Collaborative Exploration in Human-Robot Teams: What's in Their Corpora of Dialog, Video, & LIDAR Messages?

Clare R. Voss* **Taylor Cassidy**†* **Douglas Summers-Stay***
*Army Research Laboratory, Adelphi, MD 20783
†IBM T. J. Watson Research Center, Hawthorne, NY 10532
{clare.r.voss.civ,taylor.cassidy.ctr,douglas.a.summers-stay.civ}@mail.mil

Abstract

This paper briefly sketches new work-in-progress (i) developing task-based scenarios where human-robot teams collaboratively explore real-world environments in which the robot is immersed but the humans are not, (ii) extracting and constructing "multi-modal interval corpora" from dialog, video, and LIDAR messages that were recorded in *ROS bagfiles* during task sessions, and (iii) testing automated methods to identify, track, and align co-referent content both within and across modalities in these interval corpora. The pre-pilot study and its corpora provide a unique, empirical starting point for our longer-term research objective: characterizing the balance of explicitly shared and tacitly assumed information exchanged during effective teamwork.

1 Overview

Robots that are able to move into areas where people cannot during emergencies and collaboratively explore these environments by teaming with humans, have tremendous potential to impact search and rescue operations. For human-robot teams to conduct such shared missions, humans need to trust that they will be kept apprised, at a miniumum, of where the robot is and what it is sensing, as it moves about without them present.

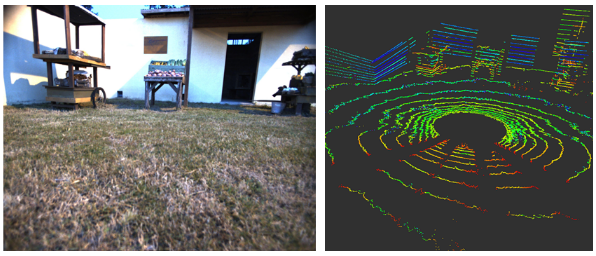


Figure 1: Outside View: Video Image & LIDAR.

To begin documenting the communication challenges humans face in taking a robot's perspective, we conducted a *pre-pilot* study[1] to record, identify and track the dialog, video, and LIDAR information that is explicitly shared by, or indirectly available to, members of human-robot teams when conducting collaborative tasks.

1.1 Approach

We enlisted colleagues to be the commander (C) or the human (R) controlling a mobile physical robot in such tasks. Neither could see the robot. Only R could "see for" the robot, via its onboard video camera and LIDAR. C and R communicated by text chat on their computers, as in this example,

R_41: I can see in the entrance.
C_42: Enter and scan the first room.

R_44: I see a door to the right and a door to the left.
C_45: Scan next open room on left.

Utterances R_41 & C_42 occur when the robot is outdoors (Fig. 1) and R_44 & C_45 occur after it moves indoors (Fig. 2). Although our approach resembles a *Wizard* **and** *Oz* paradigm (Riek, 2012),

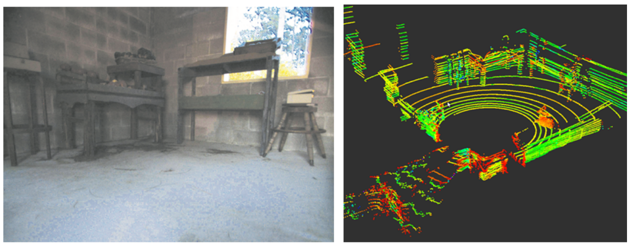


Figure 2: Inside View: Video Image & LIDAR. *Brightness and contrast of video image increased for print publication.*

[1] Statisticians say *pre-pilots* are for "kicking the tires," early-stage tests of scenarios, equipment, and data collection.

Proceedings of the of the EACL 2014 Workshop on Dialogue in Motion (DM), pages 43–47,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

with C as User and R as Wizard controlling the robot, there is no intent for R to deceive C.

In these dialog snippets, notice that the doors mentioned in R_44 are not visible in the image of that utterance's time interval and, even if they had been visible, their referents were context-dependent and ambiguous. How are the robot and human to refer to the same door? This challenge entails resolving several types of co-reference (linguistic, are they talking about the same door? visual, are they looking at the door? navigational, is one backing into a door no longer in view but previosuly stored in its map?) Successful communication on human-robot teams, *where humans send messages to direct robot movements and receive robot-processed messages as the robot navigates,* entails effective identification of named referents (such as doors), both within and across available modalities during exploratory tasks. The research question is, how might the identification and alignment of entities using combinations of (i) NLP on dialog, (ii) image processing on the video and LIDAR stream, with (iii) robot position, motion, and orientation coordinates, support more effective human-robot missions?

We conducted the pre-pilot study with ten trial sessions to collect multi-modal data from C-R and R-only scenarios (Table 1). Each session involved a single participant playing the role of R with control over the physical robot, or two participants, one person playing R and one playing C.

Team	R's Task
R only	Rotate in place and describe surroundings.
R only	Move along road, describe surroundings.
C, R	Follow C's guidance in navigating building's perimeter, describe surroundings.
C, R	Follow C's guidance in searching buildings for specified objects.

Table 1: Pre-pilot Scenarios.

Participants sat indoors and could not see the robot outside, roughly 30 meters away. In each session, R was instructed to act as though he or she were situated in the robot's position and to obey C. R was to consider the robot's actions as R's own, and to consider available video and LIDAR point cloud feeds as R's own perceptions.

1.2 Equipment

All participants worked from their own computers. Each was instructed, for a given scenario, to be either C or R and to communicate by text only. On their screen they saw a dedicated dialog (chat) window in a Linux terminal. For sessions with both C and R, the same dialog content (the ongoing sequence of typed-in utterances) appeared in the dialog window on each of their screens.

The physical robot ran under the Robot Operating System (ROS) (Quigley et al., 2009), equipped with a video camera, laser sensors, magnetometer, GPS unit, and rotary encoders. R could "see for the robot" via two ROS *rviz* windows with live feeds for video from the robot's camera and constructed 3D point cloud frames.[2] R had access to rotate and zoom functions to alter the screen display of the point cloud. C saw only a static bird's-eye-view map of the area. R remotely controlled over a network connection the robot's four wheels and its motion, using the left joystick of an X-Box controller.

1.3 Collection

During each session, all data from the robot's sensors and dialog window was recorded via the *rosbag* tool and stored in a single *bagfile*.[3] A bagfile contains typed *messages*. Each message contains a timestamp (specified at nanosecond granularity) and values for that message type's attributes. Message types *geometry_msgs/PoseStamped*, for example, contain a time stamp, a three-dimensional location vector and a four-dimensional orientation vector that indicates an estimate of the robot's location and the direction in which it is facing. The robot's rotary encoders generate these messages as the robot moves. The primary bagfile message types most relevant to our initial analyses[4] were:

1) instant_messenger/StringStamped
   that included speaker id, text utterances
2) sensor_msgs/PointCloud2
   that included LIDAR data
3) sensor_msgs/CompressedImage
   with compressed, rectified video images
4) sensor_msgs/GPS, with robot coordinates

Message types are packaged and *published* at different rates: some are published automatically at regular intervals (e.g., image frames), while others depend on R, C, or robot activity (e.g., dialog utterances). And the specific rate of publication for some message types can be limited at times by network bandwidth constraints (e.g. LIDAR data). Summary statistics for our initial pre-pilot collec-

[2] LIDAR measures distance from robot by illuminating targets with robot lasers and generates point cloud messages.

[3] http://wiki.ros.org/rosbag

[4] We omit here details of ROS *topics*, *transformation messages*, and other sensor data collected in the pre-pilot.

tion consisting of ten task sessions conducted over two days, and that together spanned roughly five hours in real-time, are presented in Table 2.

#bagfile msgs	$15,131K$	#dialog utts	434
min per sn	$140,848$	min per sn	15
max per sn	$3,030K$	max per sn	116
#tokens	$3,750$	#image msgs	$10,650$
min per sn	200	min per sn	417
max per sn	793	max per sn	$1,894$
#unique words	568	#LIDAR msgs	$8,422$
min per sn	84	min per sn	215
max per sn	176	max per sn	$2,250$

Table 2: Collection Statistics (sn = session).

2 From Collection to Interval Corpora

After collecting millions of messages in the pre-pilot with content in different modalities, the immediate research challenge has been identifying the time interval that covers the messages directly related to the content in each utterance.

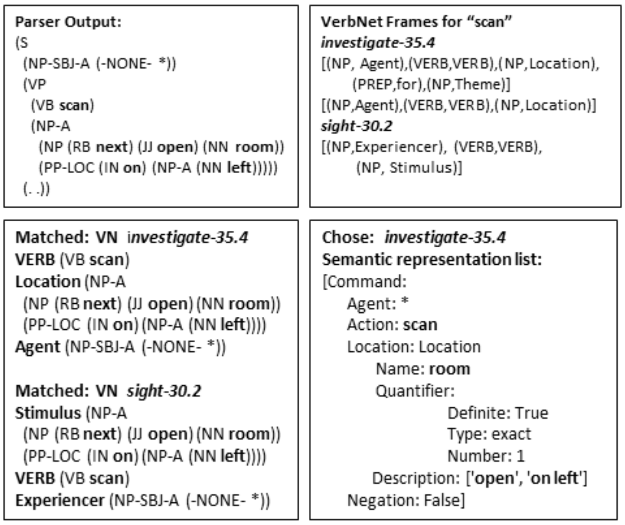
We extracted each utterance message u and its corresponding time stamp t. For a given u, we extracted the five image, five point cloud, and five GPS messages immediately preceding and the five of each immediately following u, based on message time-stamps, for a total of thirty sensor messages per utterance. These message types were published independent of the robot's movement, approximately once per second. In the second phase, we assigned the earliest and latest time stamp from the first-phase messages to delimit an interval $[t_s, t_e]$ and conducted another extraction round from the bagfile, this time pulling out all messages with time stamps in that interval as published by the rotary encoders, compass, and inertial measurement unit, only when the robot moved. The messages from both phases constitute a ten-second *interval corpus* for u.

These interval corpora serve as a first approximation at segmenting the massive stream published at nanosecond-level into units pertaining to commander-robot dialog during the task at hand. With manual inspection, we found that many automatically-constructed intervals do track relevant changes in the robot's location. For example, the latest interval in a task's time sequence that was constructed with the robot being outside a building is distinct from the first interval that covers when the robot moves inside the building.[5]

3 Corpora Language Processing

Each utterance collected from the sessions was tokenized, parsed, and semantically interpreted using SLURP (Brooks et al., 2012), a well-tested NLP front-end component of a human-robot system.[6] The progression in SLURP's analysis pipeline for utterance C_45 is shown in Figure 3.

SLURP extracts a parse tree (top-left), identifies a sub-tree that constitutes a verb-argument structure, and enumerates possibly matching sense-specific *verb frames* from VerbNet (Schuler, 2005) (bottom-left). VerbNet provides a syntactic to semantic role mapping for each frame (top-right). SLURP selects the best mapping and generates a compact semantic representation (bottom-right).[7] In this example, the correct sense of "scan" is selected (*investigate-35.4*) along with a frame that matches the syntactic parse. Overall, half the commands run through SLURP generated a semantic interpretation. Of the other half, roughly one quarter failed or had errors at parsing and the other quarter at the argument matching stage.



```
Parser Output:
(S
 (NP-SBJ-A (-NONE- *))
 (VP
  (VB scan)
  (NP-A
   (NP (RB next) (JJ open) (NN room))
   (PP-LOC (IN on) (NP-A (NN left)))))
 (. .))
```

```
VerbNet Frames for "scan"
investigate-35.4
[(NP, Agent),(VERB,VERB),(NP,Location),
     (PREP,for),(NP,Theme)]
[(NP,Agent),(VERB,VERB),(NP,Location)]
sight-30.2
[(NP,Experiencer), (VERB,VERB),
     (NP, Stimulus)]
```

```
Matched: VN investigate-35.4
VERB (VB scan)
Location (NP-A
 (NP (RB next) (JJ open) (NN room))
 (PP-LOC (IN on) (NP-A (NN left))))
Agent (NP-SBJ-A (-NONE- *))

Matched: VN sight-30.2
Stimulus (NP-A
 (NP (RB next) (JJ open) (NN room))
 (PP-LOC (IN on) (NP-A (NN left))))
VERB (VB scan)
Experiencer (NP-SBJ-A (-NONE- *))
```

```
Chose: investigate-35.4
Semantic representation list:
[Command:
   Agent: *
   Action: scan
   Location: Location
       Name: room
       Quantifier:
                 Definite: True
                 Type: exact
                 Number: 1
       Description: ['open', 'on left']
   Negation: False]
```

Figure 3: Analyses of *Scan next open room on left.*

Our next step is to augment SLURP's lexicon and retrain a parser for new vocabulary so that we can directly map semantic structures of the pre-pilot corpora into ResearchCyc[8], an extensive ontology, for cross-reference to other events and objects, already stored and possibly originated as visual input. Following McFate (2010), we will test

[5] This appears likely due to the paced descriptions in R's utterances. Another pre-pilot is needed to test this hypothesis.

[6] https://github.com/PennNLP/SLURP.

[7] Verbnet associates each frame with a conjunction of boolean semantic predicates that specify how and when event participants interact, for an event variable (not shown).

[8] ResearchCyc and CycL are trademarks of Cycorp, Inc.

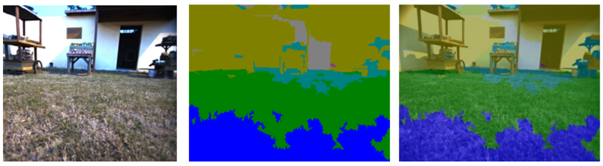


Figure 4: Outside View: Image, Zones, Overlay

the mapping of matched VerbNet frames to ResearchCyc's semantic predicates to assess its lexical coverage for our corpora.

4 Image Processing

Interval corpus images were labelled by a neural network trained for visual scene classification (Munoz, 2013) of nine material classes: dirt, foliage, grass, road, sidewalk, sky, wall, wood, and ground cover (organic debris). Figures 4 and 5 show the images from Figures 1 and 2 with two additional versions: one with colored zones for system-recognized class boundaries and another with colored zones as trasparent overlays on the original. The classes differentiate terrain types that work well with route-finding techniques that leverage them in selecting traversible paths. As the robot systems are enhanced with more sophisticated path planning software, that knowledge may be combined with recognized zones to send team members messages about navigation problems as the robot explores where they cannot go.

Accuracy is limited at the single image level: the actual grass in Figure 4 is mostly mis-classified as *dirt* (blue) along with some correctly identified *grass* (green), while the floor in Figure 5 is misclassified as *road,* although much of what shows through the window is correctly classified as *foliage.* We are experimenting with automatically assigning natural language (NL) labels to a range of objects and textures recognized in images from other larger datasets. We can retrieve labeled images stored in ResearchCyc via NL query converted into CycL, allowing a commander to, for example, ask questions about objects and regions using terms related to but not necessarily equal to the original recognition system-provided labels.

5 Related Work

We are aware of no other multi-modal corpora obtained from human-robot teams conducting exploratory missions with collected dialog, video and other sensor data. Corpora with a robot recording similar data modalities do exist (Green et al., 2006; Wienke et al., 2012; Maas et al., 2006) but for fundamentally different tasks. Tellex et al. (2011) and Matuszek et al. (2012) pair commands with formal plans without dialog and Zender et al. (2008) and Randelli et al. (2013) build multi-level maps but with a situated commander.

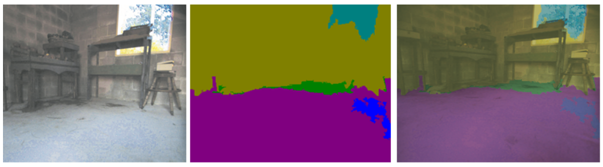


Figure 5: Inside View: Image, Zones, Overlay. *Brightness and contrast of video image and overlay increased for print publication.*

Eberhard et al. (2010)'s CReST corpus contains a set-up similar to ours minus the robot; a human task-solver wears a forward-facing camera instead. The SCARE corpus (Stoia et al., 2008) records similar modalities but in a virtual environment, where C has full access to R's video feed. Other projects yielded corpora from virtual environments that include route descriptions without dialog (Marge and Rudnicky, 2011; MacMahon et al., 2006; Vogel and Jurafsky, 2010) or referring expressions without routes (Schütte et al., 2010; Fang et al., 2013), assuming pre-existing abstractions from sensor data.

6 Conclusion and Ongoing Work

We have presented our pre-pilot study with data collection and corpus construction phases. This work-in-progress requires further analysis. We are now processing dialog utterances for more systematic semantic interpretation using disambiguated VerbNet frames that map into ResearchCyc predicates. We will run object recognition software retrained on a broader range of objects so that it can be applied to images that will be labelled and stored in ResearchCyc micro-worlds for subsequent co-reference with terms in the dialog utterances. Ultimately we want to establish in real time links across parts of messages in different modalities that refer to the same abstract entities, so that humans and robots can share their separately-obtained knowledge about the entities and their spatial relations — whether seen, sensed, described, or inferred — when communicating on shared tasks in environments.

Acknowledgments

Over a dozen engineers and researchers assisted us in many ways before, during, and after the prepilot, providing technical help with equipment and data collection, as well as participating in the prepilot. We cannot list everyone here, but special thanks to Stuart Young for providing clear guidance to everyone working with us.

References

Daniel J. Brooks, Constantine Lignos, Cameron Finucane, Mikhail S. Medvedev, Ian Perera, Vasumathi Raman, Hadas Kress-Gazit, Mitch Marcus, and Holly A. Yanco. 2012. Make it so: Continuous, flexible natural language interaction with an autonomous robot. In *Proc. AAAI*, pages 2–8.

Kathleen M. Eberhard, Hannele Nicholson, Sandra Kübler, Susan Gundersen, and Matthias Scheutz. 2010. The indiana "cooperative remote search task" (crest) corpus. In *Proc. LREC*.

Rui Fang, Changsong Liu, Lanbo She, and Joyce Y. Chai. 2013. Towards situated dialogue: Revisiting referring expression generation. In *Proc. EMNLP*, pages 392–402.

Anders Green, Helge Httenrauch, and Kerstin Severinson Eklundh. 2006. Developing a contextualized multimodal corpus for human-robot interaction. In *Proc. LREC*.

Jan F. Maas, Britta Wrede, and Gerhard Sagerer. 2006. Towards a multimodal topic tracking system for a mobile robot. In *Proc. INTERSPEECH*.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proc. AAAI*, pages 1475–1482.

Matthew Marge and Alexander I Rudnicky. 2011. The teamtalk corpus: Route instructions in open spaces. In *Proc. RSS, Workshop on Grounding Human-Robot Dialog for Spatial Tasks*.

Cynthia Matuszek, Evan Herbst, Luke S. Zettlemoyer, and Dieter Fox. 2012. Learning to parse natural language commands to a robot control system. In *Proc. ISER*, pages 403–415.

Clifton McFate. 2010. Expanding verb coverage in cyc with verbnet. In *Proc. ACL, Student Research Workshop*, pages 61–66.

Daniel Munoz. 2013. *Inference Machines: Parsing Scenes via Iterated Predictions*. Ph.D. thesis, Carnegie Mellon University.

Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully B. Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. 2009. ROS: an open-source robot operating system. In *Proc. ICRA, Workshop on Open Source Software*.

Gabriele Randelli, Taigo Maria Bonanni, Luca Iocchi, and Daniele Nardi. 2013. Knowledge acquisition through human–robot multimodal interaction. *Intelligent Service Robotics*, 6(1):19–31.

Laurel D Riek. 2012. Wizard of oz studies in hri: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1).

Karin Kipper Schuler. 2005. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Niels Schütte, John D. Kelleher, and Brian Mac Namee. 2010. Visual salience and reference resolution in situated dialogues: A corpus-based evaluation. In *Proc. AAAI, Fall Symposium: Dialog with Robots*.

Laura Stoia, Darla Magdalena Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2008. Scare: a situated corpus with annotated referring expressions. In *Proc. LREC*.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proc. AAAI*.

Adam Vogel and Daniel Jurafsky. 2010. Learning to follow navigational directions. In *Proc. ACL*, pages 806–814.

Johannes Wienke, David Klotz, and Sebastian Wrede. 2012. A framework for the acquisition of multimodal human-robot interaction data sets with a whole-system perspective. In *Proc. LREC, Workshop on Multimodal Corpora for Machine Learning*.

Hendrik Zender, O Martínez Mozos, Patric Jensfelt, G-JM Kruijff, and Wolfram Burgard. 2008. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, 56(6):493–502.