Towards Efficient Human-Robot Dialogue Collection: Moving Fido into the Virtual World

Cassidy Henry[1], Pooja Moolchandani[1], Kimberly A. Pollard[1], Clare Bonial[1], Ashley Foots[1], Ron Artstein[2], Cory Hayes[1], Clare R. Voss[1], David Traum[2], and Matthew Marge[1]

[1]U.S Army Research Laboratory, Adelphi, MD 20783
[2]USC Institute for Creative Technologies, Playa Vista, CA 90094
cassidy.r.henry.ctr@mail.mil

Abstract

Our research aims to develop a natural dialogue interface between robots and humans. We describe two focused efforts to increase data collection efficiency towards this end: creation of an annotated corpus of interaction data, and a robot simulation, allowing greater flexibility in when and where we can run experiments.

1 Introduction

Effective human-robot (H-R) teaming requires robots to engage in natural communication. Natural language (NL) dialogue allows bi-directional information exchange, with the benefit of familiarity and flexibility for humans. Our goal is to develop dialogue processing capabilities for an automated robot receiving instruction from a remote human teammate in a collaborative search-and-navigate task. The physical robot, affectionately nicknamed Fido by a participant, is a Clearpath Robotics Jackal running ROS (Quigley et al., 2009).

We follow a multiphase Wizard-of-Oz (WoZ) data collection approach (Marge et al., 2016a) to bootstrap the robot's planned language capabilities, as the technology to support teaming interactions does not yet exist. The solution cannot simply decompose into autonomous robot control and language processing. It cannot be fully autonomous as it must respond to and integrate instructions, questions, and information from human teammates into its plans. Natural language processing (NLP) relies on situated interaction based on the dynamic state and robot action, perception, and inferential capabilities, that can be neither as simple as translation to a rigid command language nor as extensive as requiring the full range of human

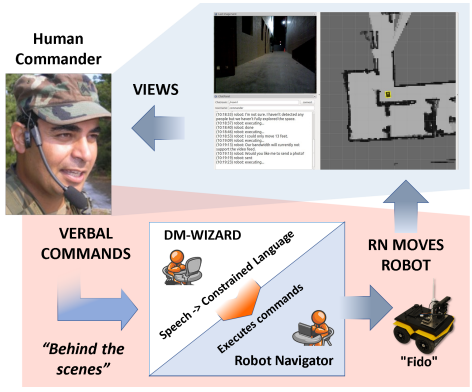

Figure 1: WoZ Experiment Setup, where a CMD gives vocal instruction to a remote robot, represented by the two-wizard setup

reasoning power and common sense knowledge. We use a two-wizard setup (Figure 1) to address this interdependence, allowing separate simulation of both NL interaction based on flexible but limited robot intelligence, and navigation controls (Marge et al., 2016b). The Dialogue Manager (DM) listens to Commander (CMD) speech, then either types back dialogue replies, or types constrained action sets to the Robot Navigator (RN) who joysticks the robot.

The first experiment phase revealed several data collection challenges, described in Section 2. In Section 3, we describe current efforts to annotate dialogue data, to be used to scope requirements for automated action and interaction, and provide training data for automating NLP components. In Section 4, we describe ongoing robot simulator development which provides the ability to collect additional training data more efficiently for both dialogue processing and robot navigation.

2 Challenges of H-R Dialogue Collection

Substantial resources are required for data collection. Due to shared test space and a single robot, we could only one run participant at a time. Four

researchers are needed for each participant for a two-hour time block. The location dependence restricts the participant pool, making sufficient participent recruitment difficult. We found more data is necessary to both capture more natural participant language use variation, and to collect sufficient training data to automate dialogue processing and robot navigation capabilities.

Two focused efforts enable us to address these issues: development of an annotated corpus of language interactions (to automate more aspects of the language process and reduce human wizard labor), and a virtual simulation replicating our physical environment (allowing greater flexibility in when and where we can run experiments).

3 Annotation of H-R Dialogue

We collected ∼10.5 hours of CMD speech in total in phase one, along with the DM text messages and RN feedback. Utterances were segmented in Praat (Boersma, 2001)

on a per-command basis, consisting of a single action or answer. There were 1668 total commands across all ten participants. Initial corpus processing included speech transcription and manual time alignment of the four streams to enable analysis of utterance relationships. Figure 2 shows the four streams (two typed, two spoken) from three interlocutors and contains: instructions, translations to the RN, feedback, clarification, and question answering. We performed several annotations, including those for dialogue moves(Marge et al., 2017), structure, and relations on the data.

Commander	DM->CMD	DM->RN	Robot Navigator
drive to the doorway			
	Which doorway?		
the doorway with the boards across it			
		move to the doorway ahead of you on the left	
	executing...		
			done
	done		
take a picture			
		image	
			image sent
	sent		

Figure 2: Excerpt of dialogue corpus, showing four message streams.

Each typed DM utterance was labeled with a dialogue move. The dialogue move categories were analyzed and condensed to form a best-representative reply for each category. We condensed DM utterances into a smaller set of specific utterances and utterance templates, which were used to build a GUI for the second phase that begins language production automation by providing tractable response data. It transforms a fully-generative task to one of selection and specification of a few details, speeding up response time and lowering the effort required to produce complex language instructions and feedback.

Utterance data is being annotated and analyzed to automate NL understanding and dialogue management modules, relying on observed behavior patterns to generate, tune, and evaluate policies for responding to the participant and translating instructions to the navigation component.

4 Moving "Fido" Into the Virtual World

Our simulation setup aims to reduce requirements to run our research program's next phase and collect more dialogue data. We developed high-fidelity replications of the robot and physical environment using ROS and Gazebo (Koenig and Howard, 2016).

The virtual Jackal was equipped with the same sensors as the physical platform. The simulated setup will host the same task, and the GUI display to CMD is the same as earlier experiments (see Figure 1, top). A point-and-click navigation system was included alongside the virtual robot, which, along with the GUI, enables a single wizard to perform both dialogue management and navigation.

Simulation thus provides several advantages for data collection: (i) faster, parallel data collection, (ii) multiple site experimentation without physical robots, and (iii) simpler control for human operators. With robot and simulation operating on the same software and emulated hardware, we expect the experiment will smoothly transition back into a physical environment for validation purposes.

Phase	DM	RN	Enviro	Gaze/ MultiSense	# Participants
Expt 1	WoZ, Typed	WoZ, Joystick	Real	No	10
Expt 2	WoZ, GUI	WoZ, Joystick	Real	No	10
Expt 3	WoZ, GUI	WoZ, Joystick	Simulated	Yes	30+
Expt 4	Automated	Automated	Simulated	Yes	30+
Expt 5	Automated	Automated	Real	Yes	10+

Figure 3: Differences across different phases of experiment and increasing number of participants.

5 Conclusion

Our overall program objective is to provide more natural ways for humans to interact and communicate with robots using language, utilizing multi-

phase data collection experiments to incrementally automate the system, towards the ultimate goal of full automation. We highlighted two focused efforts to increase data collection efficiency: corpus creation/GUI development effort and robot simulation. This corpus will help address many issues encountered in understanding and processing situated H-R dialogue. The robot simulation replicates our physical environment while allowing greater flexibility in running experiments, and allows validation of simulated results in a physical environment after completed data collection.

References

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5:9/10:341–345.

Nathan Koenig and Andrew Howard. 2016. Design and use paradigms for Gazebo, an open-source multi-robot simulator. In *Proceedings of the International Conference on Intelligent Robots and Systems*. IEEE.

Matthew Marge, Claire Bonial, Brendan Byrne, Taylor Cassidy, A. William Evans, Susan G. Hill, and Clare Voss. 2016a. Applying the Wizard-of-Oz technique to multimodal human-robot dialogue. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*.

Matthew Marge, Claire Bonial, Ashley Foots, Cory Hayes, Cassidy Henry, Kimberly A. Pollard, Ron Artstein, Clare R. Voss, and David Traum. 2017. Investigating variation of natural human commands to a robot in a collaborative navigation task. In *Proceedings of RoboNLP: Language Grounding for Robotics*.

Matthew Marge, Claire Bonial, Kimberly A. Pollard, Ron Artstein, Brendan Byrne, Susan G. Hill, Clare Voss, and David Traum. 2016b. Assessing agreement in human-robot dialogue strategies: A tale of two wizards. In *Proceedings of the Sixtheenth International Conference on Intelligent Virtual Agents*.

Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source Robot Operating System. In *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE.