A Four-Participant Group Facilitation Framework for Conversational Robots

Yoichi Matsuyama, Iwao Akiba, Akihiro Saito, Tetsunori Kobayashi
Department of Computer Science, Waseda University
27 Waseda, Shunjuku-ku, Tokyo, Japan
{matsuyama,akiba,saito}@pcl.cs.waseda.ac.jp
koba@waseda.jp

Abstract

In this paper, we propose a framework for conversational robots that facilitates four-participant groups. In three-participant conversations, the minimum unit for multiparty conversations, social imbalance, in which a participant is left behind in the current conversation, sometimes occurs. In such scenarios, a conversational robot has the potential to facilitate situations as the fourth participant. Consequently, we present model procedures for obtaining conversational initiatives in incremental steps to engage such four-participant conversations. During the procedures, a facilitator must be aware of both the presence of dominant participants leading the current conversation and the status of any participant that is left behind. We model and optimize these situations and procedures as a partially observable Markov decision process. The results of experiments conducted to evaluate the proposed procedures show evidence of their acceptability and feeling of groupness.

1 Introduction

We present a framework for conversational robots that facilitates four-participant groups with proper procedures for obtaining initiatives. Figure 1 (a) depicts a two-participant conversation. In such situations, conversational contexts including floor exchanges are commonly grounded between two interlocutors. Many dialogue systems have dealt with such two-participant situations (Raux and Eskenazi, 2009) (Chao and Thomaz, 2012). However, in three-participant conversations, as is shown in Figure 1(b), which is the minimum unit for multiparty conversation, floor exchanges cannot always be identified among the participants. Clark presented the participation structure model (Clark, 1996), drawing on Goffman's work (Goffman, 1981). In such three-participant situations, interactions between two dominant participants out of the three primarily occur (between participant A and B) and the other participant, who cannot properly get the floor to speak for a long while (cannot be promoted to be either a speaker or an addressee) tends to get left behind, even though all of them are "ratified participants" considered by the current speaker.

In terms of engagement among conversational participants, Martin et al., (Martin and White, 2005) proposed the appraisal theory that encompasses three subcategories, namely *Attitude*, *Engagement*, and *Graduation*. *Attitude* deals with expressions of affect, judgement, and appreciation. *Engagement* focuses on language use by which speakers negotiate an interpersonal space for their positions and the strategies which they uses to either acknowledge, ignore, or curtail other voices or points of view. *Graduation* focuses on the resources by which sparkers regulate the impact of these resources. Sidner et al., defined engagement as "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake" (Sidner et al., 2004). Based on these previous studies, we define engagement as the process establishing connections among participants using dialogue actions so that they can represent their own positions properly. So, the three-participant model dictates the need for one more participant who helps the participant who is left behind to engage him/her in the conversation. Conversational robots have the potential to participate in such conversations as the fourth participant, as illustrated in Figure 1 (c-1). Figure 1 (c-2) gives an example of the participants' speech activities in a certain duration. In this example, participant C's activity is relatively smaller than that of the others, and so he/she is likely to get left behind in the current conversational situation for a number of reasons. When a robot steps into the situation to coordinate, there should be proper procedures in place to obtain initiatives to control conversational contexts and to give it back to the others. If a robot naively starts to approach a participant who is left behind just after a left-behind situation is detected, it could break the current conversation. In order to coordinate situations, a facilitator (robot) must take the following procedural steps: (1) Be aware of both the presence of dominant participants leading the current conversation and the status of a participant who is left behind, (2) Obtain the initiative to control the situation and wait for approval from the others, either explicitly or implicitly, and (3) Give the floor to a suitable participant.

Research on specially situated facilitation agents in

Proceedings of the SIGDIAL 2013 Conference, pages 284–293,
Metz, France, 22-24 August 2013.

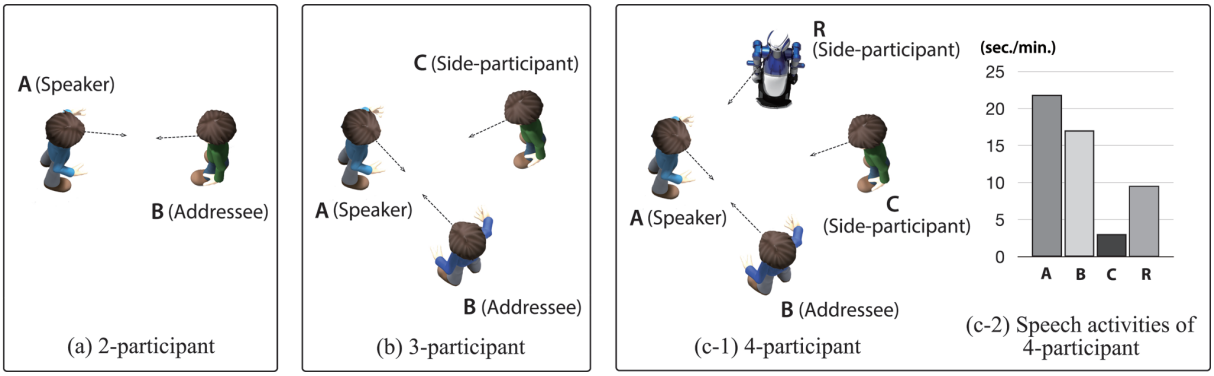


Figure 1: Types of conversations according to number of participants (dashed arrows represent their gazes): (a) Two-participant conversation model, which conventional dialogue systems have focused on. (b) Three-participant conversation model, the minimum unit for a multiparty conversation. In such multiparty conversations, social imbalance occasionally occurs. (c-1) Four-participant conversation, with a robot that regulates the imbalance situation, and (c-2) chart showing the unequal speech activities of the participants. In this case, participant C appears to have less opportunity to take the floor to speak, hence, the robot is expected to help him.

multiparty conversations has been conducted by various researchers. Matsusaka et al. pioneered the act of a physical robot participating in multiparty conversations (Matsusaka et al., 2003). We previously developed a multiparty quiz-game type facilitation system for elderly care (Matsuyama et al., 2008) and reported on the effectiveness of the existence of a robot (Matsuyama et al., 2010). Dosaka et al. developed a thought-evoking dialogue system for multiparty conversations with a quiz game task (Dohsaka et al., 2009). They reported that the existence of agents and empathic expressions are effective for user satisfaction and increase the number of user utterances. Bohus modeled engagement in multiparty conversations along Sinder's definition, namely open world dialogue (Bohus and Horvitz, 2009). In terms of facilitation, Benne et al. (Benne and Sheats, 1948) and Bales (Bales, 1950) pioneered investigations into small group dynamics, including functional facilitation roles. Kumar et al. designed a dialogue action selection model based on Bales's Socio-Emotional Interaction Categories for text-based character agents (Kumar et al., 2011).

In this paper, we propose a framework of procedural facilitation process to increase the total engagement of a group, with caring about side-effects of behaviors at the same time. The situations and procedures are modeled and optimized as a partially observable Markov decision process (POMDP), which is suitable for real-world sequential decision processes, including dialogue systems (Williams and Young, 2007). We begin by reviewing facilitation of small groups, and summarize requirement specifications for facilitation robots in the next section. In Section 3, we first describe representations of small group situations and procedures for maintaining small groups, then we discuss how to model them as POMDP. In Section 4, we give an overview of the architecture of our proposed system. We then discuss two experiments conducted to verify the efficacy of the small group maintenance procedures. Finally, we summarize our work and conclude this paper.

2 Facilitating Small Groups

2.1 Maintaining Small Groups

Benne et al. analyzed functional roles in small groups to understand the activities of individuals in small groups (Benne and Sheats, 1948). They categorized functional roles in small groups into three classes: *Group task roles*, *Group building and maintenance roles*, and *Individual roles*. The *Group task roles* are defined as "related to the task which the group is deciding to undertake or has undertaken." Those roles address concerns about the facilitation and coordination activities for task accomplishment. The *Group building and maintenance roles* are defined as "oriented toward the functioning of the group as a group." They contribute to social structures and interpersonal relations. Finally, the *Individual roles* are directed toward the individual satisfaction of each participant's individual needs. They deal with individual goals that are not relevant either to the group task or to group maintenance. Drawing on Benne's work, Bales proposed interaction process analysis (IPA), a framework for the classification of individual behavior in a two-dimensional role space consisting of a *Task area* and a *Socio-emotional area* (Bales, 1950). The roles related to the *Task area* concern behavioral manifestations that impact the management and solution of problems that a group is addressing. Examples of task-oriented activities include initiating the floor, giving information, and providing suggestions regarding a task. The roles related to the *Socio-emotional area* affect the interpersonal relationships either by supporting, enforcing, or weakening them. For instance, complementing another person to increase group cohesion and mutual trust among mem-

bers is one example of positive socio-emotional behavior. Benne's typology of functional roles is evaluated as valuable with remarkable accuracy. In this paper, we employ Benne's *Group building and maintenance roles*,[1] which are related to Bales's *Socio-emotional area*, in order to arrange the following three abstract functional roles of group maintenance:

1. **Topic Maintenance Role**: Maintaining for conflict, ideas, and topics. This person mediates the difference between other members, attempts to reconcile disagreements, and relieves tension in conflict situations. This role inherits *Compromiser*, *Harmonizer*, and *Standard setter*.

2. **Floor Maintenance Role**: Maintaining the chance for the floor in the group in a direct/indirect way. This person encourages or asks questions of the person who is not or could not get engaged in conversations, and attempts to keep the communication channel open. This role inherits *Gatekeeper*, *Expediter*, and *Encourager*.

3. **Observation Role**: Overlooking the conversation situation by finding appropriate topics, observing the motivations and moods of the participants, and comprehending the relations between participants in conversations. This person follows the conversation and comments and interprets the group's internal process. This role inherits *Observer and commentator* and *Encourager*.

2.2 Procedures for Small Group Maintenance

In order that a participant who wants to claim an initiative (we call this participant a "claimant") is transferred an initiative by the participant leading the current conversation (we call this participant a "leader"), the claimant must take procedural steps. First, the claimant must participate in the current dominant conversation the leader is leading, try to claim an initiative, and then wait for either explicit or implicit approval from the leader. Let us take the example shown in Figure 2. In the figure, participants A and B are primarily leading the current conversation. Participant C cannot get the floor to speak, and so the robot desires to give the floor to C. If the robot speaks to C directly, without being aware of A and B, the conversation might be broken, or separated into two (A-B and C-Robot), at best. In order not to break the situation, the robot should participate in the dominant conversation between A and B first, and set the stage such that the robot is approved to initiate the next situation. In this paper, we define such a state in which a person is participating in a dominant conversation as a "*Engaged*" state, and the opposite state as "*Unengaged*". Thus, in Clark's partic-

[1] Benne's *Group building and maintenance roles* are *Compromiser*, *Harmonizer*, *Standard setter*, *Gatekeeper and expediter*, *Encourager*, *Observer and commentator*, and *Follower*.

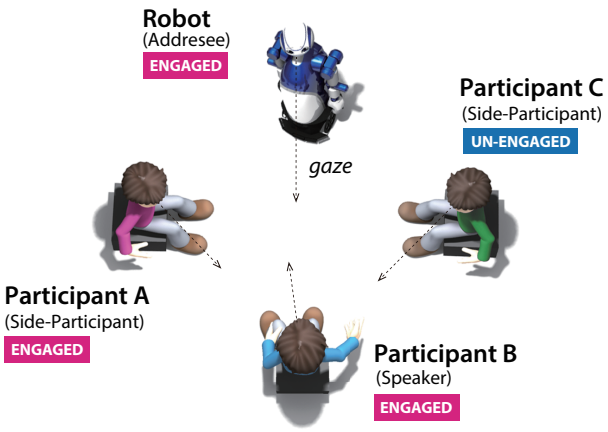


Figure 2: Four-participant conversational group. Four participants, including a robot, are talking about a certain topic. Participants A and B are leading the conversation, and mainly keep the floor. The robot also engages with A and B in line with the topic. C is an *unengaged* participant, who does not have many chances to take the floor for a while. The dashed arrows indicate the direction they are facing, assuming their gazes.

ipation structure, speaker and addressee are automatically *Engaged* participants. Side-participants are divided into *Engaged* and *Unengaged* participants based on their situations. In this paper, we assume that an *Unengaged* participant needs to respond to a *Engaged* participant's adjacency pair part to be engaged. Adjacency pairs are minimal units of conversation that are composed of two utterances by several speakers (Schegloff and Sacks, 1973). The speaking of the first utterance (the first part) provokes a responding utterance (the second part), and sometimes a third response (the third part). Understanding adjacency pairs is, therefore, essential to detecting cut-in timing.

On the basis of our discussion above, we define the following constraints for both *Engaged* and *Unengaged* participants when they address and shift current topics:

1. **Constraint of addressing**: An unengaged participant must not address the other unengaged participants directly.

2. **Constraint of topic shifting**: An engaged participant must not shift the current topic when he/she addresses the other unengaged participants.

The relationship between subjective and objective participants that are permitted to approach in the two constraints are shown in Tables 1 and 2. In the following sections, we describe a computational model that has the group maintenance functions discussed above.

3 Procedure Optimization

3.1 Representation of Engagement State

We assume only one speaker and one addressee exist at each time-step and one or two side-participants may

Table 1: Permission relationship between subjective and objective participants for the constraint of addressing. "Engaged" means a participant is assigned as a speaker or an addressee or a side-participant, who engages with the conversational group. "Unengaged" means a participant is assigned as an unengaged side-participant.

Subject \ Objective	Engaged	Unengaged
Engaged	permitted	permitted
Unengaged	permitted	NOT permitted

Table 2: Permission relationship for permission between subjective and objective participants in the constraint of topic shifting.

Subject \ Objective	Engaged	Unengaged
Engaged	permitted	NOT permitted
Unengaged	NOT permitted	NOT permitted

exist in four-participant conversations. We define side-participants as having two states: "*Engaged*" and "*Unengaged*". In the scenario shown in Figure 2, participant C may not be able to take the floor for a while. The situation probably resolves itself when the current topic is shifted. Hence, we define the depth of side-participant $Depth_{SPT}$ as the duration that a participant is assigned while the same topic continues, which represents the level of engagement.

$$Depth_{SPT_i} = Duration_{SPT_i} / Duration_{topic_j} \quad (1)$$

$$UnengagedSPT = \begin{cases} SPT_i & \text{if } Depth_{SPT_i} > Threshold \\ none & \text{otherwise} \end{cases} \quad (2)$$

The suffix i represents a participant's ID.

We also define an *Un-Engaged* participant's motivation to speak on the current topic. Thus, this state affects decision-making about topic maintenance. The amount of motivation of a participant is calculated as a linear sum of speech activities, smiling duration, and nodding duration. Further, the motivations in our current model are heuristically assumed to be binary variables.

$$Motivation_i = \begin{cases} 1 & \text{if } MotivAmount_i > Threshold \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

3.2 Procedure Optimization using POMDP

To optimize the procedures discussed above, we model the task as a partially observable Markov decision process (POMDP) (Williams and Young, 2007). Formally, a POMDP is defined as a tuple $\beta = \{S, A, T, R, O, Z, \gamma, b_0\}$, where S is a set of states describing the agent's world, A is a set of actions that the agent may take, T defines a transition probability $P(s'|s,a)$, R defines the expected reward $r(s,a)$, O is a set of observations the agent can receive about the world, and Z defines an observation probability, $P(o'|s',a)$, γ is a geometric discount factor $0 < \gamma < 1$, and b_0 is an initial belief state $b_0(s)$. At each time-step, the belief state distribution b is updated as follows:

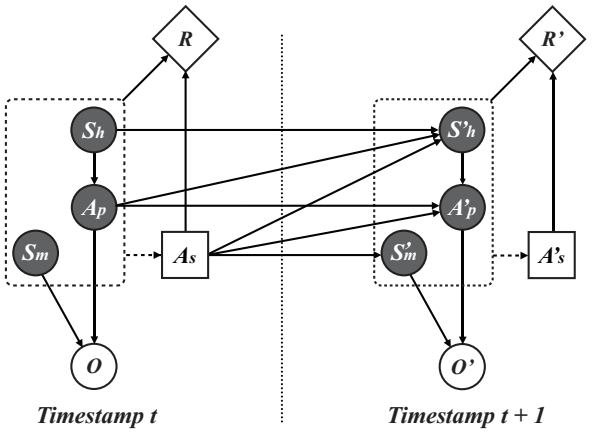


Figure 3: Influence diagram representation of the POMDP model. Circles represent random variables, squares represent decision nodes, and diamonds represent utility nodes. Shaded circles indicate random variables, while unshaded circles represent observed variables. Solid directed arcs indicate casual effect, while dashed directed arcs indicate that a distribution is used.

$$b'(s') = \gamma \cdot P(o'|s',a) \sum_s P(s'|s,a) b(s) \quad (4)$$

In this paper, we assume S can be factored into three components: the participants' engagement states S_e, the participants' motivation states S_m, and the participants' actions A_p. Hence, the factored POMDP state S is defined as

$$s = (s_e, s_m, a_p) \quad (5)$$

and the belief state b becomes

$$b = b(s_e, s_m, a_p) \quad (6)$$

To compute the transition function and observation function, a few intuitive assumptions are made:

$$\begin{aligned} P(s'|s,a) =& P(s'_e, s'_m, a'_p | s_e, s_m, a_p, a_s) \\ =& P(s'_e | s_h, s_m, a_p, a_s) \cdot \\ & P(s'_m | s'_e, s_h, s_m, a_p, a_s) \cdot \\ & P(a'_p | s'_m, s'_e, s_e, s_m, a_p, a_s) \end{aligned} \quad (7)$$

Figure 3 shows the influence diagram depiction of our proposed model. We assume conditional independence as follows: The first term in (7), which we call the *participants' engagement model* T_{S_e}, indicates how the robot engages in the current dominant conversation at

each time-step. We assume that the participants' engagement state at each time-step depends only on the previous engagement state, the participants' action, and the system action. In this paper, the *participants' engagement model* only contains the robot's engagement states because it is sufficient for the obtaining initiatives procedures . Table 3 shows the states of engagement.

$$T_{S_e} = P(s'_e|s_e, a_p, a_s) \tag{8}$$

In this paper, the probabilities of (8) were handcrafted, based on the consideration in Section 2.2 and our experiences. When the engagement state is the *Un-Engaged* state and the robot is asked by a current speaker, the state should be changed to the *Pre-Engaged* state, where the robot is awaiting the speaker's approval for the *Engaged* state. We assume that any dialogue acts from the speaker addressing the robot in the *Pre-Engaged* are approvals. Otherwise, the state will be back to the *Un-Engaged*. The *Engaged* state gradually goes down to the *Un-Engaged* state in time-steps unless the robot selects any dialogue acts.

We call the second term the *participants' motivation model* T_{S_m}. It indicates how an *Un-Engaged* participant has the motivation to take the floor at each time-step. This state implies that the participant who is left behind (target person) has a motivation to speak on the current topic. Thus, this state affects decision-making about topic shift. We assume that a participant's motivation at each time-step depends only on the previous system action. The motivations are defined as an unengaged participant's ID and a binary (true/false) variable, which is calculated by (3) .

$$T_{S_m} = P(s'_m|a_s) \tag{9}$$

We call the third term the *participants' action model* T_{A_p}. It indicates what actions the participants are likely to take at each time-step. We assume the participants' actions at each time-step depends on the previous participant's action, the previous system action, and the current robot's engagement state. As shown in Table 5, participants' actions include adjacency pair types. Understanding adjacency pairs is essential to detecting cut-in timing. In this paper, we recognize the adjacency pairs only by keyword matching using the results of speech recognition.

$$T_{A_p} = P(a'_p|s'_h, a_p, a_s) \tag{10}$$

The transition probabilities of adjacency pair types are based on a corpus we collected. We recorded two four-participant conversational groups (all participants were human subjects), where they were talked about movies. The total duration was around 60 minutes. Each utterance is segmented automatically by our speech recognition. After the recording, adjacency pair types were manually annotated for all speech segments.

We define the observation probability Z as follows:

$$Z = P(o'|s', a) = P(o'|s'_m, a'_p, a_s) \tag{11}$$

Table 3: Engagement states S_e

Engagement states	Meaning
Un-Engaged	The robot is not engaging with the current conversation.
Pre-Engaged	The robot is waiting for approval to engage with the current conversation.
Engaged	The robot is engaging with the current conversation.

Table 4: Motivation states S_m

Motivation states	Meaning
Motivated	The participant who is left behind has a motivation to speak on the current topic (interested in the current topic).
Not-Motivated	The participant who is left behind does not have any motivation to speak (not interested in the current topic).

Given the definitions above, the belief state can be updated at each time-step by substituting (8), (9), and (10) into (4).

$$b'(s'_m, a'_p) = \gamma \cdot \underbrace{P(o'|s'_m, a'_p, a_s)}_{\substack{\text{observation}\\ \text{model}}} \cdot \underbrace{P(s'_m|a_s)}_{\substack{\text{motivation}\\ \text{model}}} \cdot \underbrace{P(a'_p|s'_e, a_p, a_s)}_{\substack{\text{participants'}\\ \text{action model}}} \cdot \sum_{s_h} \underbrace{P(s'_e|s_e, a_p, a_s)}_{\substack{\text{engagement}\\ \text{model}}} \cdot b(s_m, a_p) \tag{12}$$

Table 6 shows the system actions. The system has seven actions available.

On the basis of the consideration of the constraints in Section 2.2, the reward measure includes components for both the appropriateness and inappropriateness of the robot's behaviors.

As an optimization algorithm, we employed the heuristic search value iteration (HSVI) algorithm proposed by Smith et al., which is one of point-based algorithms (Smith and Simmons, 2012).

4 System Architecture

Based on the studies on small group maintenance, we propose an architecture for conversational robots that has the capability to facilitate small groups, as shown in Figure 4. The framework primarily comprises Situation Analysis, Dialogue Management, and Sentence Generation processes.

4.1 Situation Analysis and Dialogue Management

Each time the system detects an endpoint of speech from the automatic speech recognition (ASR) module, it interprets the current situation. The Situation Analysis process includes participation roles recognition, adjacency pair part recognition, and question analysis.

Participation roles including a speaker, an addressee, and side-participants are recognized by the results of voice activity detection (VAD) and face directions recognition. The face directions are captured by depth-RGB cameras (Microsoft Kinect). In this paper, we use a hand-crafted role classifier. The speaker classification accuracy is 75.1% and the addressee classifi-

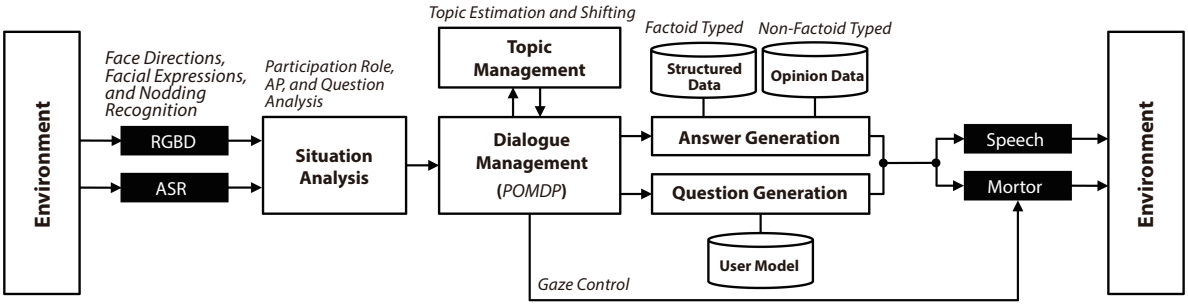


Figure 4: The architecture of the system primarily comprises the Situation Analysis, the Dialogue Management, and the Sentence Generation processes. The Situation Analysis process receives sensory information from RGBD cameras (Microsoft Kinect) and speech recognizers for each participant. The Dialogue Management process is described in Section 3. The Answer Generation process has the capability of doing additional phrasing with the robot's own opinions.

Table 5: Participants' actions A_p

Participants' actions	Meaning
first-part	A participant made an adjacency part (question)
second-part	A participant made a second adjacency part (answer)
third-part	A participant made a third adjacency part
other	A participant asked or answered the other participant
call	A participant called the robot's name

Table 6: System actions A_s

System actions	Meaning
answer	Answering the current speaker's question
question-new-topic	Asking someone a question related to a new topic
question-current-topic	Asking someone a question related to the current topic
trivia	Giving a trivia
simple-reaction	Reacting simply
nod	Nodding to the current speaker
none	Doing nothing

cation accuracy is 67.2% Adjacency pairs are recognized by the results of the participation role recognion and speech recognition. We use a hand-crafted adjacency pairs classifier. The classification accuracy is around 60%, which mostly depends on the classification accuracy of addressing for a robot. In the question analysis process, a speech utterance is interpreted with question types (5W1H interrogatives: e.g., "who," "what," "how," etc.) and predicate (verbs and adjectives). Questions are classified into two categories: Factoid type questions and Non-factoid type questions.

In the Dialogue Management process, a dialog action is selected based on abstracted conversational situation to maintain a small group, which we described in Section 3.

4.2 Sentence Generation

The Sentence Generation process consists of two components: Answer Generation and Question Generation. Based on the results of the Question Analysis process, answers are classified into two types: Factoid type answers and Non-factoid type answers (opinions). Factoid answers are generated from a structured database. In this research, we use Semantic Web technologies. After analyzing a question, it is interpreted as a SPARQL query, a resource description framework (RDF) format query language to search RDF databases. We use DBpedia as an RDF database[2].

The opinion (non-factoid type answers) generation process refers opinion data automatically collected from a large amount of reviews in the Web. The opinion generation consists of four process: document collection, opinion extraction, sentence style conversion, and sentence ranking. As an example task, we collected review documents from the Yahoo! Japan Movie site [3].

The opinion extraction consists of two processes: extraction of evaluative expressions and classification of their sentiment polarities (positive/negative). We eliminate opinions with negative sentiments because the system is expected to talk about positive contents in our conversational task. Nakagawa et al. (Nakagawa et al., 2008) used both a subjective evaluative dictionary (Higashiyama et al., 2008) and an evaluative noun dictionary (Kobayashi et al., 2007). We use an evaluative word dictionary we prepare based on their works. In order to extract evaluative expressions which can appear at any position in a sentence, we use the IOB encoding method, which has been commonly used for extent-identification tasks (Breck et al., 2007). Using IOB, each word is tagged as either (B)eginning an entity, being (I)n an entity, or being (O)utside of an entity. Based on the proposed method by Nakagawa et al, we use linear-chain conditional random fields (CRF) for the IBO encoding.

In order to preserve consistency of system's character, sentence styles are converted based on a hand-craft rule we prepare. After Japanese morphological analysis, punctuation marks and particular symbols and are eliminated. Then the last morpheme is converted.

We propose three ranking algorithms in terms of length and novelty: *Short*, *Standard* and *Diverse*. The *Short* is short length first algorithm. In this algorithm,

[2]http://ja.dbpedia.org/

[3]http://movies.yahoo.co.jp

at first, top 30% of sentences by TF-IDF score, which consists of seven to ten morphemes, are extracted. We assume top 30% of candidates is reasonably associated with a current topic. For the *Standard* and *Diverse* algorithms, at first, top 30% of sentences by TF-IDF score, which consists of fifteen to twenty morphemes, are extracted. The *Standard* algorithm is expected to contain substantial opinions or reasons, which can appeal to users about a certain topic. In this algorithm, the list is sorted by adjective term frequency. The *Diverse* algorithm is expected to express opinions or reasons with novel styles, which can be unpredictable or sometimes serendipitous to users about a certain topic. In this algorithm, the list is sorted in the inverse order by adjective term frequency.

4.3 Question Generation and User Model

The Question Generation module has two main functions: giving someone the floor and collecting users' preferences and experiences for the User Model.

The User Model is preferred for topic maintenance. A preferred new topic is decided using cosine similarity of TF-IDF scores. The topic scores (*TopicScore*) of all topics are calculated based on cosine similarities of the current topic (*CurrentTopic*), a user's topic preferences of all topics (*PreferenceTopic*), and experiences (*ExperienceTopic*) between the *CurrentTopic* and each *Topic*.

$$\begin{aligned} TopicScore_i = {} & \alpha \cos(Topic_i \cdot CurrentTopic) \\ & + \beta \left(\sum_m \cos(Topic_i \cdot PreferenceTopic_m) \right) \\ & + \gamma \left(\sum_n \cos(Topic_i \cdot ExperienceTopic_m) \right) \end{aligned} \quad (13)$$

4.4 Experimental Platform

For our experimental platform, we used the multimodal conversation robot "SCHEMA([ʃe:ma])," (Matsuyama et al., 2009) shown in Figure 2. SCHEMA is approximately 1.2[m] in height, which is the same as the level of the eyes of an adult male sitting down in a chair. It has 10 degrees of freedom for right-left eyebrows, eyelids, right-left eyes (roll and pitch) and neck (pitch and yaw). It can express anxiousness and surprise using its eyelids and control its gaze using eyes, neck, and autonomous turret. In addition, it has six degrees of freedom for each arm, which can express gestures. One degree of freedom is assigned to the mouth to indicate explicitly whether the robot is speaking or not. A computer is inside the belly to control the robot's actions, and an external computer sends commands to execute various behaviors though a WiFi network. All modules, including the ASRs and a speech synthesizer are connected to each other though a middleware called the Message-Oriented NEtworked-robot Architecture (MONEA), which we earlier produced (Nakano et al., 2006).

5 Experiments

We designed the following two experiments to evaluate the appropriateness and feeling of groupness of our proposed procedures for multiparty conversations (**experiment 1**), and the appropriateness of timing for initiating procedures (**experiment 2**). In order to cancel the effects of recognition errors, we prepared video recordings of four-participant situations (Human participant A, B, C, and a robot), just like 2. We created the following three conditions, all of which are optimized as POMDP. All subjects were native Japanese speakers recruited from Waseda University campus. They were first given a brief description of the purpose and the procedure of the conversation. They were instructed that A and B have a friendly relationship with each other, C is coming in for the first time and is feeling nervous, therefore, C is left behind in the conversation, and a robot is trying to maximize the total engagement of this situation. We also explained "a engaged situation" meant "a situation in which all participant are given their opportunities to speak something fairly."

5.1 Experiment 1: Appropriateness and Groupness by Usage of Procedures

A total of 35 subjects (23 males and 12 females) participated in this experiment. The ages of the subjects ranged between 20 and 25 years with an average age of 20.5 years. After they watched the videos, they were asked to complete questionnaires about their feeling of groupness ("For which condition did you feel a sense of groupness?") and free-form questionnaires. The following four conditions were videotaped, and the video edited at around 30 s. All videos contained the same topic ("*Princess Mononoke*"). The spatial arrangement was the same as shown in Figure 2.

Condition 1: Without procedures (without topic shifting). A robot directly asks a participant left behind without procedures claiming an initiative. As is shown in Figure 8, just after a sequence of interactions between A and B, which is segmented by a third adjacency pair part, a robot directly asks C. The topic is still maintained ("*Princess Mononoke*").

Condition 2: With procedures (without topic shifting). A robot directly asks a participant left behind with a procedure. As is shown in Figure 9, Just after a sequence of interactions between A and B, a robot asks A with the first pair part and waits for A's response (the second part). Then it finishes the interaction with A, and asks C to give a floor. In this case, topic is still maintained the current one ("*Princess Mononoke*").

Condition 3: Without procedures + topic shifting. In #6 question of Condition 1 (Figure 8), a robot initiates a new topic ("*From Up On Poppy Hill*") instead.

Condition 4: With procedures + topic shifting. In #7 question of Condition 2 (Figure 9), a robot initiates a new topic ("*From Up On Poppy Hill*") instead.

After watching the movies, they were requested to answer Likert 7-scaled questionnaires about (a) **appro-**

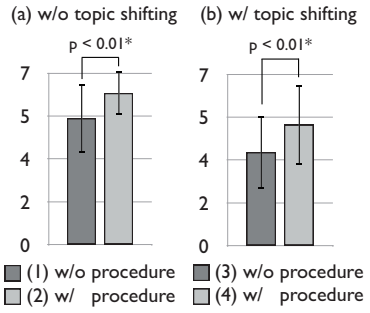


Figure 5: Result of experiment 1-a

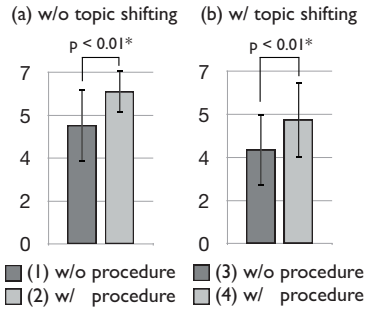


Figure 6: Result of experiment 1-b

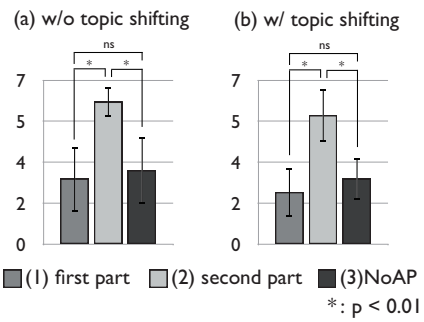


Figure 7: Result of experiment 2

priateness of procedures, (b) **Feeling of groupness**.

5.2 Experiment 2: Appropriateness of Timing of Initiating Procedures

A total of 32 subjects (21 males and 11 females) participated in this experiment. The ages of the subjects ranged between 20 and 25 years with an average age of 20.5 years. After they watched the videos, they were asked to complete questionnaires about the timing of the initiating procedures ("Which video did you feel was the most appropriate?"). The following three conditions were videotaped, and edited at around 30 s. All videos contained the same topic ("*Princess Mononoke*"). The spatial arrangement was the same as shown in Figure 2. We created three conditions:

Condition 1 (first part): Initiating a procedure just after the first adjacent pair part.

Condition 2 (second part): Initiating a procedure just after the second adjacent pair part.

Condition 3 (No AP): Out of consideration of adjacency pairs.

In conditions 1 and 2, the robot initiated its procedures just after the first and second parts, respectively. In condition 3, the robot initiated its procedure in the middle of the adjacency pairs, which is intended to show that the robot does not care about adjacency pairs. We did not consider the timings of the third part of the adjacency pair because we had already examined the appropriateness of the timing of the third part in experiment 1. After watching the movies, they were requested to answer Likert 7-scaled questionnaires about the robot's **appropriateness of behavior**.

5.3 Results and Discussions

Figure 5 shows usages of procedures are appropriate to approach a participant left behind either with or without topic shifting. The t-test result shows a significant difference between condition 1 and 2, as well as between 3 and 4 ($p < 0.01$). Figure 6 shows usages of procedures generate feelings of groupness. The t-test result also shows a significant difference between condition 1 and 2, as well as between 3 and 4 ($p < 0.01$).

Figure 7 (a) shows initiating procedures without topic shifting in timings of just after the second pair parts is more appropriate than other conditions. The result of an analysis of variance (ANOVA) shows significant differences among conditions ($F[2,26] = 34.46$, $p < 0.01$). The result of multiple comparisons with the Tukey HSD method shows a significant difference between condition 1 and 2, as well as between 2 and 3 ($p < 0.01$). Figure 7 (b) shows initiating procedures with topic shifting in timings of just after the second pair parts is more appropriate than other conditions. The result of an analysis of variance (ANOVA) shows significant differences among conditions ($F[2,26] = 42.52$, $p < 0.01$). The result of multiple comparisons with the Tukey HSD method shows a significant difference between condition 1 and 2, as well as between 2 and 3 ($p < 0.01$).

From these results, usages of procedures obtaining initiatives before approaching a participant left behind showed evidences of acceptability as a participant's behaviors, and feeling of groupness in a group. As for timings, initiating the procedures just after the second or third adjacency pair parts is felt more appropriate than the first pairs by participants.

6 Conclusions

We proposed a framework for conversational robots facilitating four-participant groups. Based on a representation of conversational situations, we presented a model of procedures obtaining conversational initiatives in incremental steps to maximize total engagement of such four-participant conversations. These situations and procedures were modeled and optimized as a partially observable Markov decision process. As the results of two experiments, usages of procedures obtaining initiatives showed evidences of acceptability as a participant's behaviors, and feeling of groupness. As for timings, initiating the procedures just after the second or third adjacency pair parts is felt more appropriate than the first pairs by participants.

7 Acknowledgements

This research was supported by the Grant-in-Aid for scientific research WAKATE-B (23700239). TOSHIBA corporation provided the speech synthesizer customized for our spoken dialogue system.

References

Robert F Bales. 1950. Interaction process analysis. *Cambridge, Mass*.

Kenneth D Benne and Paul Sheats. 1948. Functional roles of group members. *Journal of social issues*, 4(2):41–49.

Dan Bohus and Eric Horvitz. 2009. Models for multiparty engagement in open-world dialog. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–234. Association for Computational Linguistics.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2683–2688. Morgan Kaufmann Publishers Inc.

Crystal Chao and Andrea Lockerd Thomaz. 2012. Timing in multimodal turn-taking interactions: Control and analysis using timed petri nets. *Journal of Human-Robot Interaction*, 1(1).

Herbert H Clark. 1996. *Using language*, volume 4. Cambridge University Press Cambridge.

Kohji Dohsaka, Ryota Asai, Ryuichiro Higashinaka, Yasuhiro Minami, and Eisaku Maeda. 2009. Effects of conversational agents on human communication in thought-evoking multi-party dialogues. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–224. Association for Computational Linguistics.

Erving Goffman. 1981. *Forms of talk*. University of Pennsylvania Press.

Masahiko Higashiyama, Kentaro Inui, and Yuji Matsumoto. 2008. Acquiring noun polarity knowledge using selectional preferences. In *Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing*, pages 584–587.

Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. 2007. Opinion mining from web documents: Extraction and structurization. *Information and Media Technologies*, 2(1):326–337.

Rohit Kumar, Jack L Beuth, and Carolyn P Rosé. 2011. Conversational strategies that support idea generation productivity. In *in Groups, 9 th Intl. Conf. on Computer Supported Collaborative Learning, Hong Kong 160 and Rosé, 2010a) Rohit Kumar, Carolyn P. Rosé, 2010, Conversational Tutors with Rich Interactive Behaviors that support Collaborative Learning, Workshop on Opportunity*. Citeseer.

James R Martin and Peter RR White. 2005. *The language of evaluation*. Palgrave Macmillan Basingstoke and New York.

Yosuke Matsusaka, Tojo Tsuyoshi, and Tetsunori Kobayashi. 2003. Conversation robot participating in group conversation. *IEICE transactions on information and systems*, 86(1):26–36.

Yoichi Matsuyama, Hikaru Taniyama, Shinya Fujie, and Tetsunori Kobayashi. 2008. Designing communication activation system in group communication. In *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, pages 629–634. IEEE.

Yoichi Matsuyama, Kosuke Hosoya, Hikaru Taniyama, Hiroki Tsuboi, Shinya Fujie, and Tetsunori Kobayashi. 2009. Schema: multi-party interaction-oriented humanoid robot. In *ACM SIGGRAPH ASIA 2009 Art Gallery & Emerging Technologies: Adaptation*, pages 82–82. ACM.

Yoichi Matsuyama, Shinya Fujie, Hikaru Taniyama, and Tetsunori Kobayashi. 2010. Psychological evaluation of a group communication activation robot in a party game. In *Eleventh Annual Conference of the International Speech Communication Association*.

Tetsuji Nakagawa, Takuya Kawada, Kentaro Inui, and Sadao Kurohashi. 2008. Extracting subjective and objective evaluative expressions from the web. In *Universal Communication, 2008. ISUC'08. Second International Symposium on*, pages 251–258. IEEE.

Teppei Nakano, Shinya Fujie, and Tetsunori Kobayashi. 2006. Monea: message-oriented networked-robot architecture. In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 194–199. IEEE.

Antoine Raux and Maxine Eskenazi. 2009. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 629–637. Association for Computational Linguistics.

Emanuel A Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327.

Candace L Sidner, Cory D Kidd, Christopher Lee, and Neal Lesh. 2004. Where to look: a study of human-robot engagement. In *Proceedings of the 9th international conference on Intelligent user interfaces*, pages 78–84. ACM.

Trey Smith and Reid Simmons. 2012. Point-based pomdp algorithms: Improved analysis and implementation. *arXiv preprint arXiv:1207.1412*.

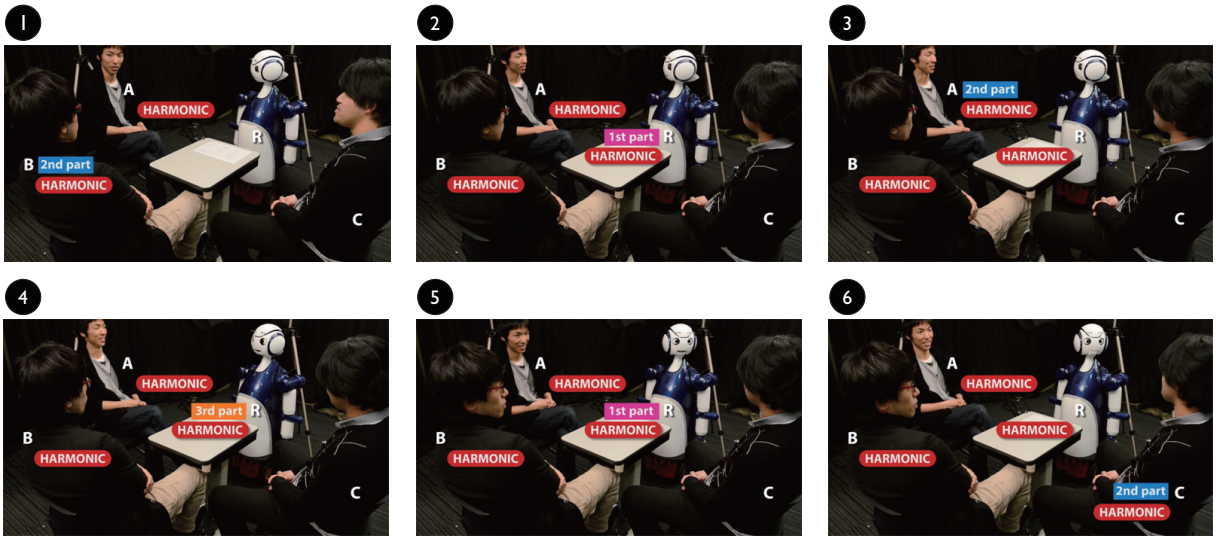
Jason Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

#	SPK → ADD	AP	Sentences
1	**A→B**	First	Have you ever watched “Princess Mononoke”?
2	**B→A**	Second	Yes, I have
3	**A→B**	First	Oh, you have?
4	**B→A**	Second	Yeah.
5	**A→B**	Third	I see
6	**R→C**	First	**Have you ever watched “Princess Mononoke”?**
7	**C→R**	Second	Yes, I have

Figure 8: Transcript of condition 1 (experiment 2)

#	SPK → ADD	AP	Sentences
1	**A→B**	1st	Have you ever watched “Princess Mononoke”?
2	**B→A**	Second	Yes, I have
3	**A→B**	Third	I see.
4	**R→A**	First	**It is one of my favorite movies among Ghibri’s**
5	**A→B**	Second	Really?
6	**B→A**	Third	Yes.
7	**R→C**	First	**Have you ever watched “Princess Mononoke”?**
8	**C→R**	Second	Yes, I have

Figure 9: Transcript of condition 2 (experiment 2)



#	SPK → ADD	AP	S_e	Sentences
				(Topic: “007 Skyfall”)
1	**A→B**	**1st**	*Un*	Let’s talk about the “Skyfall.”
2	**A→B**	**1st**	*Un*	Have you ever seen the latest one?
3	**B→A**	**2nd**	*Un*	Well, I’ve not seen that. ····· 1
4	**A→B**	**3rd**	*Un*	Oh, really.
5	**R→A**	**1st**	*Pre*	Well, I like the Bond Girl. ·····
6	**A→R**	**2nd**	*Pre*	I see.
7	**R→A**	**1st**	*Pre*	I think that movie is good because of the setting of the ”old age” for the 44-year old James Bond. ····· 2
8	**A→R**	**2nd**	*H*	Uh-huh. ····· 3
				(R is approved to obtain an initiative)
9	**R→A**	**3rd**	*H*	Yes. ····· 4
10	**R→C**	**1st**	*H*	Have you ever seen the ”Skyfall”? ····· 5
11	**C→R**	**2nd**	*H*	No, I haven’t. ····· 6
12	**A→C**	**1st**	*H*	Oh, you haven’t seen it?
13	**C→A**	**2nd**	*H*	I never seen that before.

Figure 10: Interaction scenes. The “AP” signifies adjacency pair types. At **#4**, the system recognized A’s adjacency third part and then generated a spontaneous opinion addressed to A (#5) as the first part. At that point, the system assumed the state of engagement (S_e) had changed from *Un-Engaged* to *Pre-Engaged*. After the system observed A’s second part at **#8**, it assumed it at gotten approval to obtain an initiative to control the context (*Engaged*). At **#10**, the robot asked C a question in order to give him the floor.