3c), T further actively participates in the planning process by explicitly proposing what she believes to be a valid plan for achieving C's goal. Based on this observation, we argue that, in a collaborative problem-solving environment, it is necessary to distinguish between *task initiative*, which tracks the lead in the development of the agents' plan, and *dialogue initiative*, which tracks the lead in determining the current discourse focus (Chu-Carroll and Brown, 1997a). This distinction then allows us to explain T's behavior from a response generation point of view: in (3b), T responds to C's proposal by merely taking over the dialogue initiative, i.e., changing the discourse focus to inform C of the invalidity of her proposal, while in (3c), T responds by taking over both the task and dialogue initiatives, i.e., taking the lead in the planning process by suggesting an alternative plan which she believes to be valid. In other words, by modeling both task and dialogue initiatives, a dialogue system is able to determine the appropriate distribution of task and dialogue initiatives for each dialogue turn, and thus tailor its responses to user utterances based on its model of initiatives.

An agent is said to have the *task initiative* if she is directing how the agents' task should be accomplished, i.e., if her utterances directly propose *actions* that she believes the agent(s) should perform. The utterances may propose *domain actions* (Litman and Allen, 1987) that directly contribute to achieving the agents' goal, such as *"Why don't we couple engine E2 to the boxcar that's at Elmira, and send it to Corning."*[2] On the other hand, the utterances may

[2] The majority of the examples in this paper are taken from (Gross et al., 1993).

propose *problem-solving actions* (Allen, 1991; Lambert and Carberry, 1991; Ramshaw, 1991) that contribute not directly to the agents' domain goal, but to how they will go about constructing a plan to achieve this goal, such as *"Let's look at the first [problem] first. I think they are separate."* An agent is said to have the *dialogue initiative* if she takes the conversational lead in order to establish mutual beliefs between the agents, such as mutual beliefs about a piece of domain knowledge or about the validity of a proposal. For instance, in response to agent A's proposal of sending a boxcar to Corning via Dansville, agent B may take over the dialogue initiative (but not the task initiative) by saying *"We can't go by Dansville because we've got engine E1 going on that track."* Note that although these utterances contribute indirectly to the formulation of the final plan (by suggesting that the current plan *not* be adopted), since they do not directly propose actions to be added to the plan, they affect initiative only at the dialogue level. The relationship between task and dialogue initiatives is such that when an agent takes over the task initiative, he also takes over the dialogue initiative, since a proposal of actions can be viewed as an attempt to establish the mutual belief that a set of actions be adopted. On the other hand, an agent may take over the dialogue initiative without taking over the task initiative, as in response (3b) in the sample dialogue. This agrees with Cohen et al.'s theory of initiative in which they argued that when an initiative shift occurs, the agent who took the initiative also has taken control of the conversation, but not vice versa (Cohen et al., 1998).

Our distinction between task and dialogue initiatives agrees with Jordan and Di Eugenio's view of initiative, which distinguishes between initiative, applicable to the agents' problem-solving goals, and control, pertaining to the dialogue level (Jordan and Di Eugenio, 1997). This distinction is further supported by Cohen et al.'s theory of initiative where they again separate control from initiative (Cohen et al., 1998). Novick and Sutton proposed a multi-factor model for modeling initiative based on choice of task, choice of speaker, and choice of outcome (Novick and Sutton, 1997). Our task initiative corresponds to their choice of outcome, while our dialogue initiative corresponds to their choice of task.[3] On the other hand, in Guinn's model, an agent has *task initiative* over a goal if he dictates how the agents will go about achieving the goal, while an agent has the *dialogue initiative* when both agents expect the agent to communicate next (Guinn, 1998). However, although Guinn provides distinct definitions for task initiative and dialogue initiative, he also equates task and dialogue initiatives, argues that an agent who holds the task initiative over the current mutual goal must also hold the dialogue initiative, and models only one type of initiative in his system. We contend that Guinn's framework is insufficient for modeling the type

[3] Their choice of speaker models initiative at the turn-taking level, which is not represented in our view of initiative.

of phenomenon we attempt to model in collaborative planning dialogues by distinguishing between task and dialogue initiatives. More specifically, by equating task and dialogue initiatives, Guinn's model equates being expected to communicate next (having the dialogue initiative in his model) to having control over the current goal (having the task initiative in his model). However, our earlier sample dialogue illustrated a counterexample to this view in that although T was expected to communicate next after C's question in utterance (2), she did not have control over the current goal in her response in utterance (3a).

3.2. Corpus Analysis

To analyze the distribution of task and dialogue initiatives in collaborative dialogues, we analyzed the TRAINS91 dialogues in which two agents are collaborating on planning a route for cargo shipping along a hypothetical railway system (Gross et al., 1993). The TRAINS91 corpus contains 16 dialogues based on 8 speaker pairs, and contains a total of 1000 dialogue turns. We manually labeled each dialogue turn in the corpus with two labels, *task initiative holder (TIH)* and *dialogue initiative holder (DIH)*. Each label can be assigned one of two values, *system* or *user*, depending on which agent holds the task/dialogue initiative during that turn.[4] Table I shows the distribution of task and dialogue initiatives in the TRAINS91 dialogues. The results of our analysis shows that although in the majority of turns the task and dialogue initiatives are held by the same agent, in approximately 27% of the turns the agents' behavior can be better accounted for by tracking the two types of initiatives separately.

Table I. Distribution of Task/Dialogue Initiatives

	TIH: System	**TIH: User**
DIH: System	37 (3.5%)	274 (26.3%)
DIH: User	4 (0.4%)	727 (69.8%)

To further justify the need to distinguish between task and dialogue initiatives, and to assess the reliability of our annotation, approximately 10% of the dialogues were annotated by two additional coders. In this subset of the dialogues, the original coder found that in 37% of the dialogue turns, the task and dialogue initiatives were held by different agents, while the two additional coders found that the two types of initiatives are held by different

[4] The task/dialogue initiative holder is determined on a turn by turn basis. For instance, an agent holds the task initiative during a turn as long as *some* utterance during that turn directly proposes how the agents should accomplish their goal, as in utterance (3c).

participants in 21% and 14% of the dialogue turns, respectively. Although there is a discrepancy in the percentages presented above, on average, the coders found that in 24% of the dialogue turns, the agents' behavior can be better accounted for by tracking the two types of initiatives separately, thus providing further support for the need to distinguish between task and dialogue initiatives.

Next, we used the kappa statistic (Siegel and Castellan, 1988; Carletta, 1996) to assess the level of agreement among the three coders on this subset of the dialogues. In this experiment, K is .57 for the task initiative holder agreement and K is .69 for the dialogue initiative holder agreement. Carletta reports that content analysis researchers consider $K > .8$ to be good reliability, with $.67 < K < .8$ allowing tentative conclusions to be drawn (Carletta, 1996).[5] Strictly based on this metric, our results indicate that the three coders have a reasonable level of agreement with respect to the dialogue initiative holders, but do not have reliable agreement with respect to the task initiative holders. We attribute this low level of agreement to two possible causes. First, the kappa statistic is known to be highly problematic in measuring inter-coder reliability when the likelihood of one category being chosen overwhelms that of the other (Grove et al., 1981), which is the case for the task initiative distribution in the TRAINS91 corpus, as shown in Table I. Second, the low level of agreement may be due to the fact that a small number of dialogues were used in the coding reliability test, and that these dialogues may not be representative of the corpus. Furthermore, as will be shown in Table VI, Section 6, the task and dialogue initiative distributions in TRAINS91 are not representative of collaborative dialogues. We expect that by taking a sample of dialogues whose task/dialogue initiative distributions are more representative of collaborative dialogues, we will lower the value of P(E), the probability of chance agreement in the kappa statistic, and thus obtain a higher kappa coefficient of agreement. However, we leave selecting and annotating such a subset of representative dialogues for future work.

4. Cues for Shifts in Initiative

Given that initiative shifts between dialogue participants during their interaction, we are interested in finding out the reasons that result in such shifts. Whittaker, Stenton, and Walker (Whittaker and Stenton, 1988; Walker and Whittaker, 1990) have previously identified a set of utterance intentions that serve as cues to indicate shift or lack of shift in initiative, such as prompts and questions. We examined our annotated TRAINS91 corpus and identified

[5] It is an open question as to whether or not this metric (and the kappa statistic in general) is appropriate for measuring reliability in annotating dialogue features. However, short of a better alternative, we adopt this metric in our current discussion.

Table II. Cues for Modeling Initiative Shifts

Class	Cue Type	Subtype	Effect	Initiative
Explicit	Explicit requests	give up task	both	hearer
		give up dialogue	DI	hearer
		take over task	both	speaker
		take over dialogue	DI	speaker
Discourse	End silence		both	hearer
	No new info	repetitions	both	hearer
		prompts	both	hearer
	Questions	domain	DI	speaker
		evaluation	DI	hearer
	Obligation fulfilled	task	both	hearer
		discourse	DI	hearer
Analytical	Invalidity	action	both	hearer
		belief	DI	hearer
	Suboptimality		both	hearer
	Ambiguity	action	both	hearer
		belief	DI	hearer

additional cues that may have contributed to the shift or lack of shift in task and dialogue initiatives. This resulted in eight cue types, which are grouped into three classes, based on the type of knowledge needed to recognize each cue. Table II shows the three classes, the eight cue types, their subtypes if any, whether a cue may affect merely the dialogue initiative (DI) or both the task and dialogue initiatives, and the agent expected to hold the initiative in the next turn.

A cue may affect both the task and dialogue initiatives if the cue may reasonably suggest that the hearer direct how the agents' task will be accomplished in the next dialogue turn. Examples of such cues include cases in which the speaker explicitly asks the hearer to direct the task, and those where the speaker proposes an invalid action, thereby potentially leading the hearer to correct the invalid action and proposing an alternative valid action. On the other hand, a cue only affects the dialogue initiative if the cue suggests that the hearer take the conversational lead in order to establish mutual beliefs between the agents, e.g., when the speaker attempts to verify a piece of domain knowledge.

4.1. Explicit Cues

The first cue class, *explicit cues*, includes explicit requests by the speaker to give up or take over the initiative. Explicit cues may result in shifts in both types of initiatives. For instance, consider the following dialogue segment (where the cue is highlighted in boldface):

(4) U: Yeah, so go to Bath and pick up the boxcar, bring it back to Corning and then bring it back to Elmira.

(5) S: Okay, well that's 8 hours, so you're not gaining anything by doing that.

(6) U: Okay [2sec] [sigh] [3sec] **Any suggestions***?*

(7) S: Well, there's a boxcar at Dansville and you can use that.

In utterance (4), U has both the task and dialogue initiatives, since U is explicitly proposing actions for accomplishing the agents' goal. S takes over the dialogue initiative in utterance (5) to point out the invalidity of the proposal. The explicit cue to give up task initiative, *any suggestions*, in utterance (6) suggests that S should have both the task and dialogue initiatives in the next dialogue turn, as in (7) where S proposes a (partial) solution to U's problem.

Instead of affecting both the task and dialogue initiatives, an explicit cue may be intended to affect merely the dialogue initiative, as in the following example:

(8) U: So you can start making OJ and then when the OJ is ready you load it up into the tanker car and bring it back to Avon.

(9) Okay, **summarize the plan at this point** *system.*

(10) S: Okay, lemme make sure I got all this. You wanna link the boxcar at Elmira to E2 ...

In utterance (8), U has both the task and dialogue initiatives, and in (9), U employs the give-up-dialogue cue to explicitly hand the dialogue initiative over to S by asking S to summarize the current plan, as realized in (10). Note that these cues merely indicate the speaker's intention regarding task/dialogue initiative shifts for the next dialogue turn, and it is up to the hearer to determine whether or not such shifts actually occur. For instance, instead of utterance (7), S may respond to U's cue *"any suggestions"* by saying *"it's your responsibility to make a proposal"*, i.e., responding to U's question *without* taking over the task initiative.

4.2. Discourse Cues

The second cue class, *discourse cues*, includes cues that can be recognized using linguistic information, such as the surface form of an utterance, and the recognized intentions of the utterances, such as the how the current utterance relates to prior discourse.

We have identified four types of discourse cues. The first type is perceptible silence, or pauses, observed at the end of an utterance, which has been found to correlate with discourse boundaries (Grosz and Hirschberg, 1992; Passonneau and Litman, 1993; Swerts, 1997). We believe that in the context of initiative modeling, silence at the end of an utterance may suggest that the speaker has nothing more to say in the current turn and intends to give up his task/dialogue initiative. For instance, in the following dialogue segment, the silence at the end of U's utterance led S to take over the dialogue initiative and provide what she believed to be the most relevant information at that time, even though it was not explicitly requested.

(11) U: Can we please send engine E1 over to Dansville to pick up a boxcar and then send it right back to Avon. **[3 sec]**

(12) S: Okay, it'll get back to Avon at 6.

The second type of discourse cues includes situations in which the speaker's utterances do not contribute new information that has not been conveyed earlier in the dialogue. These utterances are further classified into two groups: *repetitions*, a subset of the *informationally redundant utterances* (Walker, 1992), in which the speaker paraphrases an utterance by the hearer or repeats the utterance verbatim, and *prompts*, such as *"yeah"* and *"okay"*, where the speaker merely acknowledges the hearer's previous utterances. Repetitions and prompts also suggest that the speaker has nothing more to say and are indications that the hearer should take over the task/dialogue initiative (Whittaker and Stenton, 1988). In the following dialogue segment, utterance (14) is an informationally redundant utterance since it merely paraphrases part of utterance (13):

(13) U: Grab the tanker, pick up oranges, go to Elmira, make em into orange juice.

(14) S: Okay, then **we go to Elmira, we make orange juice***, okay.*

(15) U: And then send the orange juice back to Avon.

The third type of discourse cues includes questions which, based on anticipated responses, are divided into *domain* questions and *evaluation* questions. Domain questions are questions in which the speaker intends to obtain or verify a piece of domain knowledge. They usually merely require a direct response and thus typically do not result in a shift in dialogue initiative. Evaluation questions, on the other hand, are questions in which the speaker intends to assess the quality of a proposed plan. They often require an analysis of the proposal, and thus frequently result in a shift in dialogue initiative. In the following dialogue segment, U asks S to evaluate the feasibility of a proposed plan in utterance (16), resulting in S taking over the dialogue initiative in (17).[6] On the other hand, the domain question in utterance (18) required only a direct response, and thus did not result in an initiative shift in (19).

(16) U: **Could it take the other boxcar back to Avon to fill up with bananas?**

(17) S: Okay, we could get back to Avon by noon and that wouldn't leave enough time to get to Corning by 3.

(18) U: Right. Okay, so **how long again between Avon and Bath?**

(19) S: That's 4 hours.

The final type of discourse cue includes utterances that satisfy an outstanding task or discourse obligation. Such obligations may have resulted from a prior request by the hearer, or from an interruption initiated by the speaker himself. In either case, when the task/dialogue obligation is fulfilled, the initiative may be reverted back to the hearer who held the initiative prior to the request or interruption. As discussed in the previous section, utterance (20) suggests that S take over both the task and dialogue initiatives in the next turn. This also creates a task (and thus discourse) obligation for S to satisfy the request for suggestions by U. The utterance in boldface in (21) satisfies this outstanding obligation, thus signals that the task and dialogue initiatives should return to U, who initiated the request, as in utterance (22). Utterance (22) is an evaluation question suggesting that S take over the dialogue initiative, and at the same time creating a discourse obligation for S to provide an evaluation of the proposal. Utterance (23) satisfies this outstanding discourse obligation, signaling that the dialogue initiative be returned to U. Thus in (24), U has both the task and dialogue initiatives, and continues the planning process.

(20) U: Any suggestions?

(21) S: Well, **there's a boxcar at Dansville and you could use that***, but you'd have to change your banana plan.*

(22) U: Will the banana plan work if we get the boxcar at Bath instead of Dansville?

[6] In addition to being an evaluation question, utterance (16) also triggers the analytical cue *invalidity-action*, which will be discussed in Section 4.3.

(23) S: If you do that, the bananas will get to Corning at 3PM exactly, so **that will work**.

(24) U: Now about the oranges ...

4.3. ANALYTICAL CUES

The third class of cues, *analytical cues*, includes cues that cannot be recognized without the hearer performing an evaluation of the speaker's proposal using the hearer's private knowledge (Chu-Carroll and Carberry, 1994; Chu-Carroll and Carberry, 1995). After the evaluation, the hearer may find the proposal *invalid*, *suboptimal*, or *ambiguous*. When such a problem is detected with respect to the speaker's proposal, the hearer may initiate a subdialogue to resolve the problem, resulting in a shift in task/dialogue initiatives.[7] In the following dialogue segment, U evaluates S's proposal in (25) and believes the action to be invalid. Utterances (26a) and (26b) illustrate two alternative ways that S may correct this invalid action. In (26a), S does so by taking over both the task and dialogue initiatives to propose an action that will cause the original proposal to be valid, while in (26b), S takes over only the dialogue initiative to point out the invalidity of the proposal.

(25) U: **Let's get the tanker car to Elmira and fill it with OJ.**

(26a) S: You need to get oranges to the OJ factory.

(26b) S: You don't have OJ in Elmira.

When the evaluation of a proposal results in the hearer believing that a proposed belief is invalid, the potential effect of the detected cue *invalidity-belief* is for the hearer to take over the dialogue initiative (but not the task initiative). For example, in the following dialogue segment, S responds to U's proposed invalid belief in utterance (27) by taking over the dialogue initiative to correct the belief in (28).

(27) U: **It's shorter to Bath from Avon.**

(28) S: It's shorter to Dansville. The map is slightly misleading.

[7] Whittaker, Stenton, and Walker (Whittaker and Stenton, 1988; Walker and Whittaker, 1990) treat subdialogues initiated as a result of these cues as interruptions, motivated by their collaborative planning principle.

The cues *ambiguity-action* and *ambiguity-belief* are triggered when the hearer cannot unambiguously interpret the speaker's utterances based on her knowledge about the domain and about the world. Their effects are similar to those of *invalidity-action* and *invalidity-belief* described above. Finally, a proposal is considered *suboptimal* if the hearer believes that there exists a better way to achieve their goal than that proposed by the speaker. This would naturally lead the hearer to take over both the task and dialogue initiatives to suggest the alternative to the speaker. For instance, in the following dialogue segment, S takes over the task and dialogue initiatives in utterance (32) to suggest an alternative to U's proposal in utterance (31):

(29) U: I'm looking for Saudi Arabian Airlines on the night of the eleventh.

(30) S: Right, it's sold out.

(31) U: Is Friday open?

(32) S: Let me check here. **I'm showing economy on Pan Am is open on the eleventh.**

5. An Evidential Model for Tracking Initiative

As illustrated by our corpus analysis in Section 3.2, during the course of a dialogue, the task and dialogue initiatives switch back and forth between the dialogue participants. Our goal is to develop a model for tracking such shifts in initiative, using cues identified in the previous section, in order to allow a dialogue system to better manage its interaction with the user. Although having the task initiative means that an agent has the lead in developing the agents' plan, it does not give this agent sole control over determining the content of this plan. Instead, it merely indicates that the agent has a higher level of involvement in directing the task planning process than the other agent at the current point in the dialogue. Thus, in our model for tracking initiative, we associate with each agent a *task initiative index* and a *dialogue initiative index*, which measure the agent's levels of involvement in directing the planning process and in determining the discourse focus (and thus the likelihood that the agent holds the task initiative and dialogue initiative), respectively. In addition, we represent the effect of each cue identified in Section 4 in terms of how it affects the existing task/dialogue initiative indices. Then at the end of each dialogue turn, new initiative indices are computed based on the effects that the cues observed during that turn have on changing the current initiative indices. These new initiative indices are then used to determine the task and dialogue initiative holders for the next dialogue turn (Chu-Carroll and Brown, 1997b).

Evidently, some cues provide stronger evidence for a shift in initiative than others. For instance, an explicit request to give up initiative is a stronger cue for initiative shift than, say, silence at the end of an utterance. Thus, in developing a framework for tracking initiative, we need an underlying model that allows us to represent the amount of evidence a cue provides for a shift or lack of shift in initiative. Furthermore, since multiple cues may be observed during a dialogue turn, this model must also provide a mechanism for combining the effects of multiple cues to determine their overall effect on initiative shift. We adopt the Dempster-Shafer theory of evidence (Shafer, 1976; Gordon and Shortliffe, 1984) as such an underlying model. The reasons for this are threefold. First, the Dempster-Shafer theory allows us to use basic probability assignments to represent the effect of each cue on initiative shift, and the Dempster's combination rule allows us to easily compute a new basic probability assignments from two existing ones, i.e., to determine the combined effect of two observed cues. Second, the Dempster-Shafer theory, unlike the Bayesian model, does not require a complete set of *a priori* and conditional probabilities, which is difficult to obtain for sparse pieces of evidence. Third, the Dempster-Shafer theory distinguishes between situations in which no evidence is available to support any conclusion and those in which equal evidence is available to support each conclusion. Thus the outcome of the Dempster-Shafer model more accurately represents the *amount* of evidence available to support a particular conclusion, i.e., the *provability* of a particular conclusion (Pearl, 1990).

In the next section, we give a brief introduction to the Dempster-Shafer theory of evidence. In Section 5.2, we show how the problem of tracking initiative can be framed to utilize the Dempster-Shafer theory, discuss our training algorithm for determining the basic probability assignment used to represent the effect of a cue, and discuss the performance of this model in predicting task/dialogue initiative holders. Finally, in Section 5.3, we perform an analysis of the erroneous predictions resulting from our experiments in Section 5.2.

5.1. The Dempster-Shafer Theory of Evidence

The Dempster-Shafer Theory is a mathematical theory for reasoning under uncertainty (Shafer, 1976; Gordon and Shortliffe, 1984). It operates over a set of possible outcomes, called the *frame of discernment*, Θ. The elements in Θ are assumed to be mutually exclusive and exhaustive. Associated with each piece of evidence that may provide support for the possible outcomes is a *basic probability assignment (bpa)*. A bpa is a function that represents the impact of a piece of evidence on the subsets of Θ. It assigns a number

Table III. Intersection Tableau for Computing $m_{t1} \oplus m_{t2}$

	{hearer} (.5)	{speaker} (.3)	Θ (.2)
{hearer} (.2)	{hearer} (.1)	$\emptyset$ (.06)	{hearer} (.04)
Θ (.8)	{hearer} (.4)	{speaker} (.24)	Θ (.16)

in the range [0,1] to each subset of Θ such that the numbers sum to 1. The number assigned to the subset Θ_1 then denotes the amount of support the piece of evidence *directly* provides for the set of conclusions represented by Θ_1. For instance, suppose two bpa's, m_{t1} and m_{t2}, representing the amount of evidence that two hypothetical cues, c_1 and c_2, provide for determining whether or not a task initiative shift should occur are as follows, where Θ = {speaker,hearer}:

$$m_{t1}(\{hearer\}) = .2 \qquad m_{t1}(\Theta) = .8$$
$$m_{t2}(\{hearer\}) = .5 \qquad m_{t2}(\{speaker\}) = .3 \qquad m_{t2}(\Theta) = .2$$

m_{t1} indicates that observation of the cue c_1 supports an initiative shift to the hearer to the degree .2. The remaining belief, .8, is assigned to Θ, indicating that to the degree .8, observation of this cue does not commit to identifying whether the speaker or the hearer should have the next task initiative.

When multiple pieces of evidence are present, the theory utilizes Dempster's combination rule to compute a new bpa from the individual bpa's to represent the impact of the combined evidence. Dempster's combination rule uses a combination function, $\oplus$, to combine two bpa's, m_1 and m_2. This process involves two steps. First, $m_1 \oplus m_2(Z)$ is computed by summing all products of the form $m_1(X)\, m_2(Y)$, where X and Y run over all subsets of Θ whose intersection is Z. Second, if $m_1 \oplus m_2(\emptyset) \neq 0$, then $m_1 \oplus m_2(\emptyset)$ is assigned a value of 0 and the the other values are normalized accordingly so that they sum to 1. For instance, given m_{t1} and m_{t2} above, to obtain $m_{t1} \oplus m_{t2}$, we first compute an intersection tableau whose first column and first row are the values assigned to m_{t1} and m_{t2}, respectively. Suppose that $m_{t1}(s_i) = v_i$ is in row i and that $m_{t2}(s_j) = v_j$ is in column j. The element in entry i, j in the tableau is then $s_i \cap s_j$, and its value $v_i * v_j$. Table III shows the intersection tableau for $m_{t1} \oplus m_{t2}$.

Since $m_{t1} \oplus m_{t2}(\emptyset) \neq 0$, the resulting bpa needs to be normalized. This normalization procedure is carried out by assigning 0 to $m_{t1} \oplus m_{t2}(\emptyset)$ and dividing all other values of $m_{t1} \oplus m_{t2}$ by 1 - κ, where κ is the sum of all values assigned to $\emptyset$. In this example, 1 - κ is .94; thus the final results of applying Dempster's combination rule are as follows:

$$m_{t1} \oplus m_{t2}(\{hearer\}) = (.1 + .04 + .4)/.94 = .57$$

$m_{t1} \oplus m_{t2}(\{speaker\})$ = .24/.94 = .26
$m_{t1} \oplus m_{t2}(\Theta)$ = .16/.94 = .17

The final result of computing $m_{t1} \oplus m_{t2}$ represents the combined effect of observing both cues c_1 and c_2. It indicates that when both cues occur at the same time, they provide stronger support for the hearer taking over the initiative in the next dialogue turn than either cue alone.

5.2. Utilizing the Dempster-Shafer Theory in Tracking Initiative

As discussed earlier, our model for tracking initiative can be outlined as follows. We maintain, for each agent, a current task initiative index and a current dialogue initiative index to represent the levels of involvement that the agent has in leading the planning process and in determining the discourse focus in the current dialogue turn. At the end of each dialogue turn, new initiative indices are computed based on the current indices and the effects that the observed cues have on changing these values. These new initiative indices then become the current initiative indices for the next dialogue turn and the process repeats.

In order to utilize the Dempster-Shafer theory in this process, we represent the current initiative indices as two bpa's, m_{t-cur} and m_{d-cur}. More specifically, the bpa for representing the current task initiative indices take the form of $m_{t-cur}(\{speaker\})=x$ and $m_{t-cur}(\{hearer\})=1-x$. At the beginning of a dialogue, default bpa's are used based on the collaborative setting of the application domain. For example, in the TRAINS domain, a reasonable default setting for the initial bpa's may be $m_{t-cur}(\{user\})$ =.7; $m_{t-cur}(\{system\})$ =.3, and $m_{d-cur}(\{user\})$ =.6; $m_{d-cur}(\{system\})$ =.4. This is because the system's role is to assist the user in devising a plan to achieve his goal; thus the system should have lower task and dialogue initiative indices than the user. On the other hand, in the maptask domain (Canadian Map Task Dialogues, 1996), a reasonable default setting may be $m_{t-cur}(\{giver\})$ =1; $m_{t-cur}(\{follower\})$ =0, and $m_{d-cur}(\{giver\})$ =.6; $m_{d-cur}(\{follower\})$ =.4. This is because the instruction giver has sole control over the agents' domain actions (to follow the route on his map), but will allow the follower to ask, for instance, clarification questions when necessary.

In addition to representing the current initiative indices as bpa's, we also associate with each cue two bpa's to represent its effect on changing the values of the current task and dialogue initiative indices, respectively. For instance, the bpa that represents the effect of cue_i on changing the current task initiative bpa is $m_{t-i}(\{speaker\})=x$; $m_{t-i}(\{hearer\})=y$; $m_{t-i}(\Theta)=1-x-y$. Recall this indicates that if cue_i is observed during the current turn, then this observation provides evidence for the speaker taking over the next task initiative to the degree x, for the hearer taking over the next task initiative to the degree y, and to the degree $1 - x - y$, the observation of this cue does not commit to identifying whether the speaker or the hearer should hold the next task initiative. Once we have represented the initiative indices and the effects of cues as bpa's, then at the end of each dialogue turn, we can simply invoke Dempster's combination rule to compute two new bpa's, m_{t-new} and m_{d-new}, to represent the initiative indices for the next dialogue turn based on m_{t-cur} and m_{d-cur}, as well as m_{t-i} and m_{d-i} for each cue_i observed during the turn. However, to employ this model in initiative tracking, we need to first determine the appropriate bpa's to represent the effect of each cue. The next section describes a training algorithm for this purpose.

5.2.1. *Determining the Effect of Cues*

In order to identify the effect that each cue has on determining the next task/dialogue initiative holder, we extended our annotation of the TRAINS91 dialogues to include, in addition to the agent(s) holding the task and dialogue initiatives for each turn, a list of cues observed during that turn. We initialize the task and dialogue bpa's for each cue_i to be $m_{t-i}(\Theta)=1$ and $m_{d-i}(\Theta)=1$. In other words, we first assume that a cue will have no effect on determining the new task and dialogue initiative indices. The annotated data are then used to adjust these default bpa's based on whether or not they allow the system to correctly predict the next task and dialogue initiative holders. Figure 1 shows our training algorithm for this adjustment process.

For each turn, the task and dialogue initiative bpa's for each observed cue are used, along with the bpa's representing the current initiative indices (m_{t-cur} and m_{d-cur}), to determine the new initiative indices (step 2). The **combine** function utilizes Dempster's combination rule to combine pairs of bpa's in the given set until a final bpa is obtained to represent the cumulative effect of all given bpa's. The resulting bpa's then represent the new initiative indices, and are used to predict the task/dialogue initiative holders for the next dialogue turn (step 3). If this prediction is confirmed by the actual values in the annotated data, it indicates that the bpa's for the observed cues are appropriate in this instance. On the other hand, if the two values disagree, **Adjust-bpa** is invoked to alter the bpa's for the observed cues, and **Reset-current-bpa** is invoked to adjust the new bpa's to reflect the actual initiative holder (step 4). In our implementation of this algorithm, **Reset-current-bpa** sums the values assigned to $speaker$ and $hearer$ in each of m_{t-new} and m_{d-new}, and assigns .525 of the sum to the actual initiative holder and .475 of the sum to the other agent.[8]

[8] We ran a series of experiments to determine the optimal distribution of this sum by varying the share assigned to the actual initiative holder from .525 to .975 (at .025 intervals). Our experiment showed that values ranging between .525 and .6 yielded optimal results on tracking the distribution of task and dialogue initiatives on an 8-fold cross-validated test on the TRAINS91 corpus.

Train-bpa(annotated-data):

1. $m_{t-cur} \leftarrow$ default task initiative indices
   $m_{d-cur} \leftarrow$ default dialogue initiative indices
   cur-data $\leftarrow$ **read**(annotated-data)
   cue-set $\leftarrow$ cues in cur-data
2. /* *compute new initiative indices* */
   task-bpas $\leftarrow$ task initiative bpa's for cues in cue-set $\cup$ $\{m_{t-cur}\}$
   dialogue-bpas $\leftarrow$ dialogue initiative bpa's for cues in cue-set $\cup$ $\{m_{d-cur}\}$
   $m_{t-new} \leftarrow$ **combine**(task-bpas)
   $m_{d-new} \leftarrow$ **combine**(dialogue-bpas)
3. /* *determine predicted next initiative holders* */
   If $m_{t-new}(\{speaker\}) \geq m_{t-new}(\{hearer\})$, t-predicted $\leftarrow$ speaker
   Else, t-predicted $\leftarrow$ hearer
   If $m_{d-new}(\{speaker\}) \geq m_{d-new}(\{hearer\})$, d-predicted $\leftarrow$ speaker
   Else, d-predicted $\leftarrow$ hearer
4. /* *find actual initiative holders and compare* */
   new-data $\leftarrow$ **read**(annotated-data)
   t-actual $\leftarrow$ actual task initiative holder in new-data
   d-actual $\leftarrow$ actual dialogue initiative holder in new-data
   If t-predicted $\neq$ t-actual,
     Adjust-bpa(cue-set,task)
     Reset-current-bpa(m_{t-cur})
   If d-predicted $\neq$ d-actual,
     Adjust-bpa(cue-set,dialogue)
     Reset-current-bpa(m_{d-cur})
5. If end-of-dialogue, return
   Else, /* *swap roles of speaker and hearer* */
     $m_{t-cur}(\{speaker\}) \leftarrow m_{t-new}(\{hearer\})$
     $m_{d-cur}(\{speaker\}) \leftarrow m_{d-new}(\{hearer\})$
     $m_{t-cur}(\{hearer\}) \leftarrow m_{t-new}(\{speaker\})$
     $m_{d-cur}(\{hearer\}) \leftarrow m_{d-new}(\{speaker\})$
     cue-set $\leftarrow$ cues in new-data
     Goto step 2.

Figure 1. Training Algorithm for Determining BPA's

Adjust-bpa, which is invoked when the system's predicted task or dialogue initiative holder disagrees with the actual initiative holder, adjusts the bpa's for the observed cues in favor of the actual initiative holder. We developed three adjustment methods by varying the effect that a disagreement between the actual and predicted initiative holders will have on changing the relevant bpa's (bpa's for the observed cues). The first is the *constant-increment* method, where each time a disagreement occurs, the value assigned to the actual initiative holder in each relevant bpa is incremented by a constant (Δ), while that for Θ is decremented by Δ. The second method is the *constant-increment-with-counter* method, which associates with each bpa for each cue a counter which is incremented when the use of the bpa resulted in a correct prediction, and decremented when it resulted in an incorrect prediction. If the counter is negative, the *constant-increment* method is invoked and the counter is reset to 0. This method ensures that a bpa will only be adjusted if it has no "credit" for correct prediction in the past. The third method, *variable-increment-with-counter*, is a variation of *constant-increment-with-counter*. However, instead of using the counter to determine whether or not an adjustment is needed, the counter is used to determine the amount to be adjusted. In our implementation of the system, each time a bpa results in the system making an erroneous prediction, the value for the actual initiative holder is incremented by $\Delta/2^{count+1}$, and that for Θ decremented by the same amount. This function is selected to reflect our intention that the adjustment be inversely exponentially related to the value of the counter, i.e., the higher "credit" the bpa has for correct prediction in the past, the less it should be modified for occasional errors.

5.2.2. *Experiments and Results*

In addition to experimenting with different adjustment methods, we also varied the increment constant, Δ. For each adjustment method, we ran 19 training sessions with Δ ranging from .025 to .475, incrementing by .025 between each session. We then evaluated the system's performance based on its accuracy in predicting the task and dialogue initiative holders for each dialogue turn. We divided the TRAINS91 corpus into eight sets based on speaker/hearer pairs. For each Δ, we evaluated the system's performance in predicting the task and dialogue initiative holders using an 8-fold cross-validation. Figures 2a and 2b show our system's performance in predicting the task and dialogue initiative holders, respectively, using the three adjustment methods. These results are compared against the prediction results using a baseline method of predicting initiative holders without the use of cues, i.e., always predict that the current initiative holder will have the initiative in the next dialogue turn.

The results in Figures 2a and 2b show that in the vast majority of cases, our prediction methods yield better results than making predictions without cues (baseline strategy labeled "no cue"). Furthermore, substantial improvement is gained by the use of counters in *constant-increment-with-counter* and *variable-increment-with-counter*. This outcome agrees with our expectation that without counters, the effect of the "exceptions of the rules" may accumulate and result in erroneous predictions, hence the erratic behavior of the *constant-increment* method. With the aid of counters, the *variable-increment-with-counter* method is able to obtain substantially better and more consistent results than the *constant-increment* method. However, even by restricting the

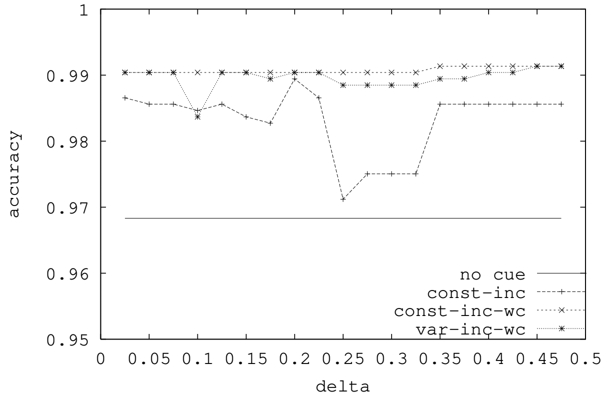

Figure 2a. Task Initiative Prediction Results

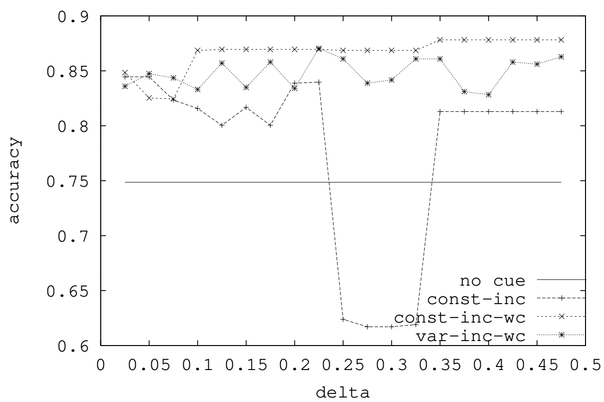

Figure 2b. Dialogue Initiative Prediction Results

increment to be inversely exponentially related to the "credit" the bpa had in making correct predictions, the exceptions of the rules still resulted in undesirable effects, hence the further improved performance by *constant-increment-with-counter*.

As Figures 2a and 2b show, the best prediction results occur using the *constant-increment-with-counter* adjustment method with Δ between .35 and .475. Tables IVa and IVb show the bpa's that result from our training procedure using Δ=.35.[9] These bpa's represent the amount of evidence each cue gives to changing the current task and dialogue initiative indices. For instance,

[9] The cues *Request-take over task* and *Request-take over dialogue* were not used in our experiments since they do not occur in our corpus.

Table IVa. Trained BPA's for Task Initiative Shift

Cue Type	**Subtype**	**Trained BPA**
Explicit request	give up task	$m_{t-gut}(\{hearer\})$ =.35; $m_{t-gut}(\Theta)$ =.65
End silence		$m_{t-es}(\Theta)$ =1
No new info	repetitions	$m_{t-rep}(\Theta)$ =1
	prompts	$m_{t-pro}(\Theta)$ =1
Obligation fulfilled	task	$m_{t-tof}(\{hearer\})$ =.35; $m_{t-tof}(\Theta)$ =.65
Invalidity	action	$m_{t-ia}(\Theta)$ =1
Suboptimality		$m_{t-sub}(\{hearer\})$ =.35; $m_{t-sub}(\Theta)$ =.65
Ambiguity	action	$m_{t-aa}(\{hearer\})$ =.35; $m_{t-aa}(\Theta)$ =.65

Table IVb. Trained BPA's for Dialogue Initiative Shift

Cue Type	**Subtype**	**Trained BPA**
Explicit request	give up task	$m_{d-gut}(\{hearer\})$ =.35; $m_{d-gut}(\Theta)$ =.65
	give up dialogue	$m_{d-gud}(\{hearer\})$ =.35; $m_{d-gud}(\Theta)$ =.65
End silence		$m_{d-es}(\Theta)$ =1
No new info	repetitions	$m_{d-rep}(\Theta)$ =1
	prompts	$m_{d-pro}(\Theta)$ =1
Questions	domain	$m_{d-dq}(\{speaker\})$ =.35; $m_{d-dq}(\{hearer\})$ =.35; $m_{d-dq}(\Theta)$ =.3
	evaluation	$m_{d-eq}(\{hearer\})$ =.35; $m_{d-eq}(\Theta)$ =.65
Obligation fulfilled	task	$m_{d-tof}(\{hearer\})$ =.35; $m_{d-tof}(\Theta)$ =.65
	discourse	$m_{d-dof}(\{hearer\})$ =.35; $m_{d-dof}(\Theta)$ =.65
Invalidity	action	$m_{d-ia}(\{hearer\})$ =.7; $m_{d-ia}(\Theta)$ =.3
	belief	$m_{d-ib}(\{hearer\})$ =.35; $m_{d-ib}(\Theta)$ =.65
Suboptimality		$m_{d-sub}(\{hearer\})$ =.35; $m_{d-sub}(\Theta)$ =.65
Ambiguity	action	$m_{d-aa}(\{hearer\})$ =.7; $m_{d-aa}(\Theta)$ =.3
	belief	$m_{d-ab}(\{hearer\})$ =.35; $m_{d-ab}(\Theta)$ =.65

the cue *suboptimality* provides support to the degree .35 that both the task and dialogue initiatives will shift to the hearer in the next turn.

The trained bpa's in Tables IVa and IVb show that explicit requests to give up task or dialogue initiative provide a moderate amount of evidence for an actual task/dialogue initiative shift to the hearer in the next dialogue turn. The degree of support, however, is not as strong as one may expect for an explicit cue. We believe that this may be the result of the the sparse data problem,[10] since our training algorithm only adjusts the existing bpa for a cue if 1) the

[10] There were only 2 explicit requests to give up task initiative and 4 explicit requests to give up dialogue initiative in the entire corpus.

Table Va. Task Initiative Prediction Errors

Cue Type	**Subtype**	**Shift**		**No-Shift**	
		error	total	error	total
Invalidity	action	2	2	0	13
Suboptimality		1	1	0	0
Ambiguity	action	3	7	1	5

cue is actually observed, and 2) the existing bpa incorrectly predicts the next initiative holder.

The trained bpa's for discourse cues indicate that the cues *end silence*, *repetitions*, and *prompts* have no effect on changing either the task or dialogue initiative index. An analysis of these cues in our corpus shows that they often occur in the user's turn and that the system often responds to them with an acknowledgment, such as *yeah* or *uh-huh*. We believe this behavior may be partially due to the role that each agent plays in the TRAINS domain, where the user is expected to make a plan and the system is expected to provide assistance in the process. As a result, the system may be more reluctant to take over the initiative when subtle hints are given, such as the discourse cues under discussion. Our results further show that *domain questions* provide an equal degree of support for a dialogue initiative shift to the hearer and for the dialogue initiative to remain with the speaker, while the remaining discourse cues provide a moderate degree of support for a task/dialogue initiative shift to the hearer.

Finally, Table IVa show that when the speaker proposes an *ambiguous* or *suboptimal* action, the cue provides a moderate degree of support for the hearer to take over the task initiative. However, the speaker's proposal of an *invalid action* has no effect on task initiative shift. We again believe that this may be due to the agents' roles in the TRAINS domain, causing the system to leave it up to the user to propose a valid alternative to an invalid action. On the other hand, Table IVb show that *invalidity-action* and *ambiguity-action* strongly support the hearer taking over the dialogue initiative, while the remaining analytical cues provide a moderate degree of support for the hearer to take over the dialogue initiative.

5.3. Error Analysis and Discussion

We analyzed the cases in which the system, using the *constant-increment-with-counter* method with $\Delta = .35$, made erroneous predictions. Tables Va and Vb, which summarize the results of our analysis, show the task and dialogue initiative prediction errors grouped according to the cue classification in Table II. For each cue type, we grouped the errors based on whether or not a

Table Vb. Dialogue Initiative Prediction Errors

Cue Type	**Subtype**	**Shift**		**No-Shift**	
		error	total	error	total
End silence		13	41	0	53
No new info	prompts	1	6	7	193
Questions	domain	13	31	0	98
	evaluation	8	28	5	7
Obligation fulfilled	discourse	12	198	1	5
Invalidity	action	5	15	0	0
	belief	6	19	0	0
Suboptimality		1	1	0	0
Ambiguity	action	6	12	0	0
	belief	3	12	0	0

shift occurred in the actual dialogue. For instance, the first row in Table Va shows that when the cue *invalidity-action* is detected, the system failed to predict both cases where a task initiative shift occurred. On the other hand, in all thirteen cases where the cue did not result in a shift in task initiative, the system made correct predictions. Table Va also shows that when an analytical cue is detected, the system correctly predicted all but one case in which there was no shift in task initiative. However, 60% of the time when an analytical cue is detected, the system failed to predict a shift in task initiative.[11] Similarly, Table Vb shows that 43% of the time when an analytical cue at the task level is observed (*invalidity-action*, *suboptimality*, and *ambiguity-action*), the system fails to predict a shift in dialogue initiative, while 29% of the time when an analytical cue at the dialogue level is observed (*invalidity-belief* and *ambiguity-belief*), the system fails to predict a shift in dialogue initiative. This suggests that while including these analytical cues improves the system's performance in tracking both task and dialogue initiatives, perhaps other cues need to be identified in order to more accurately model initiative shifts under these circumstances. We leave identifying and incorporating these additional cues for future work.

Table Vb shows that when a perceptible silence is detected at the end of an utterance, when the speaker utters a prompt, or when the speaker fulfills an outstanding discourse obligation, the system is able to correctly predict the next dialogue initiative holder in the vast majority of cases.[12] However, for

[11] In the case of suboptimal actions, we encounter the sparse data problem. Since there is only one instance of the cue in the set of dialogues, when the cue is present in the testing set, it is absent from the training set.

[12] The reason that the system may predict a shift when *end silence* and *prompts* are observed, even though their bpa's (Table IVb) indicate that such cues have no effect on dialogue initiative

the cue class *questions*, when the actual initiative shift differs from the norm, i.e., speaker retaining initiative for evaluation questions and hearer taking over initiative for domain questions, the system's performance worsens. Our analysis shows that in the case of domain questions, the erroneous predictions can be grouped into two classes. The first class involves situations in which the response to the question requires more reasoning than that typically expected of a domain question, causing the hearer to take over the dialogue initiative. For instance, in the following dialogue segment, the utterance in boldface answers the question in (33). However, since this answer was not readily available to S, S takes over the dialogue initiative in (34) in order to figure out the answer to U's question.

(33) U: We're picking up the tanker, it needs to then go back to Elmira and have those oranges immediately processed into OJ, filling the tanker and zip that off to Avon. How long will that take?

(34) S: Okay, so we get back to Corning, then we have to take the long route to Avon since the other engine is on the track going the other way, and **that'll get us to Avon at 2pm**.

The second class involves situations in which the hearer, in addition to providing a response to the speaker's question, offers information that she believes is relevant or helpful to accomplishing the agents' task. In the following dialogue segment, simply *"no"*, or *"no, there aren't"* is sufficient to answer U's question in (35). However, instead of merely answering the question, S provided extra information that he believes to be helpful to U in his response.

(35) U: Do you know if there are oranges at the orange juice factory?

(36) S: Uh **no***, the only oranges are in the warehouse.*

In the case of evaluation questions, the erroneous predictions can again be grouped into two classes. The first class involves situations in which the result of the evaluation is readily available to the hearer; thus there is no need for the hearer to take over the dialogue initiative in answering the question. For instance, in utterance (39) of the following dialogue segment, U asks S to evaluate a plan for shipping the bananas. However, since the agents have just reviewed the plan in (37) and (38) and utterances prior to them, the answer to U's question is readily available; thus S does not take over the dialogue initiative to review the proposed plan as in most other evaluation questions.

(37) U: So that would be 9am, and then if we take it by the top route that would be another 4 hours.

shifts, is because these cues often co-occur with other cues which affect dialogue initiative indices.

(38) S: So it gets there at 1pm.

(39) U: Okay, so that's a good enough plan for the bananas, right?

(40) S: Right.

The second class of errors again includes situations in which extra information is provided by the hearer, similar to the case in utterances (35) and (36) where the hearer provided extra information in response to a domain question. Based on this analysis, we believe that although it is difficult to predict when an agent may decide to provide extra information in responding to a question, taking into account the cognitive load that a question places on the hearer may allow us to more accurately predict dialogue initiative shifts. Furthermore, our observation suggests that it may be desirable for the system to take into account, in addition to the existing cues, the possibility of providing additional helpful information to user questions when determining the next task/dialogue initiative holders, if such information exists.

6. Generality of the Model

In the previous section, we discussed the performance of our system cross-validated using an annotated corpus of TRAINS91 dialogues. One interesting question to ask now is: is dialogue participants' behavior in initiative shifts particular to each domain, or is it a higher-level phenomenon that describes how conversants interact with each other in general, regardless of the application domain? To answer this question, we investigated the generality of our system by training it on the TRAINS91 dialogues, and testing it on dialogues from four other corpora.

Using the set of bpa's in Tables IVa and IVb, we evaluated the system on subsets of dialogues taken from the following four corpora: the TRAINS93 dialogues (Heeman and Allen, 1995), airline reservation dialogues (SRI Transcripts, 1992), instruction-giving dialogues (Canadian Map Task Dialogues, 1996), and non-task-oriented dialogues (Switchboard Credit Card Corpus, 1992). In addition, we applied our baseline strategy which makes predictions without the use of cues to each corpus.

Table VI shows comparisons of the features of each of the five dialogue corpora and of the system's performance on these dialogues. The first row in Table VI shows the number of turns where the *expert*[13] holds the task/dialogue initiative in each corpus, with percentages shown in parentheses in row 2. This analysis shows that the distribution of task and dialogue initiatives varies quite

[13] The *expert* is determined as follows: in the TRAINS domain, the system; in the airline reservation domain, the travel agent; in the maptask domain, the instruction giver; and in the switchboard dialogues, the agent who holds the dialogue initiative the majority of the time.

Table VI. System Performance Across Different Application Domains

Corpus (# turns)	TRAINS91 (1042)		TRAINS93 (256)		Airline (332)		Maptask (320)		Switchboard (282)	
	task	dialogue	task	dialogue	task	dialogue	task	dialogue	task	dialogue
Expert control	41 (3.9%)	311 (29.8%)	37 (14.4%)	101 (39.5%)	194 (58.4%)	193 (58.1%)	320 (100%)	277 (86.6%)	N/A	166 (59.9%)
No cue	1009 (96.8%)	780 (74.9%)	239 (93.3%)	189 (73.8%)	308 (92.8%)	247 (74.4%)	320 (100%)	270 (84.4%)	N/A	193 (68.4%)
const-inc-w-count	1033 (99.1%)	915 (87.8%)	250 (97.7%)	217 (84.8%)	316 (95.2%)	281 (84.6%)	320 (100%)	297 (92.8%)	N/A	216 (76.6%)
Absolute Improvement	2.3%	12.9%	4.4%	11.0%	2.4%	10.2%	0.0%	8.4%	N/A	8.2%
Error Reduction	71.9%	51.4%	65.7%	42.0%	33.3%	39.8%	N/A	53.8%	N/A	25.9%

significantly across different corpora, with the distribution biased toward one agent in the TRAINS and maptask corpora, and split relatively evenly in the airline and switchboard dialogues. The third row in the table shows the results of applying our baseline prediction method to each corpus. The numbers shown are correct predictions in each instance, with the corresponding percentages shown in parentheses in row 4. These results indicate the difficulty of the prediction problem in each corpus that the task/dialogue initiative distribution (rows 1 and 2) fails to convey. For instance, although the dialogue initiative is distributed approximately 30/70% between the two agents in the TRAINS91 dialogues and 40/60% in the airline reservation dialogues, row 4 in the table shows that when predicting dialogue initiative holders without the use of cues, the system achieves approximately a 75% correct prediction rate in both domains. This indicates that although the dialogue initiative distribution ratio is different for the TRAINS91 and airline reservation domains, the frequency of dialogue initiative shifts is about the same in these two domains, namely, 25% of all dialogue turns. Row 5 in the table shows the prediction results using the bpa's shown in Tables IVa and IVb. The second to the last row shows the improvement in absolute percentage points between our prediction method and the baseline prediction method, while the last row shows such improvement in terms of error reduction rate. To test the statistical significance between the results obtained by the two prediction methods, for each corpus, we applied Cochran's Q test (Cochran, 1950) to the results in rows 4 and 6. The tests show that for all corpora, the differences between the two algorithms when predicting the task and dialogue initiative holders are statistically significant ($p<.05$ and $p< 10^{-5}$, respectively).

Based on the results of our evaluation, we make the following observations. First, Table VI illustrates the generality of our prediction mechanism. Although the system's performance varies across domains, the use of cues improves the system's accuracies in predicting the task and dialogue initiative holders in all cases.[14] Second, Table VI shows the specificity of the trained bpa's with respect to application domains. When trained on the TRAINS91 dialogues, the system performs similarly well on both the TRAINS91 and TRAINS93 corpora.[15] In terms of improvement in percentage points (second to the last row in table), the system's performances on the collaborative planning dialogues (TRAINS91, TRAINS93, and airline reservation) most closely resemble one another. This suggests that the bpa's may be somewhat sensitive to application environments since they may affect how agents interpret

[14] With the exception of task initiative holder prediction in the maptask dialogues where there is no room for improvement. This is because in the maptask domain, the task specifies that the task initiative remain with one agent, the instruction giver, throughout the dialogue.

[15] Both of the TRAINS corpora contain dialogues on the planning of cargo shipping, with the participants in the TRAINS93 dialogues solving more complicated problems than those in the TRAINS91 dialogues.

cues. Third, our prediction mechanism yields better results on task-oriented dialogues (all but the switchboard dialogues). We believe this is because such dialogues are constrained by the goals; therefore, there are fewer digressions and offers of unsolicited opinion, and thus fewer unsignaled initiative shifts, as compared to the switchboard corpus.

7. Effects of Initiative Tracking on Response Generation

We argued at the beginning of this paper that in order for a dialogue system to interact with its user in a coherent and cooperative fashion, it is necessary for the system to be able to model initiative shifts between the participants during the dialogues. Furthermore, we have discussed an initiative tracking model that determines new initiative indices based on the current initiative indices and the effects that cues observed during the current dialogue turn have on changing the current indices. Since the current initiative indices in turn depend on the initiative indices of the previous dialogue turn and the effects of the cues observed during that turn, this dependency chain backtracks until we reach the default initiative indices used at the beginning at the dialogue. Thus, the initiative indices of the next dialogue turn are affected by three factors: 1) *default initiative indices*, 2) *dialogue history* (cues observed during the course of the dialogue), and 3) *effects of observed cues*. When incorporated into a response generation system, the initiative tracking model can be used to determine the initiative indices for system turns, thus allowing the system to tailor its responses based on the distribution of task and dialogue initiatives. This will then allow a generic response generation system to alter its behavior in different application domains, such as being more active in correcting invalid proposals and suggesting better alternatives in a collaborative planning domain, while being passive when playing the role of the follower in an instruction-giving domain. In the rest of this section, we show how varying the three factors that determine the initiative indices may affect the system responses to user utterances.

7.1. Default Initiative Indices

To illustrate the effect of the default initiative indices on response generation, consider two domains: the TRAINS domain and the airline reservation domain. In the TRAINS domain, the system's task is to assist the user in devising a plan to achieve his goal. A reasonable default initiative distribution for such a task may be $m_{t-cur}(\{U\})$ =.7; $m_{t-cur}(\{S\})$ =.3, and $m_{d-cur}(\{U\})$ =.6; $m_{d-cur}(\{S\})$ =.4. On the other hand, in the airline reservation domain, the system plays a more active role in helping the user make her plan; thus a reasonable default initiative distribution for this task may be

$m_{t-cur}(\{U\})$ =.6; $m_{t-cur}(\{S\})$ =.4, and $m_{d-cur}(\{U\})$ =.6; $m_{d-cur}(\{S\})$ = .4.

Given these default initial initiative indices, now consider the following dialogue segment in the TRAINS domain, where utterance (41) is the first utterance in the dialogue:

(41) U: I need to get oranges to Avon.

(42) S: Okay.

(43) U: Move engine E2 and the boxcar from Elmira to Corning, pick up oranges and send them to Avon.

The initiative indices for utterance (41) are those of the default initiative setting; thus we have,

$$m_{t-(41)}(\{U\}) = .7 \qquad m_{t-(41)}(\{S\}) = .3$$
$$m_{d-(41)}(\{U\}) = .6 \qquad m_{d-(41)}(\{S\}) = .4$$

Since no cue is observed during utterance (41), the initiative indices for utterance (42) remain the same as that for (41). The initiative indices indicate that the system should have neither the task or the dialogue initiative in response to the user's utterance. Thus, under this particular dialogue context, it is reasonable for the system to generate an acknowledgment to the user's proposal, as in (42). Utterance (42) triggers the discourse cue *prompt*. However, our trained bpa's (Tables IVa and IVb) show that *prompts* have no effect on shifts in task or dialogue initiative; thus the initiative indices for utterance (43) remain unchanged. The user's utterance in (43) triggers the cue *ambiguity-action*, since there are two possible routes from Corning to Avon. We have the task initiative indices for utterance (43) and the bpa representing the effect of the observed cue, *ambiguity-action*, on task initiative shift (taken from Table IVa, with *hearer* instantiated as the system) as follows:

$$m_{t-(43)}(\{U\}) = .7 \qquad m_{t-(43)}(\{S\}) = .3$$
$$m_{t-aa}(\{S\}) = .35 \qquad m_{t-aa}(\Theta) = .65$$

Furthermore, we have the dialogue initiative indices for utterance (43) and the bpa representing the effect of the observed cue on dialogue initiative shift as follows:

$$m_{d-(43)}(\{U\}) = .6 \qquad m_{d-(43)}(\{S\}) = .4$$
$$m_{d-aa}(\{S\}) = .7 \qquad m_{d-aa}(\Theta) = .3$$

Tables VIIa and VIIb show the intersection tableaus for computing the new task and dialogue bpa's, respectively. Based on the intersection tableaus, the initiative indices for the next dialogue turn can be computed as follows (see Section 5.1):

Table VIIa. Intersection Tableau for $m_{t-(43)} \oplus m_{t-aa}$

	{U} (.7)	{S} (.3)
{S} (.35)	∅ (.245)	{S} (.105)
Θ (.65)	{U} (.455)	{S} (.195)

Table VIIb. Intersection Tableau for $m_{d-(43)} \oplus m_{d-aa}$

	{U} (.6)	{S} (.4)
{S} (.7)	∅ (.42)	{S} (.28)
Θ (.3)	{U} (.18)	{S} (.12)

$$m_{t-(44)}(\{S\}) = (.105 + .195)/(1 - .245) = .4$$
$$m_{t-(44)}(\{U\}) = .455/(1 - .245) = .6$$

$$m_{d-(44)}(\{S\}) = (.28 + .12)/(1 - .42) = .69$$
$$m_{d-(44)}(\{U\}) = .18/(1 - .42) = .31$$

The new initiative indices indicate that the user should have the task initiative while the system should take over the dialogue initiative in the next turn. In the case of an ambiguous proposal, such an initiative distribution may lead the system to point out the ambiguity in the proposal (thus taking over the dialogue initiative), but not actually resolve the ambiguity (thus leaving the task initiative with the user), as follows:

(44) S: Would you like to go from Corning to Avon through Dansville or Bath?

Now consider a similar dialogue segment in the airline reservation domain, where utterance (45) is again the first utterance in the dialogue:

(45) U: I need to go to San Antonio on Friday.

(46) S: Okay.

(47) U: I'd like to leave Newark at around 8am.

The analysis of initiative indices for this dialogue is similar to that for the previous dialogue, where the initiative indices for utterance (47) are the same as the initial default indices:

$m_{t-(47)}(\{U\})$ =.6 $m_{t-(47)}(\{S\})$ =.4
$m_{d-(47)}(\{U\})$ =.6 $m_{d-(47)}(\{S\})$ =.4

The user's utterance in (47) again triggers the cue *ambiguity-action*, since there are no direct flights from Newark to San Antonio. Thus, computing the new initiative indices based on the initiative indices for utterance (47) and the effect of the observed cue, *ambiguity-action*, we have,[16]

$m_{t-(48)}(\{S\})$ =.51 $m_{t-(48)}(\{U\})$ =.49
$m_{d-(48)}(\{S\})$ =.69 $m_{d-(48)}(\{U\})$ =.31

The new initiative indices indicate that the system should take over both the task and dialogue initiatives in the next turn. Thus, instead of merely pointing out the ambiguity in the proposal and querying the user for disambiguation, as in utterance (44) in the previous example, the system may take a more active role in the planning process as follows:

(48) S: You can either connect through Houston or Dallas. The connection through Houston has a 45-minute layover and through Dallas there's a 75-minute layover. Would you like to connect through Houston?

These two dialogue segments illustrate how the system's response generation behavior can be affected by the initial initiative indices. Thus, one can select the default initiative indices to reflect the degree of task/dialogue initiative the system is expected to have for the particular application domain. As an extreme example, an appropriate bpa representing the default task initiative indices in the maptask domain is $m_{t-cur}(giver)$ =1. In this case, regardless of the observed cues, the instruction giver always has the task initiative when planning her utterances.

7.2. Effects of Cues

As discussed in Section 6, although the trained bpa's performed very well across different domains, they performed better in the domain in which the bpa's were trained. This suggests that perhaps different bpa's are needed to capture the different effects that a cue may have on initiative shifts in different application domains. For example, our trained bpa's in Tables IVa and IVb show that observation of the cue *invalidity-action* in the TRAINS domain has no effect on task initiative shift ($m_{t-ia-trains}(\Theta)$ =1), but strongly suggests a shift in dialogue initiative ($m_{d-ia-trains}(\{hearer\})$ =.7; $m_{d-ia-trains}(\Theta)$ = .3). However, in a domain where the dialogue participants play more equal

[16] The intersection tableaus are similar to those in Tables VIIa and VIIb, and will not be shown in subsequent examples.

roles in the planning process, such as in the airline reservation domain, more appropriate bpa's for this cue may be as follows, where observation of the cue provides a moderate amount of evidence for a shift in task initiative as well:

$$m_{t-ia-airline}(\{hearer\}) = .35 \quad m_{t-ia-airline}(\Theta) = .65$$
$$m_{d-ia-airline}(\{hearer\}) = .7 \quad m_{d-ia-airline}(\Theta) = .3$$

To illustrate how varying the effects of cues may affect a system's response to user utterances, consider the scenario in which the current initiative indices are as follows,

$$m_{t-cur}(\{U\}) = .55 \quad m_{t-cur}(\{S\}) = .45$$
$$m_{d-cur}(\{U\}) = .6 \quad m_{d-cur}(\{S\}) = .4$$

and the user says

(49) U: *We'll make OJ at Elmira, and ship it to Avon through Corning and Dansville.*

Utterance (49) triggers the cue *invalidity-action* since the track through Dansville is unavailable. Based on the current initiative indices and the effect of *invalidity-action* in the TRAINS domain (where *hearer* is instantiated as the system) shown above, the initiative indices for the next dialogue turn is as follows:

$$m_{t-(50)}(\{S\}) = .45 \quad m_{t-(50)}(\{U\}) = .55$$
$$m_{d-(50)}(\{S\}) = .69 \quad m_{d-(50)}(\{U\}) = .31$$

leading the system to take over the dialogue initiative but not the task initiative in its response to the user's proposal. In the case of an invalid proposal, this may lead to a response that points out the invalidity of the proposal but does not suggest a valid alternative, as follows:

(50) S: *We can't go by Dansville because we've got Engine 1 going on that track.*

Now consider a similar dialogue in the airline reservation domain where the user proposes:

(51) U: *I want to book a ticket from Newark to Tokyo through San Francisco on March 1.*

Again, this utterance triggers the cue *invalidity-action*, since the system believes that all flights from San Francisco to Tokyo are sold out for that date. Using the same initiative distribution for the current utterance as that for utterance (49), and the bpa's representing the effect of *invalidity-action* in the airline reservation domain ($m_{t-ia-airline}$ and $m_{d-ia-airline}$), we have the following initiative distribution for the next dialogue turn:

$m_{t-(52)}(\{S\})$ =.56 $m_{t-(52)}(\{U\})$ =.44
$m_{d-(52)}(\{S\})$ =.69 $m_{d-(52)}(\{U\})$ =.31

This will lead the system to take over both the task and dialogue initiatives when responding to the user's utterance as follows:

(52) S: All flights from San Francisco to Tokyo are full, but I do show availability from L.A. Would you like to do that instead?

Undoubtedly, the system could have used the bpa's trained for the TRAINS domain and provided a response similar to that in utterance (50), such as *"You can't go by San Francisco because all flights from San Francisco to Tokyo are sold out."* It is then up to the user to suggest an alternative city for making a connection. Although these two alternative system responses may eventually lead to the same plan for the user, we believe that the response in utterance (52) is more helpful and perhaps more efficient (if the user in fact accepts the system's proposal) in the airline reservation domain. This effect is achieved in our model by using bpa's designed specifically for the application domain.[17] As another example, in the maptask domain, the effect of the cue *request-take over task* can be reduced to nil to reflect the fact that the instruction follower cannot ever take over the task initiative.

7.3. Dialogue History

The effect of dialogue history on initiative shift, and hence response generation, can best be illustrated with examples of dialogue segments from the airline reservation domain where we hypothesize the cue *invalidity-action* to provide a moderate degree of support for a shift in task initiative, as in the previous section. Suppose that in the preceding dialogue the user has stated her goal to travel from Newark to San Francisco then to Seattle, that the current initiative indices are as follows,

$m_{t-cur}(\{U\})$ =.7 $m_{t-cur}(\{S\})$ =.3
$m_{d-cur}(\{U\})$ =.7 $m_{d-cur}(\{S\})$ =.3

and the user says

(53) U: For the Newark to San Francisco leg, I would like United 805 leaving Thursday morning at 9am.

This utterance triggers the cue *invalidity-action* since the system believes that United 805 is sold out. The new initiative indices are

[17] In this example, we hypothesized what we believe to be appropriate bpa's for the airline reservation domain. In the actual development of a system, in order to tailor effects of bpa's to specific domains, we need to annotate a corpus in the new domain and perform the training process described earlier in this paper to obtain the new bpa's.

$m_{t-(54)}(\{S\})$ =.4 $m_{t-(54)}(\{U\})$ =.6
$m_{d-(54)}(\{S\})$ =.59 $m_{d-(54)}(\{U\})$ =.41

suggesting a response in which the system will take over the dialogue initiative but not the task initiative, such as the following:

(54) S: I'm showing that United 805 is sold out.

Now suppose that the user continues with:

(55) U: In that case, I'll go with Continental flight 51, leaving at 9:20am.

Since utterance (54) triggers the cue *obligation fulfilled-dialogue* and utterance (55) again triggers the cue *invalidity-action*, the initiative indices for the next dialogue turn will be:

$m_{t-(56)}(\{S\})$ =.51 $m_{t-(56)}(\{U\})$ =.49
$m_{d-(56)}(\{S\})$ =.75 $m_{d-(56)}(\{U\})$ =.25

Thus, instead of merely pointing out the invalidity of the proposal, as in utterance (54), the system will take over both the task and dialogue initiatives and suggest a valid alternative, such as in the following:

(56) S: That flight is full as well. I'm showing that American Airlines flight 104 is open, and it leaves Newark at 9:30am. Will that be ok?

The extended dialogue in this section illustrates the effect on dialogue history on response generation, i.e., how the cumulative effect of previous cues affects system behavior. Notice that although in both utterances (53) and (55), the cue *invalidity-action* is observed, the system responded to them in different manners, namely, by taking over merely the dialogue initiative in response to the former, while taking over both the task and dialogue initiatives in the latter. This behavior reflects our expectation that in a collaborative planning process, the system may not jump in and take over the planning process at the user's first mistake, but if the system senses that the user is at a lost, e.g., making several mistakes in a row, a cooperative system will offer help by proposing a valid plan.

7.4. Discussion

In the previous sections, we showed how initiative shifts, and hence the system's response generation behavior, may be affected by the default initiative indices, the effects of observed cues, and the discourse history. We have discussed how the default initiative indices and the effects of observed cues can be affected by the application domain and the role the system plays in the domain. Furthermore, it is possible that, during its interactions with users, the system may modify its default initiative indices to more specific indices tailored to specific users, as well as adapt the bpa's that represent the effects of observed cues based on user behavior. Such adaptation will then allow the system to tailor its model of initiative tracking to individual users in future interactions. Thus, our model of initiative tracking allows the system to tailor its responses to user utterances based on the application domain, the system's role in the domain, user characteristics, as well as prior system-user interaction in the dialogue history.

8. Future Work

We are currently investigating other potential cues that may provide additional information for signaling initiative shifts. To this end, we analyzed the coverage of our existing cues. Our analysis shows that the cues shown in Table II covered 92% of the shifts in dialogue initiative in our experiment, leaving 47 unsignaled shifts which can be grouped as follows. First, there are 22 cases in which the hearer, at the end of a dialogue turn, takes over the initiative to provide unsolicited but helpful information. Second, there are 11 cases of interruptions, 7 in which the hearer completes the speaker's utterance and 4 in which the hearer interrupts the speaker and initiates a new topic. Finally, there are 14 cases in which a shift in initiative occurred as a result of a return from a previously unsignaled shift in initiative. We are currently investigating potential cues that may shed some light on the cause of such unsignaled shifts. In particular, we are looking into how prosodic information, such as intonational patterns and final syllable lengthening, may provide information on initiative shifts that our current cues, based solely on linguistic and domain knowledge, fail to convey.

At present, the only evaluation that we have performed on our initiative tracking model is an assessment of how well it can predict the task/dialogue initiative holders in naturally-occurring collaborative dialogues. However, we have not been able to evaluate the effect of this initiative model on the response generation process and the quality of the resultant dialogues. In the longer term, we plan to incorporate this initiative tracking model into the speech-to-speech transaction-based dialogue system we are currently developing. The understanding component of this dialogue system will attempt to automatically identify cues from the user's utterances in order for the system to determine the initiative indices for the next dialogue turn. The initiative indices will then be used by the response generation component to determine an appropriate response to the user's utterance. This dialogue system will then serve as a testbed for varying the system's mixed-initiative behavior in dialogue interaction and for evaluating the cooperativeness and coherence of the dialogues that ensue.

9. Conclusions

In this paper, we presented a model for tracking shifts in initiative between dialogue participants in mixed-initiative dialogue interactions. We showed why it is necessary to distinguish between task and dialogue initiatives, and discussed how this distinction allows us to model phenomena in collaborative dialogues that existing frameworks are unable to explain. We identified eight types of cues that affect shifts in initiative in dialogues, and showed how our evidential model for tracking initiative is able to predict task/dialogue initiative shifts based on the distribution of the current task/dialogue initiative, as well as the effects that observed cues have on changing the current distribution. Our experiments show that by utilizing the *constant-increment-with-counter* adjustment method in determining the basic probability assignments for each cue, the system can correctly predict the task and dialogue initiative holders in 99.1% and 87.8% of the dialogue turns, respectively, in the TRAINS91 corpus, compared to 96.8% and 74.9% without the use of cues. The differences between these results are shown to be statistically significant using Cochran's Q test. In addition, we demonstrated the generality of our model by applying it to dialogues in other collaborative dialogue corpora. The results indicate that although the basic probability assignments may be sensitive to application environments, the use of cues in the prediction process consistently provides substantial improvement in the system's performance. Finally, we showed how the use of an initiative tracking model in a dialogue system will allow the system to tailor its responses to user utterances under different circumstances, based on the application domain, system's role in the domain, dialogue history, and particular user characteristics.

Acknowledgments

We would like to thank Lyn Walker, Diane Litman, Susan Haller, Bob Carpenter, Christer Samuelsson, and the three anonymous reviewers for their comments on earlier drafts of this paper, Bob Carpenter and Christer Samuelsson for participating in the coding reliability test, Jan van Santan and Lyn Walker for discussions on statistical testing methods, as well as Jim Hieronymus and Chilin Shih for discussions on prosody.

References

Allen, J.: 1991, 'Discourse Structure in the TRAINS Project'. In: *Darpa Speech and Natural Language Workshop*. pp. 325–330.

Canadian Map Task Dialogues: 1996, 'Transcripts of DCIEM Sleep Deprivation Study, conducted by Defense and Civil Institute of Environmental Medicine, Canada, and Human Communication Research Centre, University of Edinburgh and University of Glasgow, UK'. Distrubuted by HCRC and LDC.

Carletta, J.: 1996, 'Assessing Agreement on Classification Tasks: The Kappa Statistic'. *Computational Linguistics* **22**, 249–254.

Chu-Carroll, J. and M. K. Brown: 1997a, 'Initiative in Collaborative Interactions — Its Cues and Effects'. In: *Working Notes of the AAAI-97 Spring Symposium on Computational Models for Mixed Initiative Interaction*. pp. 16–22. Also available as AAAI TR SS-97-04.

Chu-Carroll, J. and M. K. Brown: 1997b, 'Tracking Initiative in Collaborative Dialogue Interactions'. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. pp. 262–270.

Chu-Carroll, J. and S. Carberry: 1994, 'A Plan-Based Model for Response Generation in Collaborative Task-Oriented Dialogues'. In: *Proceedings of the Twelfth National Conference on Artificial Intelligence*. pp. 799–805.

Chu-Carroll, J. and S. Carberry: 1995, 'Response Generation in Collaborative Negotiation'. In: *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*. pp. 136–143.

Cochran, W. G.: 1950, 'The Comparison of Percentages in Matched Samples'. *Biometrika* **37**, 256–266.

Cohen, R., C. Allaby, C. Cumbaa, M. Fitzgerald, K. Ho, B. Hui, C. Latulipe, F. Lu, N. Moussa, D. Pooley, A. Qian, and S. Siddiqi: 1998, 'What is Initiative?'. In this issue.

Gordon, J. and E. H. Shortliffe: 1984, 'The Dempster-Shafer Theory of Evidence'. In: B. Buchanan and E. Shortliffe (eds.): *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Chapt. 13, pp. 272–292.

Gross, D., J. F. Allen, and D. R. Traum: 1993, 'The TRAINS 91 Dialogues'. Technical Report TN92-1, Department of Computer Science, University of Rochester.

Grosz, B. and J. Hirschberg: 1992, 'Some Intonational Characteristics of Discourse Structure'. In: *Proceedings of the International Conference on Spoken Language Processing*. pp. 429–432.

Grove, W. M., N. C. Andreasen, P. McDonald-Scott, M. B. Keller, and R. W. Shapiro: 1981, 'Reliability Studies of Psychiatric Diagnosis'. *Archives of General Psychiatry* **38**, 408–413.

Guinn, C. I.: 1996, 'An Analysis of Initiative Selection in Collaborative Task-Oriented Discourse'. In: *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*. pp. 278–285.

Guinn, C. I.: 1998, 'Principles of Mixed-Initiative Human-Computer Collaborative Discourse'. In this issue.

Heeman, P. A. and J. F. Allen: 1995, 'The TRAINS 93 Dialogues'. Technical Report TN94-2, Department of Computer Science, University of Rochester.

Jordan, P. W. and B. Di Eugenio: 1997, 'Control and Initiative in Collaborative Problem Solving Dialogues'. In: *Working Notes of the AAAI-97 Spring Symposium on Computational Models for Mixed Initiative Interaction*. pp. 81–84. Also available as AAAI TR SS-97-04.

Kitano, H. and C. Van Ess-Dykema: 1991, 'Toward a Plan-Based Understanding Model for Mixed-Initiative Dialogues'. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. pp. 25–32.

Lambert, L. and S. Carberry: 1991, 'A Tripartite Plan-based Model of Dialogue'. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. pp. 47–54.

Lester, J. C., B. A. Stone, and G. D. Stelling: 1998, 'Lifelike Pedagogical Agents for Mixed-Initiative Problem Solving in Constructivist Learning Environments'. In this issue.

Litman, D. and J. Allen: 1987, 'A Plan Recognition Model for Subdialogues in Conversation'. *Cognitive Science* **11**, 163–200.
Novick, D. G.: 1988, 'Control of Mixed-Initiative Discourse Through Meta-Locutionary Acts: A Computational Model'. Ph.D. thesis, University of Oregon.
Novick, D. G. and S. Sutton: 1997, 'What is Mixed-Initiative Interaction?'. In: *Working Notes of the AAAI-97 Spring Symposium on Computational Models for Mixed Initiative Interaction*. pp. 114–116. Also available as AAAI TR SS-97-04.
Passonneau, R. J. and D. J. Litman: 1993, 'Intention-Based Segmentation: Human Reliability and Correlation with Linguistic Cues'. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. pp. 148–155.
Pearl, J.: 1990, 'Bayesian and Belief-Fuctions Formalisms for Evidential Reasoning: A Conceptual Analysis'. In: G. Shafer and J. Pearl (eds.): *Readings in Uncertain Reasoning*. Morgan Kaufmann, pp. 540–574.
Pollack, M. E.: 1990, 'Plans as Complex Mental Attitudes'. In: P. R. Cohen, J. Morgan, and M. E. Pollack (eds.): *Intentions in Communication*. MIT Press, pp. 77–104.
Ramshaw, L. A.: 1991, 'A Three-Level Model for Plan Exploration'. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. pp. 36–46.
Rich, C. and C. L. Sidner: 1998, 'COLLAGEN: A Collaboration Manager for Software Interface Agents'. In this issue.
Shafer, G.: 1976, *A Mathematical Theory of Evidence*. Princeton University Press.
Siegel, S. and N. J. Castellan, Jr.: 1988, *Nonparametric Statistics for the Behavioral Sciences*. McGraw Hill.
Smith, R. W. and D. R. Hipp: 1994, *Spoken Natural Language Dialog Systems — A Practical Approach*. Oxford University Press.
SRI Transcripts: 1992, 'Transcripts derived from audiotape conversations made at SRI International, Menlo Park, CA'. Prepared by Jacqueline Kowtko under the direction of Patti Price.
Swerts, M.: 1997, 'Prosodic Features at Discourse Boundaries of Different Strength'. *Journal of the Acoustic Society of America* **101**(1), 514–521.
Switchboard Credit Card Corpus: 1992, 'Transcripts of telephone conversations on the topic of credit card use, collected at Texas Instruments'. Produced by NIST, available through LDC.
van Beek, P. G.: 1987, 'A model for generating better explanations'. In: *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*. Stanford, CA, pp. 215–220.
Walker, M. and S. Whittaker: 1990, 'Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation'. In: *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*. pp. 70–78.
Walker, M. A.: 1992, 'Redundancy in Collaborative Dialogue'. In: *Proceedings of the 15th International Conference on Computational Linguistics*. pp. 345–351.
Whittaker, S. and P. Stenton: 1988, 'Cues and Control in Expert-Client Dialogues'. In: *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*. pp. 123–130.

Vitae

Jennifer Chu-Carroll

Dr. Jennifer Chu-Carroll is a Member of Technical Staff in the Dialogue Systems Research Department at Bell Laboratories, Lucent Technologies. She received her M. Math degree in Computer Science from the University of Waterloo and her Ph.D. degree in Computer and Information Sciences from the University of Delaware. Her research interests lie in the areas of natural language processing, spoken language dialogue systems, and user modeling.

Michael K. Brown

Dr. Brown is a Member of Technical Staff at Bell Labs and Senior Member of the IEEE. He received his MSEE from the University of Michigan, producing the University's first Master's Thesis, working on control systems for ink jet printing. His PhD degree also came from the University of Michigan, Ann Arbor, working on Handwriting Recognition. Dr. Brown has worked extensively in speech understanding, robotics, and machine intelligence. He has over 50 publications and more than a dozen patents.