Heterogeneous Multi-Robot Dialogues for Search Tasks

Thomas Kevin Harris, Satanjeev Banerjee, and Alexander I. Rudnicky
Carnegie Mellon University
tkharris+, banerjee+, air+ @cs.cmu.edu

Abstract

Dialogue agents are often designed with the tacit assumption that at any one time, there is but one agent and one human, and that their communication channel is exclusive. We are interested in examining situations in which multiple heterogeneous dialogue agents need to interact with a human interlocutor, and where the communication channel becomes necessarily shared. To this end we have constructed a multi-agent dialogue test-bed on which to study dialogue coordination issues in multi-robot scenarios.

Research Goals

Our goals are to:

- **Understand the skills needed for communication in a human-robot team.** Aspects of such a language include the mechanisms for initiating conversations with yet unseen robots, understanding their capabilities, sharing and negotiating task allocation among robots, et cetera. Humans have little difficulty communicating such information among themselves; we would like to computationally reproduce these skills.
- **Develop languages for robot navigation in novel environments.** Novel environments require on-the-fly map building on the part of the robot. The robot must then communicate it's map to the human in terms that the human can understand. The human, in turn, needs to communicate goals that reference the robot's map. Rapid converge on mutually understood descriptions and labels is a necessary aspect of the problem.
- **Understand how novel objects, locations, and tasks come to be described in language.** Robots and humans will need to mutually ground their perceptions to effectively communicate about novel objects, locations, and tasks. Current speech understanding technologies do not easily support such interaction.

The Treasure–Hunt Scenario

We have selected a multi–robot–assisted treasure–hunt domain to pursue research towards these goals. This scenario is designed with some complications intended to provide real-world plausibility, and some constraints intended to make the problem tractable.

A "treasure–hunt" is a domain in which one is tasked to search a space for one or more objects of interest, and to bring those objects back to some specified location. It is a domain that has wide applicability in terms of real-world scenarios, encompassing warehouse retrieval, search–and–rescue, "broken arrow" situations, biological sample collection, et cetera.

Robots have a great potential to contribute to treasure–hunt domains in three ways:

- Robots are sometimes able to traverse areas that are difficult or dangerous for humans to explore. Space exploration and search–and–rescue operations obviously have a dramatically difficult and dangerous search space, but even mundane scenarios such as collecting golf-balls from the woods or oysters from a bay present spaces that provide difficulty for humans.
- Robots are able to explore without rest.
- Multiple parties of robots may be able to explore spaces quite rapidly.

Effectively conducting a treasure-hunt involves the mastery of several skills:

- Identifying and prioritizing search spaces.
- Moving or deploying sensors to cover search spaces.
- Analysing sensor data to classify objects or spaces as interesting or not.
- Capturing and transporting objects of interest.

Where robots are employed in treasure–hunts, we expect that human–robot communication will be an important aspect of the treasure–hunt. There are three reasons:

- We expect that in many treasure–hunt domains, robots will posses some but not not all of these skills. A robot that cannot identify and prioritize search spaces will need to be told where to look and will need to be able to indicate when it has finished searching in a space. A robot that cannot capture or transport objects will need to lead people (or other robots) toward the object of interest.

- While robots may posess one of these treasure–hunt skills, they might not be fully capable. Sensing, and especially vision, is one of the most problematic aspects of current robots. We expect that in many situations where robots may be able to identify an object of interest, they may need to confirm this object with a human through some kind of communication.
- While robots may posess one of these treasure–hunt skills, they might not be fully capable. Sensing, and especially vision, is one of the most problematic aspects of today's robots. We expect that in many situations where robots may be able to identify an object of interest, they may need to confirm this object with a human through some kind of communication.
- While robots may posess one of these treasure–hunt skills, they might not be fully capable. Sensing, and especially vision, is one of the most problematic aspects of today's robots. We expect that in many situations where robots may be able to identify an object of interest, they may need to confirm this object with a human through some kind of communication.
- While robots may posess one of these treasure–hunt skills, they might not be fully capable. Sensing, and especially vision, is one of the most problematic aspects of today's robots. We expect that in many situations where robots may be able to identify an object of interest, they may need to confirm this object with a human through some kind of communication.
- Even in cases where robots are fully capable in all of these skills, humans might still participate with robots as regular partners. Coordination is an inherent prerequisite for any useful multi-party search.

Although many communication modalities could be explored to exchange information between human and robot partners in a treasure–hunt scenario, we have focused on spoken language interfaces. Spoken language is a natural modality for humans and is the one most likely to be useful in the widest array of treasure–hunt situations. (For example, where mobilityt is necessary or where the environment is not illuminated.)

Problems with the Communication Channel

Multiple spoken dialogue agents will face some of the same main communication issues that have been active areas of research in the domain of computer networks, namely the issues of message identification, message addressing, channel contention, and session identification.

Message Identification With multiple spoken dialogue agents in an environment, there is the unintended potential (and sometimes the need) for dialogue agents to speak to each other. Dialogue agents will undoubtedly misbehave unless they can identify who is speaking. Speaker identification has had some success recently and may be employed to address this issue, but serious issues remain to be resolved, such as its scalability.

Message Addressing When an environment contains more than one spoken dialogue agent, each agent must resolve who a particular utterance is addressed to. Evidence from acoustic, linguistic, and pragmatic sources of knowledge, combined with additional information from other communication modalities such as gesture, gaze, and touch have been used to perform address resolution. Systems that perform such resolution however often make use of deep domain knowledge, or make the simplifying assumption that other dialogue agents in the environment have domains that are sufficiently different from its own.

Channel Contention With multiple spoken dialogue agents in an environment, there is the potential that they will speak simultaneously, or that they will interrupt each other. Methods currently employed to resolve these contentions often consist simply of waiting until nobody else is speaking for a second or so, and then to begin speaking until someone interrupts. In environments with one dialogue agent and one human interlocutor, this algorithm is usually sufficient, but studies of larger human group dynamics show that many important subtleties are missing. As well, high-priority messages must be able to interrupt on-going communications. Agents who do not indicate their desire to speak, or who do not introduce themselves before they speak are likely to not be understood by their listeners, or, at the very minimum, such agents are likely to add to the cognitive load of their listeners.

Sessions Identification Although communication between humans and agents is improved when those agents introduce themselves before speaking, they cannot introduce themselves before every utterance. A concept of a communication session must be developed.

An Experimental Multiple Agent Dialogue System

We have engaged in a systematic approach to finding efficient solutions through empirical experiments for the four communication issues identified above. In particular, we have developed a multi–agent dialogue (MAD) system, which can accommodate multiple dialogue agents in a single experimental framework (see Figure 1). The system works both with real robots adapted for the Carmen robot platform (Montemerlo, Roy, & Thrun, 2002), and in a simulated Carmen environment.

The front-end architecture is an instance of the Galaxy-II spoken dialogue system reference architecture (Seneff et al., 1998). We use Sphinx-II (Huang et al., 1993) for automatic speech recognition, Phoenix (Ward, 1994) for context-free grammar parsing, Helios (Bohus & Rudnicky, 2002) for confidence annotation, Ravenclaw (Bohus & Rudnicky, 2003) for dialogue management, ROSETTA (Oh & Rudnicky, 2000) for natural language generation, and Festival (Black, Clark, Richmond, & King, 2004) for text-to-speech rendering. The robots, named Bashful and Clyde (ghosts from Namco Ltd.'s Pac-Man®), each have their own Ravenclaw dialogue system. Ravenclaw is a generalized tree-based dialogue management framework that provides the

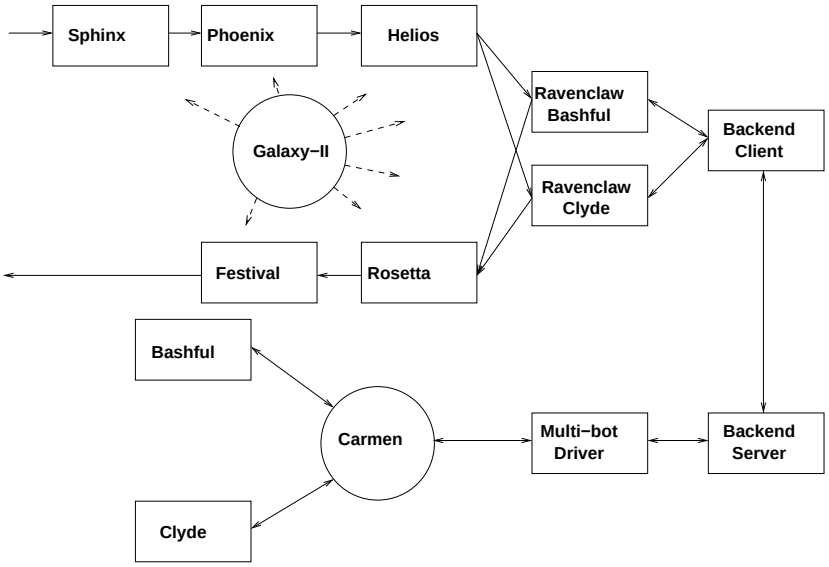

Figure 1: Multiple Agent Dialogue System

designer of a dialogue management system with mechanisms through which to specify dialogue tasks. Essentially a designer specifies the various actions that must take place in the system (e.g. the action to be taken when the user asks a robot where it is) and the flow of the dialogue.

The back-end consists of programs that use the Carmen set of libraries to communicate with the robots. The libraries currently utilized in our project include those that allow the user to send messages to the robots to get them to move a specified distance in a certain direction, and those that allow the user to set a goal position and then allow the robot to plan a route to that position.

User Studies

Goals

We have used the system described above to conduct a preliminary set of pilot studies. We have two goals in performing these studies. The first goal is to establish the *usability* of our system, that is, we ask the question: Can this system be used by a human being to successfully interact with the robots? A large system like the one above can fail to be usable for a variety of reasons: the speech recognition may be too error prone, the speech synthesis may be unintelligible, the pace of the interaction too slow, the robot navigation libraries too unreliable, etc. Our experimentation is designed to show that our system can indeed be used to interact naturally with the robots. The second goal of our studies is to experiment with a very simple mechanism for dealing with multi-agent communication issues. Specifically, we tested a simple strategy for disambiguating the intended addressee of each user utterance.

Addressee Disambiguation Algorithm

We have found that a simple algorithm for disambiguating the intended addressee of each user utterance. It can be described as follows:

- If an utterance starts with the name of a robot, then that is the robot this utterance is addressed to. We call this form of addressing *explicit addressing*, and the robot being addressed the *explicit addressee.* For example, in the utterance *Bashful, where are you?*, the form of addressing is explicit, and Bashful is the explicit addressee.
- If an utterance does not start with the name of a robot, then the last explicitly addressed robot is being addressed in this utterance. We call this form of addressing *implicit addressing*, and the robot being addressed the *implicit addressee.* For example, if the utterance above is followed by the utterance *Go ten meters north*, the form of addressing is implicit, and Bashful is the implicit addressee.

Task Description

In our experiment, users were required to navigate the two robots (Bashful and Clyde) through a maze of corridors using only the speech channel to communicate with the robots. The users were not allowed to see the robots and therefore had to rely on spoken dialogue to query the robots regarding their positions in the maze at all points of time.

Specifically, the task involved first finding out the initial positions of the two robots in the maze, and then navigating them to the point in the maze marked with an X. Figure 2 shows a map of the maze used. Participants were provided with a hard–copy of this map (without the initial locations of the robots) and were asked to mark on it the positions they believed the robots were at initially. Users were given a maximum of 30 minutes to finish the task, and were allowed to give up earlier if they wished. Note that users were not informed of the addressing mechanism described above since one of the aims of this experiment is to determine if the addressing mechanism can be intuited by the users and, in general, if it makes for naturalistic dialogue.

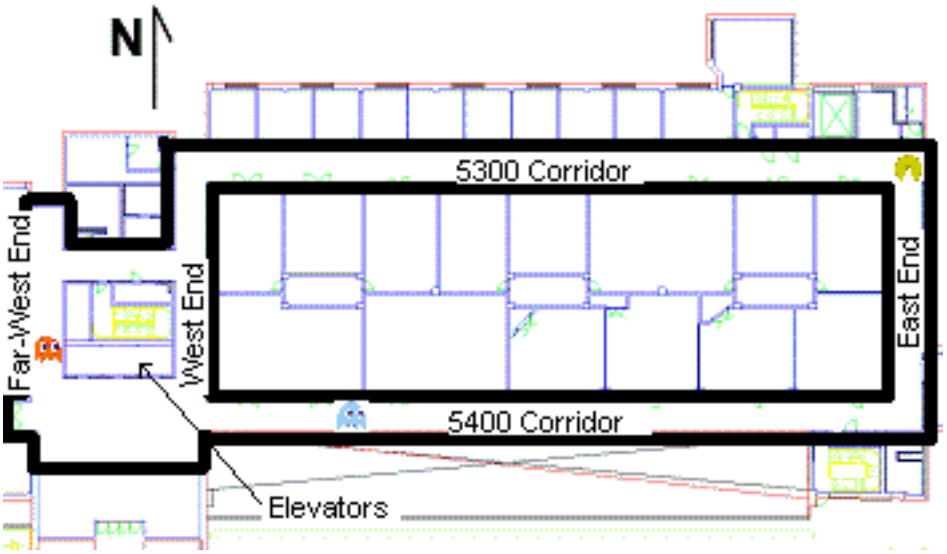

Figure 2: Maze Map

Grammar for User Utterances

We used the following simple grammar to parse the user's utterances:

`HumanReportCommand` →
`([RobotName]? report)` | `([RobotName])`

`RobotName` could be either Bashful or Clyde. The user could either utter "Bashful" or "Bashful report" to address Bashful explicitly, or just utter "report" to implicitly address the last explicitly addressed robot.

`HumanLocationQuery` →
`([RobotName]? where are you)`

This command could be used to query (either explicitly or implicitly) a robot's location.

`MoveVector` →
`([RobotName]? MOVE [Direction]? [Distance]?)`

This command could be used to direct a robot to move a certain distance along a certain direction. We restricted `distance` to be any integer distance from 1 to 20 meters, and `direction` to be either north, south, east or west. Although utterances that did not contain both a direction and a distance were parsed by the grammar, both pieces of information were needed to perform the move. Hence in a situation where the user provided only one or neither of the pieces of information, the dialogue manager would ask the user to supply the missing information.

Although users were not presented with the above grammar, they were informed that they could ask "Where are you?" and that they could instruct the robot to move between 1 and 20 meters in one of the four directions.

Robot Responses to User Utterances

The Ravenclaw dialog system used in this platform requires the designer to specify the response of the system for every parse–able user utterance. Responses may include speech output, back–end actions taken by the system, or a combination of the two. Following is a description of the system responses for each of the three families of user utterances described above.

Response to HumanReportCommand: Robots responded to this family of utterances by saying *Bashful here*, or *Clyde reporting*, etc. This dialog helped the user initiate a communication channel with a robot.

Response to HumanLocationQuery: Robots responded to this family of utterances by specifying where in the maze they were. Each part of the map a robot could be in was pre–assigned a name as shown in figure 2. The system's back–end mapped the robot's absolute (x, y) coordinates obtained from the CARMEN robot API to the corresponding area name. The system also computed the approximate distance from the closest end of the area. A typical reply to a HumanLocationQuery would be *I am now in the fifty three hundred corridor, about five meters from the east end.*

Response to MoveVector: The addressed robot responded to this family of utterances by first making sure it had a value for both the distance and the direction components. If one or both values were missing, the system engaged the user in a follow–up dialog by asking, for example, *How far do you want me to go east?* or *In which direction do you want me to go five meters?* Once both values were provided, the system used the robot's current (x, y) coordinates to compute the destination position, and then used CARMEN's autonomous navigation API to move the robot to the new position. At the same time, the robot would inform the user that it was following the command by uttering, for example, *Going five meters toward the north.*

Other Details

For the current study, we used simulated robots in a simulated environment. These robots were initially placed at the positions shown in Figure 2 for each participant in this user study. Participants used a single head–mounted close–talking microphone to speak to both robots, and the speech from both the robots was routed through a single set of speakers. To help the user to distinguish between the speech from the two robots, we used a male voice to synthesize the speech from Bashful, and a female voice to synthesize the voice from Clyde.

Results

Table 1: *Pilot–study Results*

Part. #	Task Success	Time Taken (mins)	Addressing Mechanisms Used
1	Both	28	Only explicit
2	One	21	Both forms
3	Both	28	Both forms
4	None	18	Both forms
5	One	20	Only explicit
6	One	12	Only explicit

We ran the experiment with 6 different participants. Every participant could correctly identify the approximate initial positions of the robots on the map. We defined task success as follows: A participant was completely successful (denoted as "both"), partially successful ("one") and completely unsuccessful ("none") if he or she managed to navigate both, one or none of the robots to their destinations respectively. We also measured the time taken till the end of the experiment. Table 1 shows task success and time taken for each of the 6 participants.

During the experiment we also noted what addressing mechanisms, explicit or implicit, the participant was using in his or her utterances. Three participants used only the explicit form of addressing; that is, each of their utterances was prefaced with the name of the robot. When asked at the end of the experiment whether they understood that they could engage in implicit addressing and simply chose not to, all three replied that they did not realize that implicit addressing was possible. The remaining three participants used both forms of addressing.

We also noted the strategy used by each participant in moving the robots to their final destinations. Four participants elected to simultaneously move both robots, conversing with one robot after giving the other robot a move command, for example. The two remaining participants chose to complete moving one robot to its final destination before commencing on the other robot.

Analysis and Lessons Learned

We analyzed the participants' utterances, and, for each participant, counted the number of utterances that the system recognized and processed without any errors, the number of utterances that had at least one error in them, and the number of utterances that were outside the system's grammar described in section . Utterance errors included either speech recognition errors (such as misrecognizing the word *east* as *west*) or dialog processing errors (such as ignoring a HumanLocationQuery command while the robot was in the process of engaging the user in confirmation dialog regarding a previous MoveVector command). Table 2 shows the results of this analysis. Corresponding participant numbers in tables 1 and 2 refer to the same participants. Due to a technical problem we lost the utterances of participant #1.

Table 2: *Utterance Analysis*

Part. #	# utts w/o recognition errors	# utts with recognition errors	# utts outside grammar
1	?	?	?
2	52 (77.6%)	9 (13.4%)	6 (9.0%)
3	88 (91.7%)	6 (6.2%)	2 (2.1%)
4	52 (70.3%)	18 (24.3%)	4 (5.4%)
5	103 (76.3%)	31 (23.0%)	1 (0.7%)
6	37 (43.0%)	47 (54.7%)	2 (2.3%)

Note from table 2 that more than 70% utterances for 4 subjects had no errors at all, and that less than 10% utterances for each subject were outside the system's grammar. These numbers result from the fact that the system used a very constrained grammar, making the speech recognizer's task relatively easy. The number of recognition errors seem to correspond somewhat with task success. Thus, the participant who managed to move both robots to their destination had the largest percentage of correctly recognized utterances, while the participant who had the lowest success had the second lowest percentage. Due to the small scale of this pilot experiment however, larger experiments may be needed to evaluate such hypotheses.

When asked to rate the interaction after the experiment, every participant replied that he or she found both the dialogue and the pace of the interaction naturalistic. These reports established that our implemented system can be used successfully to interact with robots. Furthermore, the fact that every participant understood the explicit addressing mechanism and half the participants understood the implicit mechanism implies that our simple addressee disambiguation algorithm is easy to understand and makes for natural dialogue.

Participants were also asked to provide feedback on any aspect of the experiment. Every participant felt that the robots do not always provide as much feedback as they could. For example in the current design when a robot is asked to go further than it can, they do not report this inability. Participants also expressed satisfaction at having two robots to work with instead of one. They felt that the pace of the interaction would have been too slow if there was only one robot, since robots take a long time to move from one point to another. Participants also felt that the set of commands that the user can issue was limiting. When the robot

was stuck against an unknown obstruction, the participants felt that more exploratory commands such as *What can you see?* would have been very useful to make progress toward the goal. Some participants felt that descriptions of the locations as spoken by the robots were sometimes unintuitive. For example, when the robot said *I am in front of the elevators about 3 meters from the east end*, the robot was referring to the east end of "in front of the elevators". Since "in front of the elevators" is not normally the type of area with clear boundaries, the subject naturally thought that "east end" was the east end of the map. Further research is necessary to determine how to describe the current location of a robot such that the description is maximally intuitive from a human's point of view.

Conclusions

Heterogeneous interface agents cannot act in concert, achieving a globally optimal interface strategy by understanding or predicting each others' behavior. Some research groups have taken the approach that constructs an aggregated spoken dialogue front-end for a community of under–specified agents. The Speech Graffiti Personal Universal Controller (Harris, 2004), which was designed explicitly with multi-agent control in mind, is such an aggregating system. This system severely limits the expressive power of natural language, however, and any aggregating spoken dialogue front-end will potentially sacrifice the integration of domain knowledge into the dialogue.

In order to directly address heterogeneous multi-agent communication problems, we have established an understanding of the issues and a platform for experimentation in that domain. The platform, with a few simple strategies, has yielded interesting lessons and results from a small pilot study.

Future Directions

We seek strategies that support the protocol-minimum solutions. We examine those sets of solutions to these communication problems that would place the smallest burden on some new robot. For example, one solution to speaking in turn would be to monitor the acoustic channel for speech, and wait for an absence of speech before speaking oneself. Another approach would be to assign one robot to be the server, and that server itself would manage the speaking turns as if the robots had monitored the speech channel for each others' acoustic utterances. Both solutions have an identical performance, but the first solution required each robot to monitor the acoustic channel while the second solution requires no monitoring of the acoustic channel. Because the second approach assumes less about the robots' capabilities, it may be preferred as a protocol.

We believe it is possible for systems to adhere to protocols when these are lightweight, orthogonal and naïve of the research topics, and the advantages are clear. Computer network protocols are one such example. Computer network protocols such as the Ethernet protocol exist in environments of heterogeneous computer systems, are lightweight protocols, are orthogonal and not committed to a particular approach to communication. The advantages of having such protocols are clear. The analogy can be carried further when one realizes that computer network protocols also address the four problem areas that we have identified for environments of multiple dialogue agents: message identification, message addressing, channel contention, and session identification.

References

Black, A., Clark, R., Richmond, K., & King, S. (2004). *The festival speech synthesis system.*

Bohus, D., & Rudnicky, A. (2002, November). *Integrating multiple knowledge sources for utterance-level confidence annotation in the CMU Communicator spoken dialog system* (Tech. Rep. No. CMU-CS-02-190). Pittsburgh, Pennsylvania: School of Computer Science, Carnegie Mellon University.

Bohus, D., & Rudnicky, A. I. (2003). Ravenclaw: Dialog management using hierarchical task decomposition and an expectation agenda. In *Eurospeech.* Geneva, Switzerland.

Harris, T. (2004). *The speech graffiti personal universal controller*. Unpublished master's thesis, Carnegie Mellon University, Pittsburgh.

Huang, D., Alleva, F., Hon, H. W., Hwang, M. Y., Lee, K. F., & Rosenfeld, R. (1993). The Sphinx-II speech recognition system: An overview. *Computer, Speech, and Language*, *7*(2), 137–148.

Montemerlo, M., Roy, N., & Thrun, S. (2002). *Carnegie mellon robot navigation toolkit.*

Oh, A. H., & Rudnicky, A. (2000, May). Stochastic language generation for spoken dialogue systems. In *Anlp/naacl workshop on conversational systems* (pp. 27–32).

Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., & Zue, V. (1998). Galaxy-II: A reference architecture for conversational system development. In *Proceedings of the international conference on spoken language processing.*

Ward, W. (1994, September). Extracting information from spontaneous speech. In *Proceedings of the international conference on spoken language processing.*