

ScoutBot: A Dialogue System for Collaborative Navigation

Stephanie M. Lukin¹, Felix Gervits^{2*}, Cory J. Hayes¹, Anton Leuski³, Pooja Moolchandani¹, John G. Rogers III¹, Carlos Sanchez Amaro¹, Matthew Marge¹, Clare R. Voss¹, David Traum³

¹ U.S. Army Research Laboratory, Adelphi, MD 20783

² Tufts University, Medford MA 02155

³ USC Institute for Creative Technologies, Playa Vista, CA 90094

stephanie.m.lukin.civ@mail.mil, felix.gervits@tufts.edu, traum@ict.usc.edu

Abstract

ScoutBot is a dialogue interface to physical and simulated robots that supports collaborative exploration of environments. The demonstration will allow users to issue unconstrained spoken language commands to ScoutBot. ScoutBot will prompt for clarification if the user's instruction needs additional input. It is trained on human-robot dialogue collected from Wizard-of-Oz experiments, where robot responses were initiated by a human wizard in previous interactions. The demonstration will show a simulated ground robot (Clearpath Jackal) in a simulated environment supported by ROS (Robot Operating System).

1 Introduction

We are engaged in a long-term project to create an intelligent, autonomous robot that can collaborate with people in remote locations to explore the robot's surroundings. The robot's capabilities will enable it to effectively use language and other modalities in a natural manner for dialogue with a human teammate. This demo highlights the current stage of the project: a data-driven, automated system, ScoutBot, that can control a simulated robot with verbally issued, natural language instructions within a simulated environment, and can communicate in a manner similar to the interactions observed in prior Wizard-of-Oz experiments. We used a Clearpath Jackal robot (shown in Figure 1a), a small, four-wheeled unmanned ground vehicle with an onboard CPU and inertial measurement unit, equipped with a camera and light detection and ranging (LIDAR) mapping that

*Contributions were primarily performed during an internship at the Institute for Creative Technologies.

readily allows for automation. The robot's task is to navigate through an urban environment (e.g., rooms in an apartment or an alleyway between apartments), and communicate discoveries to a human collaborator (the Commander). The Commander verbally provides instructions and guidance for the robot to navigate the space.

The reasons for an intelligent robot collaborator, rather than one teleoperated by a human, are twofold. First, the human resources needed for completely controlling every aspect of robot motion (including low-level path-planning and navigation) may not be available. Natural language allows for high-level tasking, specifying a desired plan or end-point, such that the robot can figure out the details of how to execute these natural language commands in the given context and report back to the human as appropriate, requiring less time and cognitive load on humans. Second, the interaction should be robust to low-bandwidth and unreliable communication situations (common in disaster relief and search-and-rescue scenarios), where it may be impossible or impractical to exercise real-time control or see full video streams. Natural language interaction coupled with low-bandwidth, multimodal updates addresses both of these issues and provides less need for high-bandwidth, reliable communication and full attention of a human controller.

We have planned the research for developing ScoutBot as consisting of five conceptual phases, each culminating in an experiment to validate the approach and collect evaluation data to inform the development of the subsequent phase. These phases are:

1. Typed wizard interaction in real-world environment
2. Structured wizard interaction in real environment
3. Structured wizard interaction in simulated environment
4. Automated interaction in simulated environment
5. Automated interaction in real environment

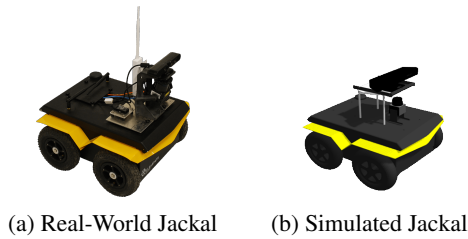


Figure 1: Jackal Robot

The first two phases are described in [Marge et al. \(2016\)](#) and [Bonial et al. \(2017\)](#), respectively. Both consist of Wizard-of-Oz settings in which participants believe that they are interacting with an autonomous robot, a common tool in Human-Robot Interaction for supporting not-yet fully realized algorithms ([Riek, 2012](#)). In our two-wizard design, one wizard (the Dialogue Manager or DM-Wizard) handles the communication aspects, while another (the Robot Navigator or RN-Wizard) handles robot navigation. The DM-Wizard acts as an interpreter between the Commander and robot, passing simple and full instructions to the RN-Wizard for execution based on the Commander instruction (e.g., the Commander instruction, *Now go and make a right um 45 degrees* is passed to the RN-Wizard for execution as, *Turn right 45 degrees*). In turn, the DM-Wizard informs the Commander of instructions the RN-Wizard successfully executes or of problems that arise during execution. Unclear instructions from the Commander are clarified through dialogue with the DM-Wizard (e.g., *How far should I turn?*). Additional discussion between Commander and DM-Wizard is allowed at any time. Note that because of the aforementioned bandwidth and reliability issues, it is not feasible for the robot to start turning or moving and wait for the Commander to tell it when to stop - this may cause the robot to move too far, which could be dangerous in some circumstances and confusing in others. In the first phase, the DM-Wizard uses unconstrained texting to send messages to both the Commander and RN-Wizard. In the second phase, the DM-Wizard uses a click-button interface that facilitates faster messaging. The set of DM-Wizard messages in this phase were constrained based on the messages from the first phase.

This demo introduces ScoutBot automation of the robot to be used in the upcoming phases: a simulated robot and simulated environment to support the third and fourth project phases, and initial automation of DM and RN roles, to be used in the

fourth and fifth phases. Simulation and automation were based on analyses from data collected in the first two phases. Together, the simulated environment and robot allow us to test the automated system in a safe environment, where people, the physical robot, and the real-world environment are not at risk due to communication or navigation errors.

2 System Capabilities

ScoutBot engages in collaborative exploration dialogues with a human Commander, and controls a robot to execute instructions and navigate through and observe an environment. The real-world Jackal robot measures 20in x 17in x 10in, and weights about 37lbs (pictured in [Figure 1a](#)). Both it and its simulated counterpart (as seen in [Figure 1b](#)) can drive around the environment, but cannot manipulate objects or otherwise interact with the environment. While navigating, the robot uses LIDAR to construct a map of its surroundings as well as to indicate its position in the map. The Commander is shown this information, as well as static photographs of the robot’s frontal view in the environment (per request) and dialogue responses in a text interface.

Interactions with ScoutBot are collaborative in that ScoutBot and the Commander exchange information through dialogue. Commanders issue spoken navigation commands to the robot, request photos, or ask questions about its knowledge and perception. ScoutBot responds to indicate when commands are actionable, and gives status feedback on its processing, action, or inability to act. When ScoutBot accepts a command, it is carried out in the world.

[Figure 2](#) shows a dialogue between a Commander and ScoutBot. The right two columns show how the Commander’s language was interpreted, and the action the robot takes. In this dialogue, the Commander begins by issuing a single instruction, *Move forward* (utterance #1), but the end-point is underspecified, so ScoutBot responds by requesting additional information (#2). The Commander supplies the information in #3, and the utterance is understood as an instruction to move forward 3 feet. ScoutBot provides feedback to the Commander as the action is begins (#4) and upon completion (#5). Another successful action is executed in #6-8. Finally, the Commander’s request to know what the robot sees in #9 is interpreted as a re-

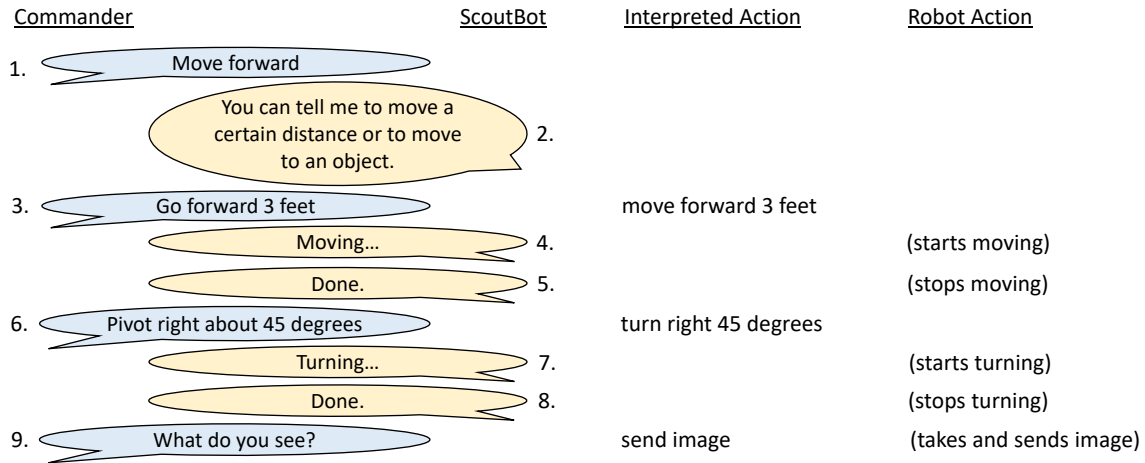


Figure 2: Dialogue between a Commander and ScoutBot

quest for a picture from the robot’s camera, which is taken and sent to the Commander.

ScoutBot accepts unconstrained spoken language, and uses a statistical text classifier trained on annotated data from the first two phases of the project for interpreting Commander instructions. Dialogue management policies are used to generate messages to both the Commander and interpreted actions for automated robot navigation. Our initial dialogue management policies are fairly simple, yet are able to handle a wide range of phenomena seen in our domain.

3 System Overview

ScoutBot consists of several software systems, running on multiple machines and operating systems, using two distributed agent platforms. The architecture utilized in simulation is shown in Figure 3. A parallel architecture exists for a real-world robot and environment. Components primarily involving language processing use parts of the Virtual Human (VH) architecture (Hartholt et al., 2013), and communicate using the Virtual Human Messaging (VHMSG), a thin layer on top of the publicly available ActiveMQ message protocol¹. Components primarily involving the real or simulated world, robot locomotion, and sensing are embedded in ROS². To allow the VHMSG and ROS modules to interact with each other, we created ros2vhmsg, software that bridges the two messaging architectures. The components are described in the remainder of this section.

The system includes several distinct worksta-

¹<http://activemq.apache.org>

²<http://www.ros.org>

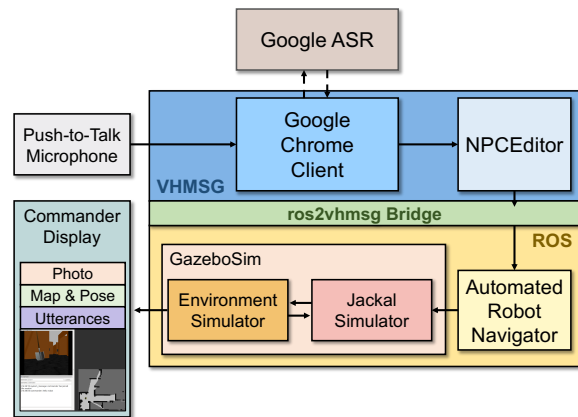


Figure 3: ScoutBot architecture interfacing with a simulated robot and environment: Solid arrows represent communications over a local network; dashed arrows indicate connections with external resources. Messaging for the spoken language interface is handled via VHMSG, while robot messages are facilitated by ROS. A messaging bridge, ros2vhmsg, connects the components.

tion displays for human participants. The Commander display is the view of the collaborative partner for the robot (lower left corner of Figure 3). This display shows a map of the robot’s local surroundings, the most recent photo the robot has sent, and a chat log showing text utterances from the robot. The map is augmented as new areas are explored, and updated according to the robot’s position and orientation. There are also displays for experimenters to monitor (and in the case of Wizards-of-Oz, engage in) the interaction. These displays show real-time video updates of what the robot can see, as well as internal communication and navigation aids.

3.1 VHMSG Components

VHMSG includes several protocols that implement parts of the Virtual Human architecture. We use the protocols for speech recognition, as well as component monitoring and logging. The NPCEditor and other components that use these protocols are available as part of the ICT Virtual Human Toolkit (Hartholt et al., 2013). These protocols have also been used in systems, as reported by Hill et al. (2003) and Traum et al. (2012, 2015). In particular, we used the adaptation of Google’s Automated Speech Recognition (ASR) API used in Traum et al. (2015). The NPCEditor (Leuski and Traum, 2011) was used for Natural Language Understanding (NLU) and dialogue management. The new `ros2vhmsg` component for bridging the messages was used to send instructions from the NPCEditor to the automated RN.

3.1.1 NPCEditor

We implemented NLU using the statistical text classifier included in the NPCEditor. The classifier learns a mapping from inputs to outputs from training data using cross-language retrieval models (Leuski and Traum, 2011). The dialogues collected from our first two experimental phases served as training data, and consisted of 1,500 pairs of Commander (user) utterances and the DM-Wizard’s responses. While this approach limits responses to the set that were seen in the training data, in practice it works well in our domain, achieving accuracy on a held out test set of 86%. The system is particularly effective at translating actionable commands to the RN for execution (e.g., *Take a picture*). It is robust at handling commands that include pauses, fillers, and other disfluent features (e.g., *Uh move um 10 feet*). It can also handle simple metric-based motion commands (e.g., *Turn right 45 degrees, Move forward 10 feet*) as well as action sequences (e.g., *Turn 180 degrees and take a picture every 45*). The system can interpret some landmark-based instructions requiring knowledge of the environment (e.g., *Go to the orange cone*), but the automated RN component does not yet have the capability to generate the low-level, landmark-based instructions for the robot.

Although the NPCEditor supports simultaneous participation in multiple conversations, extensions were needed for ScoutBot to support multi-floor interaction (Traum et al., 2018), in which

two conversations are linked together. Rather than just responding to input in a single conversation, the DM-Wizard in our first project phases often translates input from one conversational floor to another (e.g., from the Commander to the RN-Wizard, or visa versa), or responds to input with messages to both the Commander and the RN. These responses need to be consistent (e.g. translating a command to the RN should be accompanied by positive feedback to the Commander, while a clarification to the commander should not include an RN action command). Using the dialogue relation annotations described in Traum et al. (2018), we trained a hybrid classifier, including translations to RN if they existed, and negative feedback to the Commander where they did not. We also created a new dialogue manager policy that would accompany RN commands with appropriate positive feedback to the commander, e.g., response of “Moving..” vs “Turning...” as seen in #4 and #7 in Figure 2.

3.1.2 `ros2vhmsg`

`ros2vhmsg` is a macOS application to bridge the VHMSG and ROS components of ScoutBot. It can be run from a command line or using a graphical user interface that simplifies the configuration of the application. Both VHMSG and ROS are publisher-subscriber architectures with a central broker software. Each component connects to the broker. They publish messages by delivering them to the broker and receive messages by telling the broker which messages they want to receive. When the broker receives a relevant message, it is delivered to the subscribers.

`ros2vhmsg` registers itself as a client for both VHMSG and ROS brokers, translating and reissuing the messages. For example, when `ros2vhmsg` receives a message from the ROS broker, it converts the message into a compatible VHMSG format and publishes it with the VHMSG broker. Within `ros2vhmsg`, the message names to translate along with the corresponding ROS message types must be specified. Currently, the application supports translation for messages carrying either text or robot navigation commands. The application is flexible and easily extendable with additional message conversion rules. `ros2vhmsg` annotates its messages with a special metadata symbol and uses that information to avoid processing messages that it already published.



Figure 4: Real-world and simulated instances of the same environment.

3.2 ROS Components

ROS provides backend support for robots to operate in both real-world and simulated environments. Here, we describe our simulated testbed and automated navigation components. Developing automated components in simulation allows for a safe test space before software is transitioned to the real-world on a physical robot.

3.2.1 Environment Simulator

Running operations under ROS makes the transition between real-world and simulated testing straightforward. Gazebo³ is a software package compatible with ROS for rendering high-fidelity virtual simulations, and supports communications in the same manner as one would in real-world environments. Simulated environments were modeled in Gazebo after their real-world counterparts as shown in Figure 4. Replication took into consideration the general dimensions of the physical space, and the location and size of objects that populate the space. Matching the realism and fidelity of the real-world environment in simulation comes with a rendering trade-off: objects requiring a higher polygon count due to their natural geometries results in slower rendering in Gazebo and a lag when the robot moves in the simulation. As a partial solution, objects were constructed starting from their basic geometric representations, which could be optimized accordingly, e.g., the illusion of depth was added with detailed textures or shading on flat-surfaced objects. Environments rendered in Gazebo undergo a complex workflow to support the aforementioned requirements. Objects and their vertices are placed in an environment, and properties about collisions and gravity are encoded in the simulated environment.

3.2.2 Jackal Simulator

Clearpath provides a ROS package with a simulated model of the Jackal robot (Figure 1b) and

³<http://gazebosim.org/>

customizable features to create different simulated configurations. We configured our simulated Jackal to have the default inertial measurement unit, a generic camera, and a Hokuyo LIDAR laser scanner.

As the simulated Jackal navigates through the Gazebo environment, data from the sensors is relayed to the workstation views through rviz⁴, a visualization tool included with ROS that reads and displays sensor data in real-time. Developers can select the data to display by adding a panel, selecting the data type, and then selecting the specific ROS data stream. Both physical and simulated robot sensors are supported by the same rviz configuration, since rviz only processes the data sent from these sensors; this means that an rviz configuration created for data from a simulated robot will also work for a physical robot if the data stream types remain the same.

3.2.3 Automated Robot Navigator

Automated robot navigation is implemented with a python script and the ROSPY package⁵. The script runs on a ROS-enabled machine running the simulation. The script subscribes to messages from NPCEditor, which are routed through ros2vhmsg. These messages contain text output from the NLU classifier that issue instructions to the robot based on the user’s unconstrained speech.

A mapping of pre-defined instructions was created, with keys matching the strings passed from NPCEditor via ros2vhmsg (e.g., *Turn left 90 degrees*). For movement, the map values are a ROS TWIST message that specifies the linear motion and angular rotation payload for navigation. These TWIST messages are the only way to manipulate the robot; measures in units of feet and degrees cannot be directly translated. The mapping is straightforward to define for metric instructions. Included is basic coverage of moving forward or backward between 1 and 10 feet, and turning right and left 45, 90, 180, or 360 degrees. Pictures from the robot’s simulated camera can be requested by sending a ROS IMAGE message. The current mapping does not yet support collision avoidance or landmark-based instructions that require knowledge of the surrounding environment (e.g., *Go to the nearby chair*).

⁴<http://wiki.ros.org/rviz>

⁵<http://wiki.ros.org/rosipy>

4 Demo Summary

In the demo, visitors will see the simulated robot dynamically move in the simulated environment, guided by natural language interaction. Visitors will be able to speak instructions to the robot to move in the environment. Commands can be given in metric terms (e.g., #3 and #6 in Figure 2), and images requested (e.g., #9). Undecipherable or incomplete instructions will result in clarification subdialogues rather than robot motion (e.g., #1). A variety of command formulations can be accommodated by the NLU classifier based on the training data from our experiments. Visualizations of different components can be seen in: <http://people.ict.usc.edu/~traum/Movies/scoutbot-acl2018demo.wmv>.

5 Summary and Future Work

ScoutBot serves as a research platform to support experimentation. ScoutBot components will be used in our upcoming third through fifth development phases. We are currently piloting phase 3 using ScoutBot’s simulated environment, with human wizards. Meanwhile, we are extending the automated dialogue and navigation capabilities.

This navigation task holds potential for collaboration policies to be studied, such as the amount and type of feedback given, how to negotiate to successful outcomes when an initial request was underspecified or impossible to carry out, and the impact of miscommunication. More sophisticated NLU methodologies can be tested, including those that recognize specific slots and values or more detailed semantics of spatial language descriptions. The use of context, particularly references to relative landmarks, can be tested by either using the assumed context as part of the input to the NLU classifier, transforming the input before classifying, or deferring the resolution and requiring the action interpreter to handle situated context-dependent instructions (Kruijff et al., 2007).

Some components of the ScoutBot architecture may be substituted for different needs, such as different physical or simulated environments, robots, or tasks. Training new DM and RN components can make use of this general architecture, and the resulting components aligned back with ScoutBot.

Acknowledgments

This work was supported by the U.S. Army.

References

- Claire Bonial, Matthew Marge, Ashley Foots, Felix Gervits, Cory J Hayes, Cassidy Henry, Susan G Hill, Anton Leuski, Stephanie M Lukin, Pooja Moolchandani, Kimberly A. Pollard, David Traum, and Clare R. Voss. 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*.
- Arno Hartholt, David Traum, Stacy C Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. 2013. All Together Now Introducing the Virtual Human Toolkit. In *Proc. of Intelligent Virtual Agents*, pages 368–381. Springer.
- Randall Hill, Jr., Jonathan Gratch, Stacy Marsella, Jeff Rickel, William Swartout, and David Traum. 2003. Virtual Humans in the Mission Rehearsal Exercise System. *KI Embodied Conversational Agents*, 17:32–38.
- Geert-Jan M Kruijff, Pierre Lison, Trevor Benjamin, Henrik Jacobsson, and Nick Hawes. 2007. Incremental, Multi-Level Processing for Comprehending Situated Dialogue in Human-Robot Interaction. In *Symposium on Language and Robots*.
- Anton Leuski and David Traum. 2011. NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32(2):42–56.
- Matthew Marge, Claire Bonial, Kimberly A Pollard, Ron Artstein, Brendan Byrne, Susan G Hill, Clare Voss, and David Traum. 2016. Assessing Agreement in Human-Robot Dialogue Strategies: A Tale of Two Wizards. In *Proc. of Intelligent Virtual Agents*, pages 484–488. Springer.
- Laurel Riek. 2012. Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, 1(1).
- David Traum, Priti Aggarwal, Ron Artstein, Susan Foutz, Jillian Gerten, Athanasios Katsamanis, Anton Leuski, Dan Noren, and William Swartout. 2012. Ada and Grace: Direct Interaction with Museum Visitors. In *Proc. of Intelligent Virtual Agents*.
- David Traum, Kallirroi Georgila, Ron Artstein, and Anton Leuski. 2015. Evaluating Spoken Dialogue Processing for Time-Offset Interaction. In *Proc. of Special Interest Group on Discourse and Dialogue*, pages 199–208. Association for Computational Linguistics.
- David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, Cory Hayes, and Susan Hill. 2018. Dialogue Structure Annotation for Multi-Floor Interaction. In *Proc. of Language Resources and Evaluation Conference*.