Turn-Taking in Commander-Robot Navigator Dialog

T. Cassidy and **C. Voss** and **D. Summers-Stay**
Computational and Information Sciences Directorate
Army Research Laboratory
Adelphi, MD

Abstract

We seek to develop a robot that will be capable of teaming with humans to accomplish physical exploration tasks that would not otherwise be possible in dynamic, dangerous environments. For such tasks, a human commander needs to be able to communicate with a robot that moves out of sight and relays information back to the commander. What is the best way to determine how a human commander would interact in a multi-modal spoken dialog with such a robot to accomplish tasks? In this paper, we describe our initial approach to discovering a principled basis for coordinating turn-taking, perception, and navigational behavior of a robot in communication with a commander, by identifying decision phases in dialogs collected in a WoZ framework. We present two types of utterance annotation with examples applied to task-oriented dialog between a human commander and a human "robot navigator" who controls the physical robot in a realistic environment similar to expected actual conditions. We discuss core robot capabilities that bear on the robot navigator's ability to take turns while performing a "find the building doors" task at hand. The paper concludes with a brief overview of ongoing work to implement these decision phases within an open-source dialog management framework, constructing a task tree specification and dialog control logic for our application domain.

1 Introduction

Our goal is to build a robot that is both mobile and communicative, and functions as a useful team member with humans to explore potentially hazardous environments. This paper focuses on the question of when and how it is appropriate for such a robot to begin a new turn in a dialog. We define a dialog turn as a series of utterances and/or paralinguistic moves (e.g. sending a picture message from a video camera), or a long pause.

Given that such a robot does not yet exist, our broad approach is as follows:[1]

1. Record and analyze spoken dialog between a human commander (C) and a human "robot-navigator" (RN), charged with navigating a physical robot (R) to perform well-defined tasks, as well as the robot's corresponding perceptual and motoric sensor data. Determine core robot capabilities (e.g. object recognition and tracking) needed to automate RN's coordination of dialog, perceptual, and navigational behaviors.
2. Build a dialog manager (DM) that draws on core capabilities to automate turn-taking patterns observed in data. Both core and dialog capabilities may be automated, or isolated and emulated and performed by one or more Wizards.
3. Evaluate that system for its impact on C's utterances, turn-taking, and C's effectiveness in directing and completing the task by way of the DM.

This paper presents progress made on the first part. We present a new annotation scheme for dialog turns specific to the human-robot navigation domain, and apply it to previously collected task-oriented, multi-modal dialogs between C and RN (wherein the latter controlled the robot with a joystick). We follow the long-standing approach of (Sacks, Schegloff, and Jefferson 1974), as well as the similar, recent approach of (Thomasz and Chao 2011) in factoring apart the domain–independent organization of dialog structure from the domain-specific aspects of the expressed dialog content. We discuss how RN coordinates dialog, perceptual, and navigational behaviors in an annotated dialog excerpt, as well as what core robot capabilities could be brought to bear on automating this coordination.

The task performed in the collected dialogs was to explore a series of buildings and courtyards, identify all doors of all the buildings in an assigned region, and transmit images of those doors. The robot traveled on four wheels and was equipped with a video camera and occupancy grid built live during the exploration. C and RN were told that doors of courtyards were *not* to be counted. This caveat required RN and C to communicate to ensure C had seen each door in sent images with enough surrounding area to verify seeing a building door.[2]

[1] This design falls under the Wizard of Oz (WoZ) framework. Note C is aware that a human RN controls the robot, whereas most WoZ studies attempt to convince a human that a robot is autonomous.

[2] Eight complete dialogs (and accompanying data) were collected for this "find the building doors" task with four participants, totalling 103 minutes. See (Summers-Stay, Cassidy, and Voss 2014) for details of study and corpus. The format of this study

The following properties characterize the setting in which C and RN interacted:

1. C and RN speak to each other via an audio channel but are at different locations, i.e., not face-to-face
2. C and RN neither see nor hear the physical robot, but C knows RN is navigating R.
3. RN can, with the push of a button, transmit to C, individual snapshot images from R's onboard streaming video camera or 2-D map updates constructed from R's onboard streaming LIDAR

We took a data-driven approach to design our annotation schema. Our aim was to identify phases within the task that correlate with observed speech and behavior patterns. Two of the authors annotated dialogs.

2 Annotations

Decision Phases

Figure 1 traces the sequence of decisions in the dialog phase annotation starting from the top arrow. Initially C must decide (phase 0) whether they need further location information (move to phase 1) or are ready to issue a command (move to phase 2), after which RN decides when to act (decision phase 3) and stop (proceeding to phase 3-inc when continued execution is not feasible, not safe, or not consistent; or transitioning to phase 4 when done executing command). Interruptions by C or RN may occur at any time in dialog. The dashed lines within a decision phase indicate possible dialog turns before achieving the end state.

Tasks and Cycles

We distinguish two high-level tasks within each dialog of our collection and annotate each turn as belonging to Task "Sight-Door" (S) or "Verify-Door" (V). We indicate that a turn belongs to S when it is part of the robot's exploration of the environment while both C and RN are looking for a new building door in the images and LIDAR maps, but have not yet explicitly said either that (i) they have seen one or that (ii) they think, or are not quite sure, they may have seen one.

Once a door is sighted or the speaker (C or RN) believes they may have seen evidence for such a sighting and, crucially, they say so explicitly in an utterance turn, then all subsequent turns are annotated as belonging to V until C accepts an image of the (possible) door and describes it. To facilitate the analysis of a dialog we also label each block of adjacent S turns followed by a block of V turns a "Cycle".

An example of our different annotations appears in Figure 3 on a portion of a longer original full dialog that contained six full cycles and one partial cycle (due to the end of the recording session).

3 Analysis and Discussion

The ability to engage in natural turn-taking is a prerequisite for a useful, mobile, and communicative robot team member. Simply identifying all building doorways could be

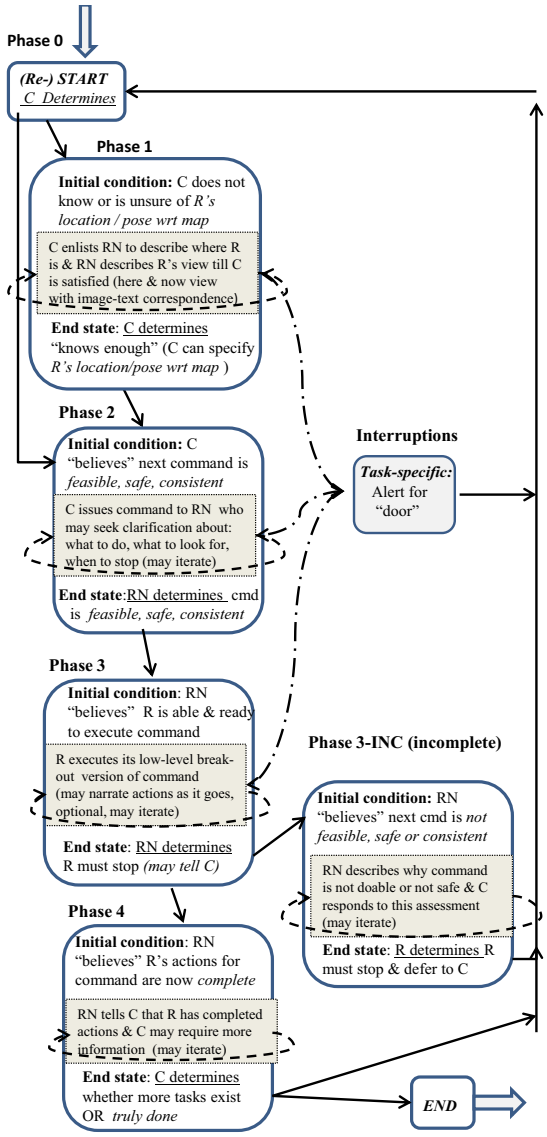

Figure 1: Decision Phases

accomplished using a non-interactive drone. However, in an emergency scenario we need robots to relay information about dynamic environments in a timely, natural way so that we can adjust their search and exploration tasks accordingly. The need for human-like interaction necessitates:

1. Correctly timing reports of perceptual phenomena.
2. Quickly processing perceptual data to identify a variety of perceptual phenomena with accurate confidence values.
3. Linguistic ability to generate important information about events in a natural way.
4. The ability to choose the correct modality for conveying information (speech, image, map)
5. The ability to convey uncertainty with respect to classification of perceptual phenomena.

was motivated by pre-pilot results described in (Voss, Cassidy, and Summers-Stay 2014).

6. The ability to keep track of multiple objects of the same type while moving throughout a large scale environment.

The following walk through of the dialog in Figure 3 contains parenthetical allusions to the related capabilities enumerated above. (The utterance number is leftmost on each line, the phase number corresponding to a phase Figure 1 follows on the right. Task and Cycle numbers for each utterance appear on the right side.)

In lines 14-16 RN interrupts execution of a command upon recognizing a doorway. Because they are in S there is a premium on reporting new door sightings and the interruption is therefore appropriate (1). Had they already been pursuing a door sighting (V) RN would likely have waited to point out the new sighting. Note that RN failed to notice the door immediately upon passing it (2) and was therefore presented with the non-trivial task of adequately situating the sighting as a past event (3). The longer it takes RN to be confident enough in a door sighting, more events might unfold making it difficult to naturally explain the sighting.

In lines 17-22 they complete CYCLE 1, and C prompts RN to indicate the command has been completed in phase 4 by means of transmitting a map and picture (4). The command in lines 23-24 does not provide sufficient evidence that C intends to home in on a particular doorway sighting (contrast this command with "... and see if *that's* another door on the side of building 2", which would indicate C viewed a particular map region as a potential doorway sighting). RN then conveys a level of uncertainty in 26 through hedging. When the distaste for giving false alarms must be traded for timeliness, it is crucial to be able to convey uncertainty in this way (2).

C begins the next cycle by directing RN to explore a frontier C deems likely to contain doors. Here, it would be appropriate to call out a door sighting not in the area specified by C ("on your left") since they are still in S, but RN should decide whether to first check the area picked out by C before making any such reports (2).

When C says "the indentation" in 36 he speaks in terms of a map feature, whereas the participants normally speak in terms of what map features depict (e.g. "wall" instead of "line"). Deciding whether to generate or understand language in terms of media qua media or in terms of what media depicts is on one hand orthogonal to the turn-taking problem. However, participants did sometimes speak in these terms, and a robust system that accommodates this way of speaking should vary its approach to clarification and elaboration accordingly. Whether toggling between these modes of reference would aide a DM in disambiguation is a matter of future research (5).

In line 38 RN indicates that the command in line 36 has been completed by declaring that RN (as R) has entered a courtyard. While there are a variety of ways to express R's state at that point, note that prior to the two doorway sightings (lines 16, 26) and subsequent classifications R was instructed to "follow [a] wall through the courtyard doorway" (see line 14), indicating that the act of entering a courtyard would be a good reason to seize the floor (1). Finally in lines 39-41 RN must infer that C is likely to be unsure of R's location, thus prompting a shift to phase 1 during the next turn in which RN offers help in the form of media information (4).

4 Ongoing and Future Work

Currently our ongoing work focuses on implementing the decision phases from Figure 1 within an open-source dialog management framework, by constructing a task specification and dialog control logic for our application domain. In particular we draw on the RavenClaw dialog management framework (Bohus and Rudnicky 2009) and ideas (Raux and Eskenazi 2007) for combining multi-modal information from the real-world via an Interaction Manager.

We plan to operationalize our decision phases and task models as part of a RavenClaw (Bohus and Rudnicky 2009) Task Tree. The RavenClaw dialog management architecture separates task specification from dialog execution. The former uses a task tree whose nodes are agents that perform subtasks, and whose structure dictates the canonical dialog state sequencing. The latter is a "dialog engine" that traverses the task tree to instantiate agents, organizing them in a stack that determines what sort of content is expected from the user and in what order it is expected. In practice, task tree agents *per se* are leaf nodes, while higher order "agency" nodes are used to impose structure on the dialog flow.

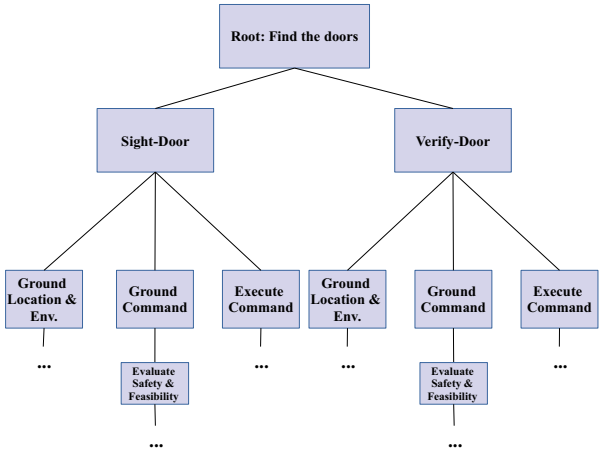

Figure 2: RavenClaw Task Tree

Figure 2 shows an initial sketch of our task tree. The dialog engine traverses the tree from left to right, thus starting in the S task. Phases 1-4 (see Figure 1) will be mapped to agencies. Specific actions and associated dialog subtasks for a given phase, including the content in beige boxes in Figure 1 as well as the roles played by phase 0 and the Interruptions phase, will be implemented as agents. (In Figure 2 agents are currently represented by ellipses). For example, the Ground Location and Environment agency under S would contain an agent that handles grounding the location of an initial door sighting, before jumping to the V subtask. We anticipate that most agents will have the ability to react to new door-sightings. Note that in principle, the dialog en-

	Decision Phases				Who	Utterance/info transfer	Task	Cycle
14		2			C	ok go forward until you reach the wall and then turn left and follow that wall forward through the courtyard doorway	S	1
15		2			R	ok I will try to do that	S	1
			3		*R*	**(Navigating)**	S	1
16	interruption				R	I think I passed a **door** a little bit ago on my left side	V	1
17		2			C	ok, go back to the door facing the door	V	1
18		2			C	and send me a picture	V	1
19		2			C	and also send me an updated map	V	1
			3		*R*	**(Navigating)**	V	1
20				4	*R*	**<<Picture & Map Transmitted>>**	V	1
				4		ok,		
21	1				C	I can see where we are now	V	1
22	1				C	it looks like we are facing a door on building	V	1
23		2			C	ok turn right and follow that wall around the corner	S	2
24		2			C	and see if there's another door on the side of building 2	S	2
25		2			R	ok	S	2
			3		*R*	**(Navigating)**		2
26	interruption				R	I'm detecting what may be a **door**	V	2
27		2			C	ok, turn and face it	V	2
28		2			C	and send me a photo	V	2
29		2			R	ok	V	2
30				4	*R*	**<<Picture Transmitted>>**	V	2
31	1				C	it looks like a closed door	V	2
32		2			C	why don't you back up a few body lengths	V	2
33		2			C	and take another shot	V	2
			3		*R*	**(Navigating)**	V	2
34				4	*R*	**<<Picture Transmitted>>**	V	2
35	1				C	yeah, that's a closed door on building 2	V	2
36		2			C	ok turn um 180 degrees and go forward uh past the indentation on your left	S	3
37		2			C	looking for a door on your left	S	3
			3		*R*	**(Navigating)**	S	3
38			INC		R	I think I've just entered through a co..through a.. into a courtyard	S	3
39	interruption				R	do you want me to send you an updated	S	3
40	1				C	yeah,	S	3
41	1				C	and send me another picture too	S	3
42	1				*R*	**<<Picture & Map Transmitted>>**	S	3

Figure 3: Dialog Turns with Decision Phase (1–4), Speaker, Utterance/Info Transfer, Tasks (Search, Verify), Cycles

gine has the ability to push agencies to the stack from any part of the task tree. In this effort we aim to construct a DM that properly models the sequence of perception, action, and communication events.

In future work, we will explore another cue of significance in turn-taking that goes beyond event sequencing: precise timing. All speech and robot sensor data are aligned with time stamps, which will allow us to collect the timing of turns when information is exchanged, both non-verbal (image and map) and verbal in our dialogs.

Acknowledgments

We would like to thank Matthew Marge for his inciteful comments.

References

Bohus, D., and Rudnicky, A. I. 2009. The ravenclaw dialog management framework: Architecture and systems. *Computer Speech & Language* 23(3):332–361.

Raux, A., and Eskenazi, M. 2007. A multi-layer architecture for semi-synchronous event-driven dialogue management. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Sacks, H.; Schegloff, E.; and Jefferson, G. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50 (4):696–735.

Summers-Stay, D.; Cassidy, T.; and Voss, C. 2014. Joint Navigation in Commander/Robot Teams: Dialog & Task Performance When Vision is Bandwidth-Limited. In *COLING 2014 Proceedings of the Third Workshop on Vision and Language (VL'14)*.

Thomasz, A. L., and Chao, C. 2011. Turn taking based on information flow for fluent human-robot interaction. *AI Magazine* 32 (4):53–63.

Voss, C.; Cassidy, T.; and Summers-Stay, D. 2014. Collaborative Exploration in Human-Robot Teams: What's in Their Corpora of Dialog, Video, & LIDAR Messages? In *Proceedings of EACL Dialog in Motion Workshop*.