Human-Robot Dialogue and Collaboration in Search and Navigation

Claire Bonial[1], Stephanie M. Lukin[1], Ashley Foots[1], Cassidy Henry[1], Matthew Marge[1], Kimberly A. Pollard[1], Ron Artstein[2], David Traum[2], Clare R. Voss[1]
[1]U.S. Army Research Laboratory, [2]USC Institute for Creative Technologies
Adelphi MD 20783, Playa Vista CA 90094
claire.n.bonial.civ@mail.mil

Abstract

Collaboration with a remotely located robot in tasks such as disaster relief and search and rescue can be facilitated by grounding natural language task instructions into actions executable by the robot in its current physical context. The corpus we describe here provides insight into the translation and interpretation a natural language instruction undergoes starting from verbal human intent, to understanding and processing, and ultimately, to robot execution. We use a 'Wizard-of-Oz' methodology to elicit the corpus data in which a participant speaks freely to instruct a robot on what to do and where to move through a remote environment to accomplish collaborative search and navigation tasks. This data offers the potential for exploring and evaluating action models by connecting natural language instructions to execution by a physical robot (controlled by a human 'wizard'). In this paper, a description of the corpus (soon to be openly available) and examples of actions in the dialogue are provided.

Keywords: human-robot interaction, multiparty dialogue, dialogue structure annotation

1. Introduction

Efficient communication in dynamic environments is needed to facilitate human-robot collaboration in many shared tasks, such as navigation, search, and rescue operations. Natural language dialogue is ideal for facilitating efficient information exchange, given its use as the mode of communication in human collaboration on these and similar tasks. Although the flexibility of natural language makes it well-suited for exchanging information about changing needs, objectives, and physical environments, one must also consider the complexity of interpreting human intent from speech to an executable instruction for a robot. In part because this interpretation is so complex, we are developing a human-robot dialogue system using a bottom-up, phased 'Wizard-of-Oz' (WoZ) approach. It is bottom-up in the sense that we do not assume that we can know *a priori* how humans would communicate with a robot in a shared task. Instead, the phased WoZ methodology, in which humans stand in for technological components that do not yet exist, allows us to gather human-robot communication data, which in turn will be used in training the automated components that will eventually replace our human wizards.

Here, we describe the details of our data collection methodology and the resulting corpus, which can be used in connecting spoken language instructions to actions taken by a robot (action types and a sample of spoken instructions are given in Table 1), as well as relevant images and video collected on-board the robot during the collaborative search and navigation task. Thus, this corpus offers potential for exploring and evaluating models for representing, interpreting and executing actions described in natural language.

Action Type Action Sub-Type	IU N	 %
Command	1243	94
Send-Image	443	52
"take a photo of the doorway to your right"		
"take a photo every forty five degrees"		
Rotate	406	47
"rotate left twenty degrees"		
"turn back to face the doorway"		
Drive	358	42
"can you stop at the second door"		
"move forward to red pail"		
Stop	29	3
"wait"		
"stop there"		
Explore	7	1
"explore the room"		
"find next doorway on your left"		
Request-Info	34	4
"how did you get to this building last time"		
"what type of material is that in front of you"		
Feedback	28	3
"essentially I don't need photos behind you"		
"no thank you not right now"		
Parameter	14	2
"the doorway with the boards across it"		
"the room that you're currently in"		
Describe	5	1
"watch out for the crate on your left"		

Table 1: Actions distribution over all Instruction Units (IU: see Section 3.1.) in the corpus (N=858). (Percent sum is greater than 100% as an IU may have one or more actions).

2. Corpus Collection Methodology

Our WoZ methodology facilitates a data-driven understanding of how people talk to robots in our collaborative domain. Similar to DeVault et al. (2014), we use the WoZ methodology only in the early stages of a multi-stage development process to refine and evaluate the domain and provide training data for automated dialogue system components. In all stages of this process, participants communicating with the 'robot' speak freely, even as increas-

ing levels of automation are introduced in each subsequent stage or 'experiment.' The iterative automation process utilizes previous experiments' data.

Currently, we are in the third experiment of the ongoing series, and our corpus includes data and annotations from the first two experiments. The first two experiments use two wizards: a Dialogue Manager Wizard (DM-Wizard, DM) who sends text messages and a Robot Navigator Wizard (RN-Wizard, RN) who teleoperates the actual robot. A naïve participant (unaware of the wizards) is tasked with instructing a robot to navigate through a remote, unfamiliar house-like environment, and asked to find and count objects such as shoes and shovels. The participant is seated at a workstation equipped with a microphone and a desktop computer displaying information collected by the robot: a map of the robot's position and its heading in the form of a 2D occupancy grid, the last still-image captured by the robot's front-facing camera, and a chat window showing the 'robot's' responses. This layout is shown in Figure 1. Note that although video data is collected on-board the robot, this video stream is not available to the participant, mimicking the challenges of collaborating with a robot in a low bandwidth environment. Thus, the participant's understanding of the environment is based solely upon still images that they request from the robot, the 2d map, and natural language communications with the robot.

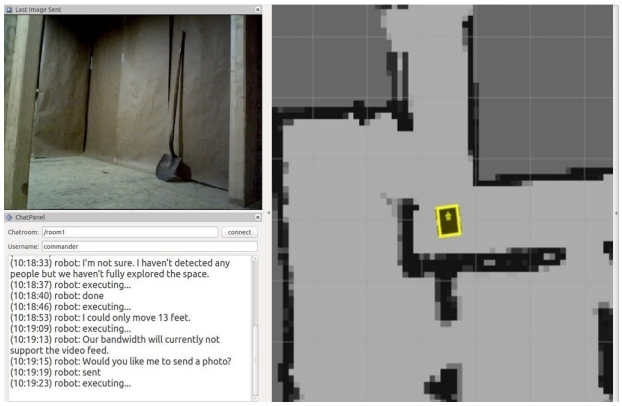

Figure 1: Participant's interface in experiments: photo from robot requested by participant (top left), chat window with text communications from 'robot' (bottom left), dynamically-updating 2D map of robot's location (right).

At the beginning of the study, the participant is given a list of the robot's capabilities: the robot understands basic object properties (e.g., most object labels, color, size), relative proximity, some spatial terms, and location history. The overall task goal is told explicitly to participants, and a worksheet with task questions is handed to the participant before they begin the exploration. For example, participants are aware that they will be asked to report the number of doorways and shovels encountered in the environment and to answer analysis questions, such as whether or not they believe that the space has been recently occupied. The participant may refer back to this worksheet, and to the list of robot capabilities, at any time during the task. To encourage as wide a range of natural language as possible, experimenters do not provide sample robot instructions. The participant is told that they can speak naturally to the robot to complete tasks.

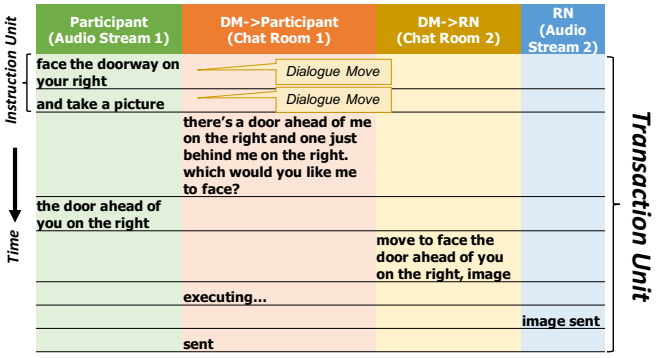

Figure 2: An interaction with one *transaction unit* (see 3.1.), showing the dialogue flow from the participant's spoken instructions to the robot's action and feedback.

In reality, the participant is speaking not to a robot, but to an unseen DM-Wizard who listens to the participant's spoken instructions and responds with text messages in the chat window. There are two high-level response options:

i If the participant's instructions are clear and executable in the current physical environment, then the DM-Wizard passes a simplified text version of the instructions to the RN-Wizard, who then joysticks the robot to complete the instructions and verbally acknowledges completion to the DM-Wizard over a private audio stream.

ii If the instructions are problematic in some way, due to ambiguity or impossibility given either the current physical context or the robot's capabilities, then the DM-Wizard responds directly to the participant in text via the chat window to clarify the instructions and/or correct the participant's understanding of the robot's capabilities.

Figure 2 shows an example *transaction unit* of the multi-party information exchange.

We engage each participant in three sessions: a training task and two main tasks. The training task is simpler in nature than the main tasks, and allows the participant to become acquainted with verbally commanding a robot. The main tasks, lasting 20 minutes each, focus on slightly different search and analysis subtasks and start in distinct locations within a house-like environment. The subtasks were developed to encourage participants to treat the robot as a teammate who helps search for certain objects, but also to tap into participants' own real-world knowledge to analyze the environment.

In Experiment 1, our goal was to elicit a full range of communications that may arise. The DM-Wizard typed free-text responses to the participant following guidelines established during piloting that governed the DM-Wizard's real-time decision-making (Marge et al., 2016). Ten subjects participated in Experiment 1.

In Experiment 2, instead of free responses, the DM-Wizard constructs a response by selecting buttons on a graphical user interface (GUI). Each button press sends a pre-defined text message, mapped from the free responses, to either the participant or RN-Wizard (Bonial et al., 2017). The GUI also supports templated text messages where the the DM-Wizard fills in a text-input field, for example to specify how many feet to go forward in a move command: "Move forward ___ feet."

To create Experiment 2's GUI, data from all ten Experiment 1 participants were analyzed to compose a communication set balancing tractability for automated dialogue and full domain coverage, including recovery from problematic instructions. 99.2% of Experiment 1 utterances were covered by buttons on the GUI (88.7% were exact matches, 10.5% were partial text-input matches) which included 404 total buttons. Buttons generated participant-directed text such as "processing. . ." "How far southeast should I go?" and "Do you mean the one on the left?" as well as RN-directed text such as "turn to face West," "move to cement block," and "send image."

Experiment 2 included ten new participants and was conducted exactly like Experiment 1, aside from the use of the DM-Wizard's GUI. The switch from free-typing to a GUI is a step in the progression toward increasing automation; i.e. it represents one step closer to 'automating away' the human wizards. The GUI buttons constrain DM-Wizard responses to fixed and templatic messages in order to provide tractable training data for an eventual automated dialogue system. Thus, executable instructions from Experiment 2 participants were translated using this limited set when passed to the RN-Wizard. This difference between Experiments 1 and 2 is evident in the corpus and the example in Figure 6 to follow.

3. Corpus Details

We are preparing the release of our Experiment 1 and 2 data, which comprises 20 participants and about 20 hours of audio, with 3,573 participant utterances (continuous speech) totaling 18,336 words, as well as 13,550 words from DM-Wizard text messages. The corpus includes speech transcriptions from participants as well as the speech of the RN-Wizard. These transcriptions are time-aligned with the DM-Wizard text messages passed to the participant and to the RN-Wizard. We are also creating videos that align additional data streams: the participant's instructions, the text messages to both the participant and the RN-Wizard passed via chat windows, the dynamically updating 2D map data, still images taken upon participant request, and video taken from on-board the robot throughout each experimental session (as mentioned in the previous section, video is collected but is never displayed to the participant in order to simulate a low band-width communication environment). We are exploring various licensing possibilities in order to release as much of this data as possible.

3.1. Annotations

The corpus includes dialogic annotations alongside the original data streams. The goal of these annotations is to illuminate dialogue patterns that can be used as features in training the automated dialogue system. Although there are standard annotation schemes for both dialogue acts (Bunt et al., 2012) and discourse relations (Prasad and Bunt, 2015) (and our annotations do overlap with both of these) we found that existing schemes do not fully address the issues of dialogue structure. Of particular interest to us, and not previously addressed in other schemes, are cases in which the units and relations span across multiple conversational floors. Full details on the annotations can be found in Traum et al. (2018) and Marge et al. (2017). This discussion will be limited to annotations that help to summarize what action types are requested in the instructions and carried out by the robot. We discuss three levels of dialogue structure, from largest to smallest: *transaction units*, *instruction units*, and actions or *dialogue-moves*. Each of these is defined below.

Each dialogue is annotated as a series of higher-level *transaction units* (TU). A TU is a sequence of utterances aiming to achieve a task intention. Each TU contains a participant's initiating message and then subsequent messages by the participant and wizards to complete the transaction, either by task execution or abandonment of the task in favor of another course of action.

Within TUs, we mark *instruction units* (IU). An IU comprises all participant speech to the robot within a transaction unit before robot feedback. Each IU belongs to exactly one TU, so that each transaction's start (e.g., a new command is issued) marks a new IU. An IU terminates when the robot replies to the request, or when a new transaction is initiated.

To analyze internal IU structure, we annotate participant-issued finer-grained actions with *dialogue-moves*. Specific to the robot navigation domain, these include *commands*, with subtypes such as *command:drive* or *command:rotate*. Our schema supports clarifications and continuations of participant-issued actions, which are annotated as being linked to the initial action. The relationships of IUs, TUs, and dialogue moves is exemplified in both Figure 2 and Figure 3.

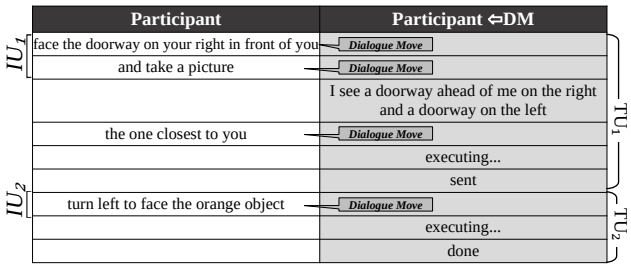

Figure 3: Annotation structures on human-robot dialogue, shown over participant and DM-Wizard streams.

3.2. Actions in the Data

We analyzed the selection of dialogue-moves that participants issued in their IUs. Participants often issued more than one dialogue-move per IU (mean = 1.6 dialogue-moves per IU, s.d. = 0.88, min = 1, max = 8). Unsurpris-

ingly, the *command* dialogue-move was the most frequent across IUs (appearing in 94% of all IUs). Table 1 summarizes the dialogue move types in the corpus, and gives a sense of the action types requested of the robot to complete search and navigation tasks (full description found in Marge et al. (2017)).

Actions are initiated by participant verbal instructions, then translated into a simplified text version passed by the DM-Wizard to the RN-Wizard, who carries out physical task execution. Throughout an interaction, feedback is passed up from both the RN-Wizard to the DM-Wizard and from the DM-Wizard to the participant. This feedback is crucial for conveying action status: indicating first that the instructions were heard and understood, then that they are being executed, and finally that they are completed.

For each clear, unambiguous instruction (as opposed to instructions that require clarifying dialogue between the DM-Wizard and participant), there are three realizations or interpretations of a single action:

i Participant's instruction for action, expressed in spoken language;

ii DM-Wizard's translation into simplified text message for RN;

iii RN-Wizard's execution of text instruction with physical robot, evident to participant via motion on the 2D map.

In addition to these perspectives on an action, a full TU also includes the RN-Wizard's confirmation of execution, spoken to the DM-Wizard, and finally the DM-Wizard's translation of this confirmation to the participant in a text message. Here, we provide several examples of this 'translation' process from our data, ranging from explicit, simple instructions to more complex and opaque instructions.

In many cases, the participant provides instructions that are simple and explicit, such that there is little change in the instructions from the spoken language to the text version the DM-Wizard sends to the RN-Wizard(Figure 4). Furthermore, in most of these simple cases, the action carried out seems to match the participant intentions given that no subsequent change or correction is requested by the participant.

Participant (Audio Stream 1)	DM->Participant (Chat Room 1)	DM->RN (Chat Room 2)	RN (Audio Stream 2)
turn ninety degrees to the left			
	ok		
		turn left 90 degrees	
	turning...		
			done
	done		

Figure 4: A simple and explicit action carried out.

In other cases, the instructions are less explicit in how they should be translated into robot action. For example, in Figure 5, the request for the robot to "Take a picture of what's behind you" implicitly requires first turning around 180 degrees before taking the picture. Our human DM-Wizard has no problem recognizing the need for this implicit action, but in the future, associating queries regarding "behind [X]" with particular actions will require nuanced spatial understanding in our automated system. Other instructions mentioning "behind" do not require the implicit turn, such as: "Can you go around and take a photo behind the TV?" An adequate system requires the sophistication to tease apart distinct spatial meanings in different physical contexts.

Participant (Audio Stream 1)	DM->Participant (Chat Room 1)	DM->RN (Chat Room 2)	RN (Audio Stream 2)
take a picture of what's behind you			
		turn 180, photo	
	executing...		
			image sent

Figure 5: Here, the instructions must be decomposed into the prerequisite actions needed to achieve the final goal.

Given the use of the GUI in Experiment 2, some instructions that appeared to be straightforward and explicit required a great deal of translation to be properly conveyed using the limited set of fixed and templatic action messages available to the DM-Wizard. For example, in Figure 6, the participant requests that the robot move to a clear destination (a yellow cone), stopping to take pictures every two feet along the way. The instruction must be broken into sub-actions, as there is no fixed message or template in the interface to express it in its entirety. Thus, the instruction to move two feet and send a photo is repeated eight times before reaching the destination.

Participant (Audio Stream 1)	DM->Participant (Chat Room 1)	DM->RN (Chat Room 2)	RN (Audio Stream 2)
move toward the yellow cone			
and take a photo every two feet			
	processing. . .		
		turn to face yellow cone	
		then. . .	
	Repeated 8 Iterations	move forward 2 feet	
		then. . .	
		send image	
			done and sent
		move forward 2 feet	
		then. . .	
		send image	

Figure 6: These instructions must be decomposed into simpler robot actions repeated 8 times (2 iterations shown).

Other instructions remain challenging due to their opacity and demand for pragmatic knowledge. Figure 7 provides an example that draws upon the robot's history of actions: "do the same." Determining which of the robot's preceding actions in a complex series of actions should be included in "the same" relies upon a sophisticated understanding of

both the physical context and discourse structure (i.e. what portion of the previous utterance done in a past location should be done in a new location?).

Participant (Audio Stream 1)	DM->Participant (Chat Room 1)	DM->RN (Chat Room 2)	RN (Audio Stream 2)
go into the center of the room in front of you			
and then take a picture at the <pause> east south west and north position			
		move into the center of the room in front of you, take photos at east, south, west, north positions	
	executing...		
			done
	done		
go into the room behind you			
and do the same			

Figure 7: The DM-Wizard, and in the future, the robot, must determine what is indicated by "same."

4. Conclusions & Future Work

The corpus collected will inform both the action space of possible tasks and required parameters in human-robot dialogue. As such, our 'bottom-up' approach empirically defines the range of possible actions. At the same time, we are exploring symbolic representations of the robot's surroundings, derived from the objects discussed in the environment, their locations, and the referring expressions used to ground those objects. For natural language instructions to map to robot actions, we are implementing plan-like specifications compatible with autonomous robot navigation. *Primitives* such as rotations and translations, along with absolute headings (e.g., cardinal directions, spatial language), will complement the action space. Possible techniques to leverage include both supervised and unsupervised methods of building these representations from joint models of robot and language data.

We have trained a preliminary automated dialogue manager using the Experiment 1 and 2 data, but are continuing to collect data in simulation to improve the results (Henry et al., 2017). The system currently relies on string divergence measures to associate an instruction with either a text version to be sent to the RN-Wizard or a clarification question to be returned to the participant. The challenging cases described in this paper demonstrate that a deeper semantic model will be necessary. Associating instructions referring to "behind [X]" or "do that again" with the appropriate actions in context will require modeling aspects of the discourse structure and physical environment that go far beyond string matching alone.

Furthermore, we are just beginning to tackle precise action execution methods (Moolchandani et al., 2018). Even if an action's overall semantics are understood, ambiguous attributes remain. For example, precisely where and in what manner should a robot move relative to a door when requested to do so?

This research provides data for associating spoken language instructions to actions taken by the robot, as well as images/video captured along the robot's journey. Our approach resembles that of *corpus-based robotics* (Lauria et al., 2001), whereby a robot's action space is directly informed from empirical observations, but our work focuses on data collection of bi-directional communications about actions. Thus, this data offers value for refining and evaluating action models. As we continue to explore the annotations and models needed to develop our own dialogue system, we invite others to utilize this data in considering other aspects of action modeling in robots (release scheduled for the coming year).

5. Bibliographical References

Bonial, C., Marge, M., Foots, A., Gervits, F., Hayes, C. J., Henry, C., Hill, S. G., Leuski, A., Lukin, S. M., Moolchandani, P., Pollard1, K. A., Traum, D., and Voss, C. R. (2017). Laying Down the Yellow Brick Road: Development of a Wizard-of-Oz Interface for Collecting Human-Robot Dialogue. *Proc. of AAAI Fall Symposium Series*.

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., Belis-Popescu, A., and Traum, D. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proc. of LREC*, Istanbul, Turkey, May.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S. C., Fabrizio, M., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and Morency, L.-P. (2014). SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *Proc. of AAMAS*.

Henry, C., Moolchandani, P., Pollard, K., Bonial, C., Foots, A., Hayes, C., Artstein, R., Voss, C., Traum, D., and Marge, M. (2017). Towards Efficient Human-Robot Dialogue Collection: Moving Fido into the Virtual World. *Proc. of ACL Workshop Women and Underrepresented Minorities in Natural Language Processing*.

Lauria, S., Bugmann, G., Kyriacou, T., Bos, J., and Klein, E. (2001). Training Personal Robots Using Natural Language Instruction. *IEEE Intelligent Systems*, 16:38–45.

Marge, M., Bonial, C., Byrne, B., Cassidy, T., Evans, A. W., Hill, S. G., and Voss, C. (2016). Applying the Wizard-of-Oz Technique to Multimodal Human-Robot Dialogue. In *Proc. of RO-MAN*.

Marge, M., Bonial, C., Foots, A., Hayes, C., Henry, C., Pollard, K., Artstein, R., Voss, C., and Traum, D. (2017). Exploring Variation of Natural Human Commands to a Robot in a Collaborative Navigation Task. *Proc. of ACL Workshop RoboNLP: Language Grounding for Robotics*.

Moolchandani, P., Hayes, C. J., and Marge, M. (2018). Evaluating Robot Behavior in Response to Natural Language. *To appear in the Companion Proceedings of the HRI Conference*.

Prasad, R. and Bunt, H. (2015). Semantic relations in discourse: The current state of iso 24617-8. In *Proc. of 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 80–92.

Traum, D., Henry, C., Lukin, S., Artstein, R., Gervits, F., Pollard, K., Bonial, C., Lei, S., Voss, C., Marge, M., Hayes, C., and Hill, S. (2018). Dialogue Structure Annotation for Multi-Floor Interaction. In *Proc. of LREC*.