Models for Multiparty Engagement in Open-World Dialog

Dan Bohus
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
dbohus@microsoft.com

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
horvitz@microsoft.com

Abstract

We present computational models that allow spoken dialog systems to handle multi-participant engagement in open, dynamic environments, where multiple people may enter and leave conversations, and interact with the system and with others in a natural manner. The models for managing the engagement process include components for (1) sensing the engagement state, actions and intentions of multiple agents in the scene, (2) making engagement decisions (*i.e.* whom to engage with, and when) and (3) rendering these decisions in a set of coordinated low-level behaviors in an embodied conversational agent. We review results from a study of interactions "in the wild" with a system that implements such a model.

1 Introduction

To date, nearly all spoken dialog systems research has focused on the challenge of engaging single users on tasks defined within a relatively narrow context. Efforts in this realm have led to significant progress including large-scale deployments that now make spoken dialog systems common features in the daily lives of millions of people. However, research on dialog systems has largely overlooked important challenges with the initiation, maintenance, and suspension of conversations that are common in the course of natural communication and collaborations among people. In (Bohus and Horvitz, 2009) we outlined a set of core challenges for extending traditional *closed-world* dialog systems to systems that have competency in *open-world dialog*. The work described here is part of a larger research effort aimed at addressing these challenges, and constructing computational models to support the core interaction skills required for open-world dialog. In particular, we focus our attention in this paper on the challenges of managing *engagement* – "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake", *cf.* Sidner et al. (2004) in open-world settings.

We begin by reviewing the challenges of managing engagement in the open-world in the next section. In Section 3, we survey the terrain of related efforts that provides valuable context for the new work described in this paper. In Section 4, we introduce a computational model for multiparty situated engagement. The model harnesses components for sensing the engagement state, actions, and intentions of people in the scene for making high-level engagement decisions (whom to engage with, and when), and for rendering these decisions into a set of low-level coordinated behaviors (*e.g.*, gestures, eye gaze, greetings, etc.). Then, we describe an initial observational study with the proposed model, and discuss some of the lessons learned through this experiment. Finally, in Section 6, we summarize this work and outline several directions for future research.

2 Engagement in Open-World Dialog

In traditional, single-user systems the engagement problem can often be resolved in a relatively simple manner. For instance, in telephony-based applications, it is typically safe to assume that a user is engaged with a dialog system once a call has been received. Similarly, push-to-talk buttons are often used in multimodal mobile applications. Although these solutions are sufficient and even natural in closed, single-user contexts, they become inappropriate for open-world systems that must operate continuously in open, dynamic environments, such as robots, interactive billboards, or embodied conversational agents.

Interaction in the open-world is characterized by two aspects that capture key departures from assumptions traditionally made in spoken dialog systems (Bohus and Horvitz, 2009). The first one is the *dynamic, multiparty* nature of the interaction, *i.e.*, the world typically contains not just one, but multiple agents that are relevant

Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 225–234,
Queen Mary University of London, September 2009.

to the interactive system. Engagements in open worlds are often dynamic and asynchronous, *i.e.* relevant agents may enter and leave the observable world at any time, may interact with the system and with each other, and their goals, needs, and intentions may change over time. Managing the engagement process in this context requires that a system explicitly represents, models, and reasons about multiple agents and interaction contexts, and maintains and leverages long-term memory of the interactions to provide support and assistance.

A second important aspect that distinguishes open-world from closed-world dialog is the *situated* nature of the interaction, *i.e.,* the fact that the surrounding physical environment provides rich, streaming context that is relevant for conducting and organizing the interactions. Situated interactions among people often hinge on shared information about physical details and relationships, including structures, geometric relationships and pathways, objects, topologies, and communication affordances. The often implicit, yet powerful physicality of situated interaction, provides opportunities for making inferences in open-world dialog systems, and challenges system designers to innovate across a spectrum of complexity and sophistication. Physicality and embodiment also provide important affordances that can be used by a system to support the engagement process. For instance, the use of a rendered or physically embodied avatar in a spoken dialog system provides a natural point of visual engagement between the system and people, and allows the system to employ natural signaling about attention and engagement with head pose, gaze, facial expressions, pointing and gesturing.

We present in this paper methods that move beyond the realm of closed-world dialog with a *situated multiparty engagement model* that can enable a computational system to fluidly engage, disengage and re-engage one or multiple people, and support natural interactions in an open-world context.

3 Related Work

The process of engagement between people, and between people and computational systems has received a fair amount of attention. Observational studies in the sociolinguistics and conversational analysis communities have revealed that engagement is a complex, mixed-initiative, highly-coordinated process that often involves a variety of non-verbal cues and signals, (Goffman, 1963; Kendon, 1990), spatial trajectory and proximity (Hall, 1966; Kendon, 1990b), gaze and mutual attention (Argyle and Cook, 1976), head and hand gestures (Kendon, 1990), as well as verbal greetings.

A number of researchers have also investigated issues of engagement in human-computer and human-robot interaction contexts. Sidner and colleagues (2004) define engagement as "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake", and focus on the process of maintaining engagement. They show in a user study (Sidner et al., 2004; 2005) that people directed their attention to a robot more often when the robot made engagement gestures throughout the interaction (*i.e.* tracked the user's face, and pointed to relevant objects at appropriate times in the conversation.) Peters (2005; 2005b) uses an alternative definition of engagement as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction," and present the high-level schematics for an algorithm for establishing and maintaining engagement. The algorithm highlights the importance of mutual attention and eye gaze and relies on a heuristically computed "interest level" to decide when to start a conversation. Michalowski and colleagues (2006) propose and conduct experiments with a model of engagement grounded in proxemics (Hall, 1966) which classifies relevant agents in the scene in four different categories (*present, attending, engaged* and *interacting*) based on their distance to the robot. The robot's behaviors are in turn conditioned on the four categories above.

In our work, we follow Sidner's definition of engagement as a process (Sidner et al., 2004) and describe a computational model for *situated multiparty engagement*. The proposed model draws on several ideas from the existing body of work, but moves beyond it and provides a more comprehensive framework for managing the engagement process in a dynamic, open-world context, where multiple people with different and changing goals may enter and leave, and communicate and coordinate with each other and with the system.

4 Models for Multiparty Engagement

The proposed framework for managing engagement is centered on a reified notion of *interaction*, defined here as a basic unit of sustained, interactive problem-solving. Each interaction involves two or more participants, and this number may vary in time; new participants may join an existing interaction, or current participants may leave an interaction at any point in time. The system is actively engaged in at most one interaction at a time (with one or multiple participants), but it can simultaneously keep track of additional, suspended interactions. In this context, engagement is viewed as the process subsuming the joint, coordinated activities by which participants *initiate, maintain, join, abandon, suspend, resume,* or *terminate* an interaction. Appendix A shows by means of an example the various stages of an interaction and the role played by the engagement process.

Successfully modeling the engagement process in a situated, multi-participant context requires that the system (1) senses and reasons about the engagement state,

actions and intentions of multiple agents in the scene, (2) makes high-level engagement control decisions (*i.e.* about whom to engage or disengage with, and when) and (3) executes and signals these decisions to the other participants in an appropriate and expected manner (*e.g.* renders them in a set of coordinated behaviors such as gestures, greetings, etc.). The proposed model subsumes these three components, which we discuss in more detail in the following subsections.

4.1 Engagement State, Actions, Intentions

As a prerequisite for making informed engagement decisions, a system must be able to recognize various engagement cues, and to reason about the engagement actions and intentions of relevant agents in the scene. To accomplish this, the sensing subcomponent of the proposed engagement model tracks over time three related engagement variables for each agent a and interaction i: the engagement state $ES_a^i(t)$, the engagement action $EA_a^i(t)$ and the engagement intention $EI_a^i(t)$.

The engagement state, $ES_a^i(t)$, captures whether an agent a is engaged in interaction i and is modeled as a deterministic variable with two possible values: *engaged* and *not-engaged*. The state is updated based on the joint actions of the agent and the system (see Figures 3 and 4). Since engagement is a collaborative process, the transitions to the *engaged* state require that both the agent and the system take either an *engage* action (if the agent was previously not engaged) or a *maintain* action (if the agent was already engaged); we discuss these actions in more detail shortly. On the other hand, disengagement can be a unilateral act: an agent transitions to the *not-engaged* state if either the agent or the system take a *disengage* action or a *no-action*.

The second engagement variable, $EA_a^i(t)$, models the actions that an agent takes to initiate, maintain or terminate engagement. There are four engagement actions: *engage*, *no-action*, *maintain*, *disengage*. The first two are possible only from the *not-engaged* state, while the last two are possible only from the *engaged* state. The engagement actions are estimated based on a conditional probabilistic model of the form:

$$P(EA_a^i(t)|ES_a^i(t), EA_a^i(t-1), SEA_a^i(t-1), \Psi(t))$$

The inference is conditioned on the current engagement state, on the previous agent and system actions, and on additional sensory evidence $\Psi(t)$. $\Psi(t)$ includes the detection of explicit engagement cues such as: salutations (*e.g.* "Hi!", "Bye bye"); calling behaviors (*e.g.* "Laura!"); the establishment or the breaking of an F-formation (Kendon, 1990b), *i.e.* the agent approaches and positions himself in front of the system and attends to the system; an expected, opening dialog move (*e.g.* "Come here!"). Note that each of these cues is explicit, and marks a committed engagement action.

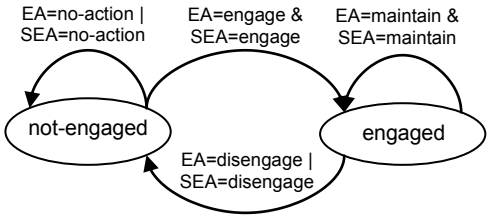


Figure 2. Engagement state transition diagram. EA is the agent's engagement action; SEA is the system's action.

A third variable in the proposed model, $EI_a^i(t)$, tracks the engagement intention of an agent with respect to a conversation. Like the engagement state, the intention can either be *engaged* or *not-engaged*. Intentions are tracked separately from actions since an agent might intend to engage or disengage the system, but not yet take an explicit engagement action. For instance, let us consider the case in which the system is already engaged in an interaction and another agent is waiting in line to interact with the system. Although the waiting agent does not take an explicit, committed engagement action, she might still intend to engage in a new conversation with the system once the opportunity arises. She might also signal this engagement intention via various cues (*e.g.* pacing around, glances that make brief but clear eye contact with the system, etc.) More generally, the engagement intention variable captures whether or not an agent would respond positively should the system initiate engagement. In that sense, it roughly corresponds to Peters' (2005; 2005b) "interest level", *i.e.* to the value the agent attaches to being engaged in a conversation with the system.

Like engagement actions, engagement intentions are inferred based on a direct conditional model:

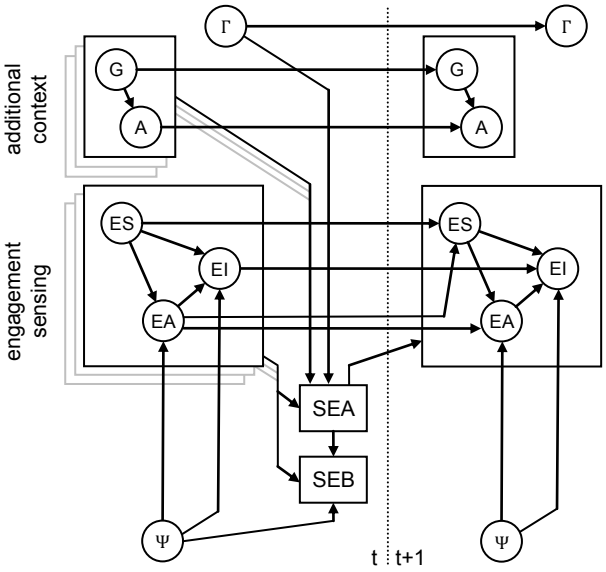


Figure 3. Graphical model showing key variables and dependencies in managing engagement.

$$P(EI_a^i(t)|ES_a^i(t), EA_a^i(t), SEA_a^i(t-1), EI_a^i(\mathrm{t}-1), \Psi(t))$$

This model leverages information about the current engagement state, the previous agent and system actions, the previous engagement intention, as well as additional evidence $\Psi(t)$ capturing implicit engagement cues. Such cues include the spatiotemporal trajectory of the participant and the level of sustained mutual attention. The models for inferring engagement actions and intentions are generally independent of the application. They capture the typical behaviors and cues by which people signal engagement, and, as such, should be reusable across different domains. In other work (Bohus and Horvitz, 2009b), we describe these models in more detail and show how they can be learned automatically from interaction data.

4.2 Engagement Control Policy

Based on the inferred state, actions and intentions of the agents in the scene, as well as other additional evidence to be discussed shortly, the proposed model outputs high-level engagement actions, denoted by SEA decision node in Figure 3. The action-space on the system side contains the same four actions previously discussed: *engage*, *disengage*, *maintain* and *no-action*. Each action is parameterized with a set of agents $\{a_k\}$ and an interaction i. Additional parameters that control the lower level execution of these actions, such as specific greetings, waiting times, urgency, etc. may also be specified. The actual execution mechanisms are discussed in more detail in the following subsection.

In making engagement decisions in an open-world setting, a conversational system must balance the goals and needs of multiple agents in the scene and resolve various tradeoffs (for instance between continuing the current interaction or interrupting it temporarily to address another agent), all the while observing rules of social etiquette in interaction. Apart from the detected engagement state, actions and intentions of an agent $\boldsymbol{E}_a^i = \langle ES_a^i, EA_a^i, EI_a^i \rangle$, the control policy can be enhanced through leveraging additional observational evidence, including high-level information $\boldsymbol{H}_a$ about the various agents in the scene, such as their long-term goals and activities, as well as other global context ($\boldsymbol{\Gamma}$), including the multiple tasks at hand, the history of the interactions, relationships between various agents in the scene (*e.g.* which agents are in a group together), etc. For instance, a system might decide to temporarily refuse engagement even though an agent takes an *engage* action, because it is currently involved in a higher priority interaction. Or, a system might try to take the initiative and engage an agent based on the current context (*e.g.* the system has a message to deliver) and activity of the agent (*e.g.* the agent is passing by), even though the agent has no intention to engage.

Engagement control policies have therefore the form,

$$\pi_{SEA}(\{\boldsymbol{E}_a^i\}_{a,i}, \{\boldsymbol{H}_a\}_a, \boldsymbol{\Gamma})$$

where we have omitted the time index for simplicity. In contrast to the models for inferring engagement intentions and action, the engagement control policy can often be application specific. Such policies can be authored manually to capture the desired system behavior. We will discuss a concrete example of this in Section 5.2. In certain contexts, a more principled solution can be developed by casting the control of engagement as an optimization problem for scheduling collaborations with multiple parties under uncertainties about the estimated goals and needs, the duration of the interactions, time and frustration costs, social etiquette, etc. We are currently exploring such models, where the system also uses information-gathering actions (*e.g.* "Are the two of you together?" "Are you here for X?," etc.), based on value-of-information computations to optimize in the nature and flow of attention and collaboration in multiparty interactions.

4.3 Behavioral Control Policy

At the lower level, the engagement decisions taken by the system have to be executed and rendered in an appropriate manner. With the use of a rendered or physical embodied agent, these actions are translated into a set of coordinated lower-level behaviors, such as head gestures, making and breaking eye contact, facial expressions, salutations, interjections, etc. The coordination of these behaviors is governed by a behavioral control policy, conditioned on the estimated engagement state, actions and intentions of the considered agents, as well as other information extracted from the scene:

$$\pi_{SEB}(SEA, \{\boldsymbol{E}_a^i\}_{a,i}, \Psi)$$

For example, in the current implementation, the *engage* system action subsumes three sub-behaviors performed in a sequence: *EstablishAttention*, *Greeting*, and *Monitor*. First, the system attempts to establish sustained mutual attention with the agent(s) to be engaged. This is accomplished by directing the gaze towards the agents, and if the agent's focus of attention is not on the system, triggering an interjection like "Excuse me!" Once mutual attention is established, on optional *Greeting* behavior is performed; a greeting can be specified as an execution parameter of the *engage* action. Finally, the system enters a *Monitor* behavior, in which it monitors for the completion of engagement. The action completes successfully once the agent(s) are in an engaged state. Alternatively if a certain period of time elapses and the agent(s) have not yet transitioned to the engaged state, the *engage* system action completes with failure (which is signaled to the engagement control layer).

Like the high-level engagement control policies, the behavioral control policies can either be authored manually, or learned from data, either in a supervised (*e.g.*

from a human-human interaction corpus) or unsupervised learning setting. Also, like the engagement sensing component, the behavioral control component is decoupled from the task at hand, and should be largely reusable across multiple application domains.

5 Observational Study

As an initial step towards evaluating the proposed situated multiparty engagement models, we conducted a preliminary observational study with a spoken dialog system that implements these models. The goals of this study were (1) to investigate whether a system can use the proposed engagement models to effectively create and conduct multiparty interactions in an open-world setting, (2) to study user behavior and responses in this setting, and (3) to identify some of the key technical challenges in supporting multiparty engagement and dialog in open-world context. In this section, we describe this study and report on the lessons learned.

5.1 Experimental platform

Studying multiparty engagement and more generally open-world interaction poses significant challenges. Controlled laboratory studies are by their very nature closed-world. Furthermore, providing participants with instructions, such as "Go interact with this system", or "Go join the existing interaction" can significantly prime and alter the engagement behaviors they would otherwise display upon encountering the system in an unconstrained setting. This can in turn cast serious doubts on the validity of the results. Open-world interaction is best observed in the open-world.

To provide an ecologically valid basis for studying situated, multiparty engagement we therefore developed a conversational agent that implements the proposed model, and deployed it in the real-world. The system, illustrated in Figure 4, takes the form of an interactive multi-modal kiosk that displays a realistically rendered avatar head which can interact via natural language. The avatar can engage with one or more participants and plays a simple game, in which the users have to respond to multiple-choice trivia questions.

The system's hardware and software architecture is illustrated in Figure 4. Data gathered from a wide-angle camera, a 4-element linear microphone array, and a 19" touch-screen is forwarded to a scene analysis module that fuses the incoming streams and constructs in real-time a coherent picture of the dynamics in the surrounding environment. The system detects and tracks the location of multiple agents in the scene, tracks the head pose for engaged agents, tracks the current speaker, and infers the focus of attention, activities, and goals of each agent, as well as the group relationships among different agents. An in-depth description of the hardware and scene analysis components falls beyond the scope of this paper, but details are available in (Bohus and Horvitz, 2009). The scene analysis results are forwarded to the control level, which is structured in a two-layer reactive-deliberative architecture. The reactive layer implements and coordinates various low-level behaviors, including engagement, conversational floor management and turn-taking, and coordinating spoken and gestural outputs. The deliberative layer plans the system's dialog moves and high-level engagement actions.

Overall, the game task was purposefully designed to minimize challenges in terms of speech recognition or dialog management, and allow us to focus our attention on the engagement processes. The avatar begins the interactions by asking the engaged user if they would like to play a trivia game. If the user agrees, the avatar goes through four multiple-choice questions, one at a time. After each question, the possible answers are displayed on the screen (Figure 4) and users can respond by either speaking an answer or by touching it. When the answer provided by the user is incorrect, the system provides a short explanation regarding the correct answer before moving on to the next question.

The system also supports multi-participant interactions. The engagement policy used to attract and engage

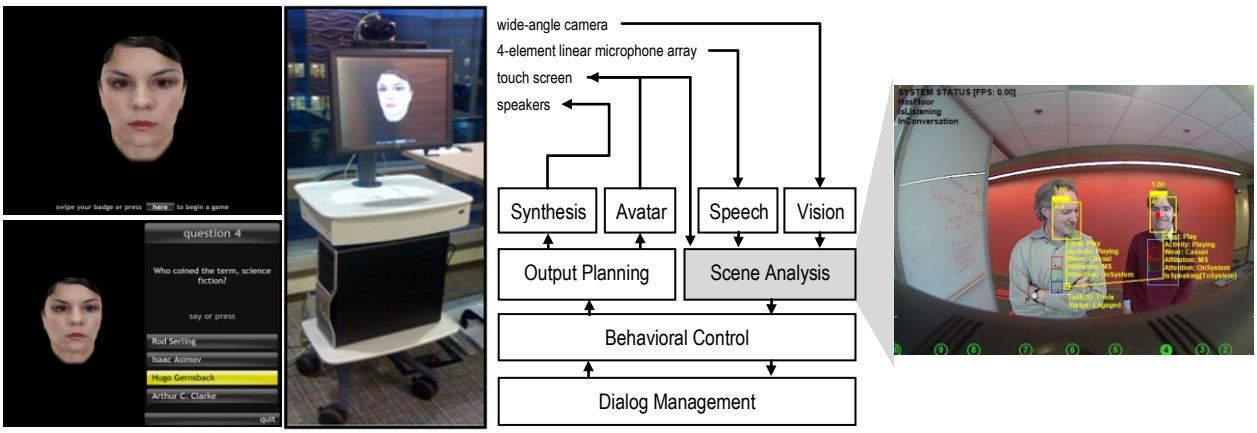


Figure 4. Trivia game dialog system: prototype, architectural overview, and runtime scene analysis

multiple users in a game is the focus of this observational study, and is discussed in more detail in the next subsection. Once the system is engaged with multiple users, it uses a multi-participant turn taking model which allows it to continuously track who the current speaker is, and who has the conversational floor (Bohus and Horvitz, 2009). At the behavioral level, the avatar orients its head pose and gaze towards the current speaker, or towards the addressee(s) of its own utterances. During multiplayer games, the avatar alternates between the users when asking questions. Also, after a response is received from one of the users, the avatar confirms the answer with the other user(s), *e.g.* "Do you agree with that?" A full sample interaction with the system is described in Appendix A, and the corresponding video is available online (Situated Interaction, 2009).

5.2 Multiparty Engagement Policy

The trivia game system implements the situated, multiparty engagement model described in Section 4. The sensing and behavioral control components are application independent and were previously described. We now describe the system's engagement policy, which is application specific.

As previously discussed, apart from using the inferred engagement state, actions and intentions for the agents in the scene, the proposed model also uses information about the high-level goals and activities of these agents when making engagement decisions. Specifically, the system tracks the goal of each agent in the scene, which can be *play*, *watch*, or *other*, and their current activity, which can be *passing-by, interacting, playing, watching,* or *departing*. The goal and activity recognition models are application specific, and in this case are inferred based on probabilistic conditional models that leverage information about the spatiotemporal trajectory of each agent and their spoken utterances, as well as global scene information (*e.g.* is the system engaged in an active interaction, etc.).

Initially, when the system is idle, it uses a conservative engagement policy and waits for the user to initiate engagement via an explicit action. Such actions include the user approaching and entering in an F-formation (Kendon, 1990b) with the system, *i.e.* standing right in front of it, swiping their badge, or pushing the start button (in the idle state the GUI displays "swipe your badge or press here to begin" below the avatar head).

While engaged in an interaction, the system attempts to engage bystanders in an effort to create a collaborative, multi-participant game. In this case, the engagement policy is conditioned on the inferred activities of the agents in the scene. Specifically, if a *watching* bystander is detected, the system temporarily disengages the current participant, and engages and attempts to "convince" the watching bystander to join the existing game. The prompts in this side interaction depend on the current game context, as shown in Table 1. If the watching bystander agrees to join in, the system adds him to the existing interaction, and continues a multi-participant game (see Table 1.A.) Conversely, if the bystander refuses, the system re-engages the previous participant and resumes the single-user game (see Table 1.B.) Additional examples are available in Appendix A.

If a *watching* bystander B is detected during the first 3 questions		
	and the *engaged* participant E has not made any mistakes	
A		[S to E]: Hold on a second. [S to B]: Hi. Excuse me, would you like to join in? [B]: Okay. [S to E&B]: Great. So let's see how well you can do together. Next question […]
	and the engaged participant E has just made a mistake, and the system did not previously engage the bystander B.	
B		[S to E]: Perhaps you can use some help. Hold on a second. [S to B]: Hi. Do you think you could help out? [B]: No thanks [S to B]: Okay. Well, I guess you can just watch for now and if you wait for a bit we'll be done and you can start a new game yourself [S to E]: Guess you're still on your own. Next question […]
	and the engaged participant E has just made a mistake, and the system did previously attempt to engage the bystander B.	
C		[S to E]: I think you could really use some help. Hold on a second. [S to B]: Are you sure you don't want to help out? Come on, this is fun. [B]: Sure [S to E&B]: Great. So let's see how well you can do together. Next question […]
If a *watching* bystander B is detected during the last question		
D		[S to E]: Excuse me for one moment. [S to B]: We're almost done here. If you wait for a bit we can start a new game right after [S to E]: Sorry about that […]

Table 1. Multiparty engagement policy

Finally, if the system is already engaged and a *watching* bystander is detected but only during the last question, the system engages them temporarily to let them know that the current game will end shortly and, if they wait, they can also start a new game (see Table 1.D).

5.3 Results and Lessons Learned

We deployed the system described above for 20 days near one of the kitchenettes in our building. The system attracted attention of passer-bys with the tracking motion of its virtual face that followed people as they passed by. Most people that interacted with the system did so for the first time; only a small number of people interacted several times. No instructions were provided for interacting with the system. We shall now review results from analysis of the collected data.

Throughout the 20 days of deployment, the system engaged in a total of 121 interactive events. Of these, in 54 cases (44%), a participant engaged the system but did not play the game. Typically, the participant would approach and enter in an F-formation with the system,

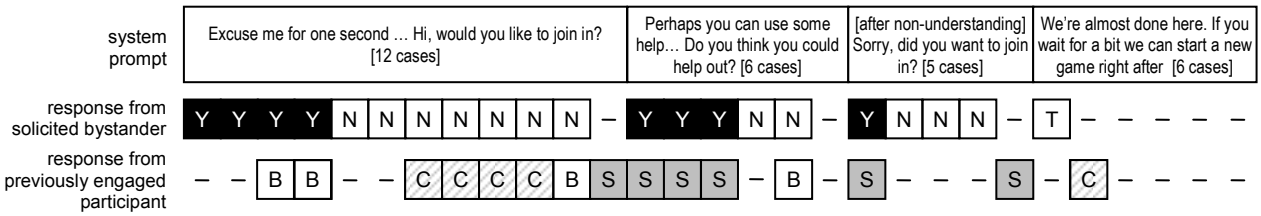


Figure 5. System multiparty engagement actions and responses from bystanders and already engaged participants. For bystander responses, Y denotes a positive response; N denotes a negative response; – denotes no response. For responses from previously engaged participant, B denotes utterances addressed to the bystander, C denotes side comments, S denotes responses directed to the system

but, once the system engaged and asked if they would like to play the trivia game, they responded negatively or left without responding. In 49 cases (40%), a single participant engaged and played the game, but no bystanders were observed during these interactions. In one case, two participants approached and engaged simultaneously; the system played a multi-participant game, but no other bystanders were observed. Finally, in the remaining 17 cases (14% of all engagements, 25% of actual interactions), at least one bystander was observed and the system engaged in multiparty interaction. These multiparty interactions are the focus of our observational analysis, and we will discuss them in more detail.

In 2 of these 17 cases, bystanders appeared only late in the interaction, after the system had already asked the last question. In these cases, according to its engagement policy, the system notified the bystander that they would be attended to momentarily (see Table 1.D), and then proceeded to finish the initial game. In 8 of the remaining 15 cases (53%), the system successfully persuaded bystanders to join the current interaction and carried on a multi-participant game. In the remaining 7 cases (47%), bystanders turned down the offer to join the existing game. Although this corpus is still relatively small, these statistics indicate that the system can successfully engage bystanders and create and manage multi-participant interactions in the open world.

Next, we analyzed more closely the responses and reactions from bystanders and already engaged participants to the system's multiparty engagement actions. Throughout the 17 multiparty interactions, the system planned and executed a total of 23 engagement actions soliciting a bystander to enter the game, and 6 engagement actions letting a bystander know that they will be engaged momentarily. The system actions and responses from bystanders and engaged participants are visually summarized in Figure 5, and are presented in full in Appendix B. Overall, bystanders successfully recognize that they are being engaged and solicited by the system and respond (either positively or negatively) in the large majority of cases (20 out of 23). In 2 of the remaining 3 cases, the previously engaged participant responded instead of the bystander; finally, in one case the bystander did not respond and left the area.

While bystanders generally respond when engaged by the system, the system's engagement actions towards bystanders also frequently elicits spoken responses from the already engaged participants; this happened in 14 out of 23 cases (61%). The responses are sometimes addressed to the system *e.g.* "Yes he does," or towards the bystander, *e.g.* "Say yes!", or they reflect general comments, *e.g.* "That's crazy!" These results show that, when creating the side interaction to solicit a bystander to join the game, the system should engage both the bystander and the existing user in this side interaction, or at least allow the previous user to join this side interaction (currently the system engages only the bystander in this interaction; see example from Appendix A.)

Furthermore, we noticed that, in several cases, bystanders provided responses to the system's questions even prior to the point the system engaged them in interaction (sometimes directed toward the system, sometimes toward the engaged participant.) We employed a system-initiative engagement policy towards bystanders in the current experiment. The initiative being taken by participants highlights the potential value of implementing a mixed-initiative policy for engagement. If a relevant response is detected from a bystander, this can be interpreted as an engagement action (recall from subsection 4.1 that engagement actions subsume expected opening dialog moves), and a mixed-initiative policy can respond by engaging the bystander, *e.g.* "Did you want to join in?" or "Please hang on, let's let him finish. We can play a new game right after that." This policy could be easily implemented under the proposed model.

We also noted side comments by both bystander and the existing participant around the time of multiparty engagement. These remarks typically indicate surprise and excitement at the system's multiparty capabilities. Quotes include: "That's awesome!", "Isn't that great!", "That's funny!", "Dude!", "Oh my god that's creepy!", "That's cool!", "It multitasks!", "That is amazing!", "That's pretty funny", plus an abundance of laughter and smiles. Although such surprise might be expected today with a first-time exposure to an interactive system that is aware of and can engage with multiple parties, we believe that expectations will change in the future, as these technologies become more commonplace.

Overall, this preliminary study confirmed that the system can effectively initiate engagement in multiparty settings, and also highlighted several core challenges for managing engagement and supporting multiparty interactions in the open world. A first important challenge we have identified is developing robust models for tracking the conversational dynamics in multiparty situations, *i.e.* identifying who is talking to whom at any given point. Secondly, the experiment has highlighted the opportunity for using more flexible, mixed-initiative engagement policies. Such policies will rely heavily on the ability to recognize engagement intentions; in (Bohus and Horvitz, 2009b), we describe the automated learning of engagement intentions from interaction data. Finally, another lesson we learned from these initial experiments is the importance of accurate face tracking for supporting multiparty interaction. Out of the 17 multiparty interactions, 7 were affected by vision problems (e.g. the system momentarily lost a face, or swapped the identity of two faces); 4 of these were fatal errors that eventually led to interaction breakdowns.

6 Summary and Future Work

We have described a computational model for managing engagement decisions in open-world dialog. The model harnesses components for sensing and reasoning about the engagement state, actions, and intentions of multiple participants in the scene, for making high-level engagement control decisions about who and when to engage, and for executing and rendering these actions in an embodied agent. We reviewed an observational study that showed that, when weaved together, these components can provide support for effectively managing engagement, and for creating and conducting multiparty interactions in an open-world context.

We believe that the components and policies we have presented provide a skeleton for engagement and interaction in open-world settings. However, there are important challenges and opportunities ahead. Future research includes developing methods for fine tuning and optimizing each of these subcomponents and their interactions. Along these lines, there are opportunities to employ machine learning to tune and adapt multiple aspects of the operation of the system. In (Bohus and Horvitz, 2009b) we introduce and evaluate an approach to learning models for inferring engagement actions and intentions online, through interaction. On another direction, we are investigating the use of decision-theoretic approaches for optimizing mixed-initiative engagement policies by taking into account the underlying uncertainties, the costs and benefits of interruption versus continuing collaboration, queue etiquette associated with expectations of fairness, etc. Another difficult challenge is the creation of accurate low-level behavioral models, including the fine-grained control of pose, gesture, and facial expressions. Developing such methods will likely have subtle, yet powerful influences on the effectiveness of signaling and overall grounding in multiparty settings. We believe that research on these and other problems of open-world dialog will provide essential and necessary steps towards developing computational systems that can embed interaction deeply into the natural flow of everyday tasks, activities, and collaborations.

Acknowledgments

We thank George Chrysanthakopoulos, Zicheng Liu, Tim Paek, Cha Zhang, and Qiang Wang for discussions and feedback in the development of this work.

References

M. Argyle and M. Cook, 1976, *Gaze and Mutual Gaze*, Cambridge University Press, New York

D. Bohus and E. Horvitz, 2009a, *Open-World Dialog: Challenges, Directions and Prototype*, to appear in KRPD'09, Pasadena, CA

D. Bohus and E. Horvitz, 2009b, *An Implicit-Learning Based Model for Detecting Engagement Intentions*, submitted to SIGdial'09, London, UK

E. Goffman, 1963, *Behaviour in public places: notes on the social order of gatherings*, The Free Press, New York

E.T. Hall, 1966, *The Hidden Dimension: man's use of space in public and private*, New York: Doubleday.

A. Kendon, 1990, *A description of some human greetings*, Conducting Interaction: Patterns of behavior in focused encounters, Studies in International Sociolinguistics, Cambridge University Press

A. Kendon, 1990b, *Spatial organization in social encounters: the F-formation system*, Conducting Interaction: Patterns of behavior in focused encounters, Studies in International Sociolinguistics, Cambridge University Press

M.P. Michalowski, S. Sabanovic, and R. Simmons, *A spatial model of engagement for a social robot*, in 9th IEEE Workshop on Advanced Motion Control, pp. 762-767

C. Peters, C. Pelachaud, E. Bevacqua, and M. Mancini, 2005, *A model of attention and interest using gaze behavior*, Lecture Notes in Computer Science, pp. 229-240.

C. Peters, 2005b, *Direction of Attention Perception for Conversation Initiation in Virtual Environments*, in Intelligent Virtual Agents, 2005, pp. 215-228.

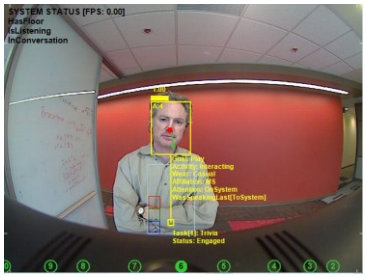
C.L. Sidner, C.D. Kidd, C. Lee, and N. Lesh, 2004, *Where to Look: A Study of Human-Robot Engagement*, IUI'2004, pp. 78-84, Madeira, Portugal

C.L. Sidner, C. Lee, C.D. Kidd, N. Lesh, and C. Rich, 2005, *Explorations in engagement for humans and robots*, Artificial Intelligence, 166 (1-2), pp. 140-164

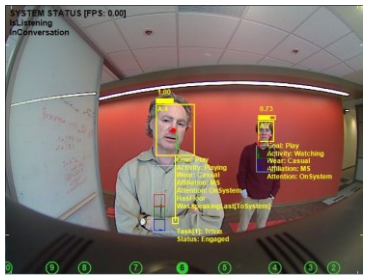
Situated Interaction, 2009, Project page: http://research.microsoft.com/~dbohus/research_situated_interaction.html

R. Vertegaal, R. Slagter, G.C.v.d.Veer, and A. Nijholt, 2001, *Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes*, CHI'01

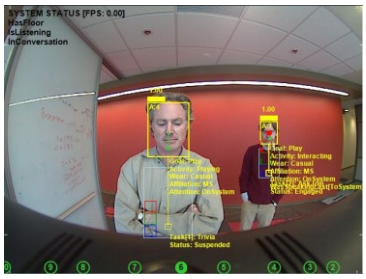
Appendix A. Sample multiparty interaction with trivia game dialog system (not part of the experiment)



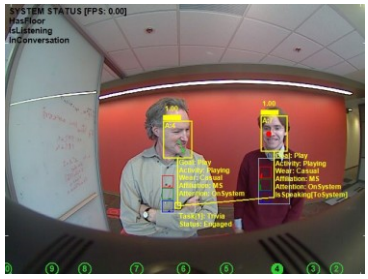
first person engages - around time t2



bystander appears – prior to t_3

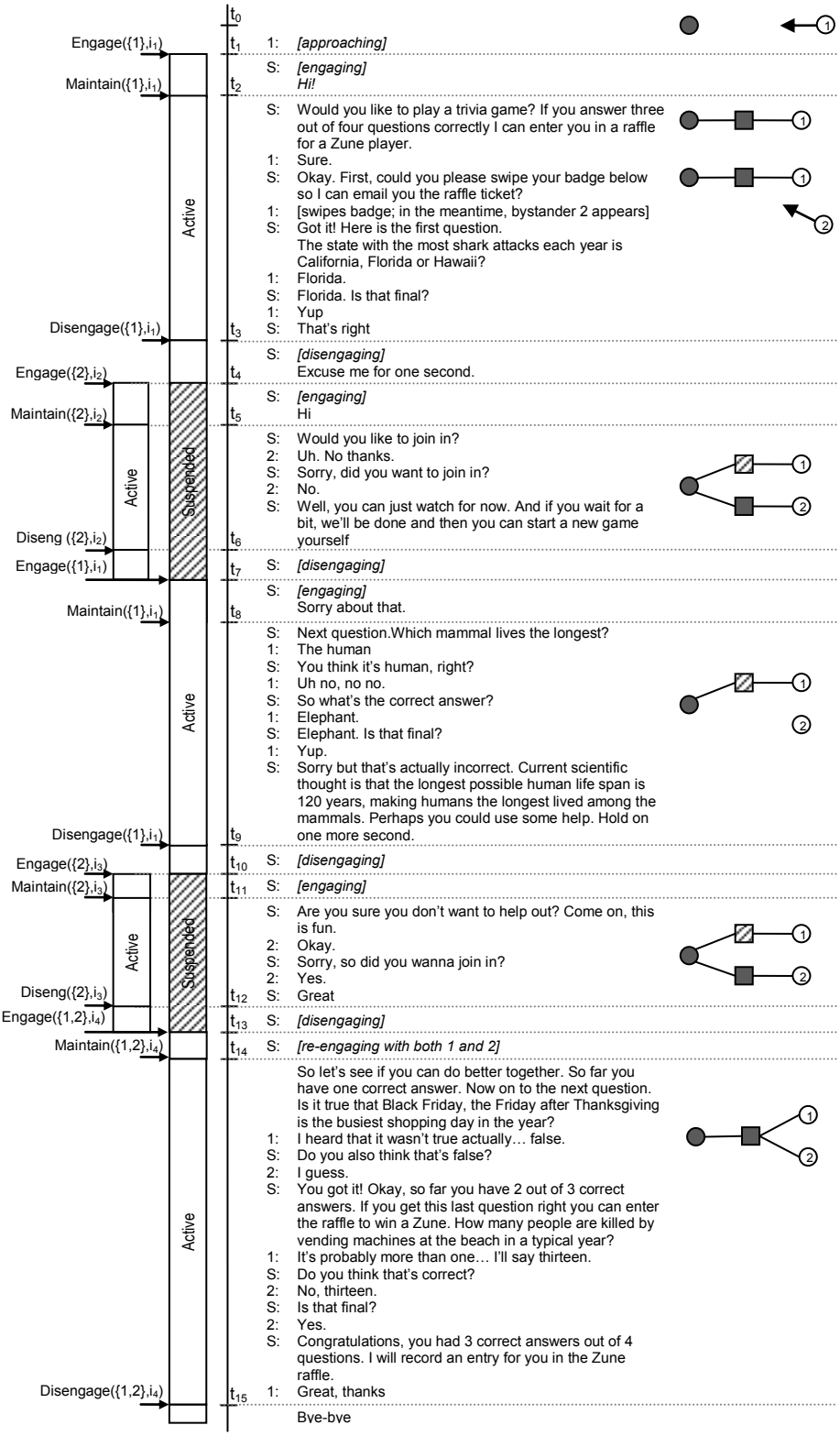


system engages bystander ~ t_5



participants play together ~ t_{14}

system
agent
active conversation
suspended conversation
other conversation



Appendix B. User responses to multiparty engagement actions.

S denotes the system, E denotes the already engaged participant, B denotes a *watching* bystander.

Actions and Resposes	Response from B	Response from E	Timing
[S to E]: *Hold on one second.* [S to B]: *Excuse me, would you like to join in?* **4 positive answers from B** **7 negative answers from B** **1 no answer from B (E answers)**	*Yes*		B only
	Yes	*Say yes*	Overlap
	Sure		B only
	Yes	[to B]: *Would you like to join in?*	E first
	No	*That's crazy!*	B first
	Oh, no. No + [moves away]	*That's funny!*	B first
	No thank you		B only
	No	*No?*	B first
	Woah, no.	*That's cool!*	B first
	No + [moves away] + *That's pretty funny.*	[laughs looking at B]	B first
	[laughs]	[laughs] *Yes. Oh yes.*	E only
[S to E]: *Perhaps you could use some help. Excuse me for one second.* [S to B]: *Hi, do you think you could help out?* **3 positive answers from B** **2 negative answers from B** **1 no-answer from B (moves away)**	*Yes.*	*Yes.*	B first
	Yes	*Yes he does.*	Overlap
	[laughs] + *No.*		B only
	[to E]: *Isn't that weird?* [to S]: *No.* [to E]: *Isn't that great?*	[to B]: *That is amazing!*	B first
	[laughs] + [moves out]	*Quit*	E only
	[laughs] + *Sure*	*Sure*	B first
If the initial response from B was not understood by the system, system asks one more time [S to B]: *Sorry, did you want to join in?* **1 positive answer from B** **3 negative answer from B** **1 no-answer from B (E answers)**	*No.* *Please.*	*Yes, I don't know, help me!*	B first
	No.		B only
	No.		B only
	No.		B only
		No.	E only
[S to B]: *We're almost done here. If you wait for a bit we can start a new game right after.* **1 answer from B** **1 answer from E** **4 no-answer from either B or E**	*Great, thanks.*		B only
		That's awesome	E only