Jointly Improving Parsing and Perception for Natural Language Commands through Human-Robot Dialog

Jesse Thomason,[*] **Aishwarya Padmakumar,**[†] **Jivko Sinapov,**[‡]
Nick Walker,[*] **Yuqian Jiang,**[†] **Harel Yedidsion,**[†] **Justin Hart,**[†]
Peter Stone,[†] and **Raymond J. Mooney**[†]
[*]Paul G. Allen School of Computer Science and Engineering, University of Washington
[†]Department of Computer Science, University of Texas at Austin
[‡]Department of Computer Science, Tufts University

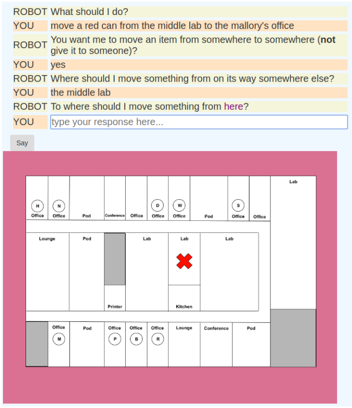
Abstract

In this work, we present methods for parsing natural language to underlying meanings, and using robotic sensors to create multi-modal models of perceptual concepts. We combine these steps towards language understanding into a holistic agent for jointly improving parsing and perception on a robotic platform through human-robot dialog. We train and evaluate this agent on Amazon Mechanical Turk, then demonstrate it on a robotic platform initialized from that conversational data. Our experiments show that improving both parsing and perception components from conversations improves communication quality and human ratings of the agent.

1 Introduction

Pre-programming robots with fixed language understanding components limits them, since different speakers and environments use and elicit different words. Rather than force humans to use language that robots around them can understand, robots should dynamically adapt—continually learning new language constructions and perceptual concepts as they are used in context.

Untrained human users providing natural language commands to robots expect world knowledge, perceptual knowledge, and semantic understanding from verbal robots. We present a holistic system for jointly improving parsing and perception on a robotic system for natural language commands through human-robot dialog. This learning agent uses clarification questions in a conversation with a human partner to understand language commands. The agent induces additional training data for a semantic parser, similar to prior work (Thomason et al., 2015), strengthening its parsing over time. The agent also uses opportunistic active learning (Thomason et al., 2017) to ask questions about nearby objects to refine multi-modal perceptual concept models (Thomason et al., 2016) on-the-fly during command dialogs.



ROBOT What should I do?
YOU move a red can from the middle lab to the mallory's office
ROBOT You want me to move an item from somewhere to somewhere (**not** give it to someone)?
YOU yes
ROBOT Where should I move something from on its way somewhere else?
YOU the middle lab
ROBOT To where should I move something from here?
YOU type your response here...

Figure 1: Mechanical Turk web interface.

We evaluate this language understanding agent on Mechanical Turk (Figure 1), and implement the agent on a physical robot.[1]

2 Conversational Agent

We implement and evaluate a conversational dialog agent that uses a semantic parser to translate human utterances into semantic meaning representations, then grounds those meaning representations using both a static knowledge base of facts about an office environment and perceptual concept models that consider multi-modal representations of physical objects.[2]

We build multi-modal concept models to con-

[1] The demonstration video can be viewed at https://youtu.be/PbOfteZ_CJc

[2] The source code for this conversational dialog agent, as well as the experiments described in the following section, can be found at https://github.com/thomason-jesse/grounded_dialog_agent

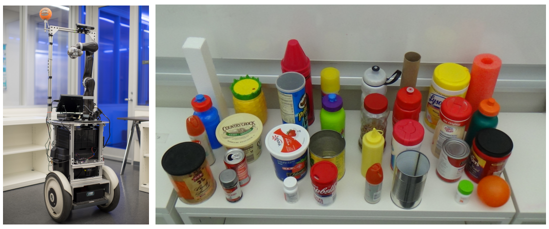


Figure 2: The robot used in our experiment, and the objects explored by the robot for grounding perceptual predicates.

nect robot perception to concept labels. We use multi-modal feature representations across various sensorimotor contexts by exploring those objects with a robot arm, performing behaviors such as grasping, lifting, and dropping objects, in addition to extracting visual information from them (Figure 2) (Sinapov et al., 2016; Thomason et al., 2016).

We connect these feature representations of objects to language labels by learning discriminative classifiers on the feature spaces for each perceptual language concept (e.g., *red* or *heavy*) (Sinapov et al., 2014; Thomason et al., 2016).

Since there are multiple possible groundings for ambiguous utterances like *the office* and varied confidences for perceptual concept models on different objects, we create a confidence distribution over the possible groundings for a semantic parse. This confidence probability distribution is used as part of an update procedure for helping the agent understand the user's intent during dialog.

We implement a conversational dialog agent for command understanding similar to that in previous work (Thomason et al., 2015). The differences between this agent and the previous one are: 1) grounding semantic parses in both static knowledge and perceptual knowledge; 2) dynamically adding new ontological predicates for novel perceptual concepts; and 3) leveraging opportunistic active learning for refining perceptual concept models on-the-fly.

Dialog begins with a human user commanding the robot to perform a task. The agent maintains a belief state modeling the unobserved true task in the user's mind, and uses the language signals from the user to infer it. The command is first parsed by the agent's semantic parser, then grounded against static and perceptual knowledge. These groundings are used to update the agent's belief state, and the agent engages in a clarification dialog to refine that belief.

When describing objects in the real world, humans can use words the agent has never heard before. In this work, we examine the lexical neighbors of unknown words in word-embedding space, and ask the user directly whether the unseen words are perceptual in nature if any neighbors are. We then check for synonyms (e.g., users marked *tall* as a synonym for *long*), and add a new perceptual concept if no known words are suitable synonyms (e.g., *red*, a neighbor of *yellow*, was added in our experiments in this way).

We introduce opportunistic active learning questions as a sub-dialog routine for the agent, in which it can query about objects *local* to the human and the robot to refine its perceptual concept models before applying them to the *remote* test object items, similar to previous work (Thomason et al., 2017).

3 Experiments

Mechanical Turk users instruct the agent in three tasks: navigation, delivery, and relocation. We deploy the agent in a simulated office environment populated by rooms, people, and object items. We randomly split the set of possible tasks into initialization (10%), train (70%), and test sets (20%).

Sixteen users (graduate students at the university across several fields) engaged with a faux-agent on initialization set tasks using the web interface. We used these commands as a scaffold on which to build an ontology, lexicon, and initial utterance-semantic parse pairs. Of them, 44 pairs, D_0, were used to train an initial parser.

We use these initial parsing resources to create a baseline agent $\mathcal{A}_1$ with a parser $\mathcal{P}_1$ trained only on the initialization pairs D_0 mentioned above and concept models for several predicates $P_{c,1}$, but with no initial object examples against which to train them.

Three training phases i are carried out by agent $\mathcal{A}_i$, after which parser $\mathcal{P}_{i+1}$ and concept predicates $P_{c,i+1}$ are trained to instantiate agent $\mathcal{A}_{i+1}$. Agent $\mathcal{A}_4$ with parser $\mathcal{P}_4$ and perception models $P_{c,4}$ is tested by interacting with users trying to accomplish tasks from the unseen test set of tasks. We also test an ablation agent, $\mathcal{A}_4^*$, with parser $\mathcal{P}_1^*$ and perception models $P_{c,4}$ (trained perception with simply initialized parser).

Table 1 gives a breakdown of the numbers of workers who engaged with our HITs.

Condition	**Number of Workers**					
	Submitted HIT	Completed Tasks	Vetted	Nav. Correct	Del. Correct	Rel. Correct
Train (A_1, A_2, A_3)	297	162	113	36	44	18
Untrained (A_1)	150	67	44	17	22	10
Test* (A_4^*)	148	83	50	20	29	10
Test (A_4)	143	79	42	16	23	10

Table 1: Breakdown of the number of workers in our experiment. We here count only workers that **submitted** the HIT with the correct code. Workers that **completed** all tasks and the survey finished the HIT entirely. **Vetted** workers' data was kept for evaluation after basic checks. The Train condition (A_1, A_2, A_3 agents) draws from the training set of tasks, while the Untrained (A_1 untrained agent), Test* (A_4^* agent with trained perception and untrained parser), and Test (A_4 agent with trained parser and perception) conditions draw from the test set of tasks.

For our embodied demonstration, we use the BWIBot (Khandelwal et al., 2014, 2017), equipped with a Kinova MICO arm, an Xtion ASUS Pro camera, a Hokuyo lidar, a Blue Snowball microphone, and a speaker. Speech transcripts are provided by the Google Speech API[3] and speech synthesis is performed with the Festival Speech Synthesis System.[4] Tabletop perception, required for both the dialog interaction and the execution of the resulting command, is implemented with RANSAC (Fischler and Bolles, 1981) plane fitting and Euclidean clustering as provided by Point Cloud Library (Rusu and Cousins, 2011).

Figure 4 gives quantitative and qualitative measures of performance in Mechanical Turk, as well as some snapshots from the embodied demonstration. The agent acquires new perceptual concept models (25 in total), and synonym words for existing concepts, during the three phases of training. Table 3 shows the learned perceptual concept model for *can* on test objects.

4 Conclusion

In this article, we presented a holistic system for jointly improving semantic parsing and grounded perception on a robotic system for interpreting natural language commands during human-robot dialog. We show, via a large-scale Mechanical Turk experiment, that users are better able to communicate tasks and rate the system more usable after this dialog-based learning procedure. We embody this learning agent in a physical robot platform to demonstrate its learning abilities for the non-visual word *rattling*.

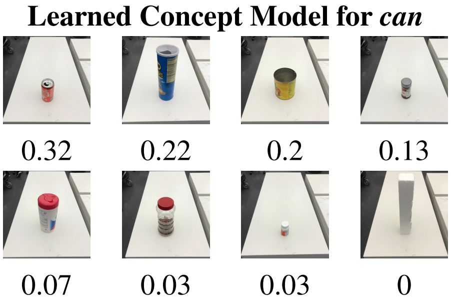

Learned Concept Model for *can*

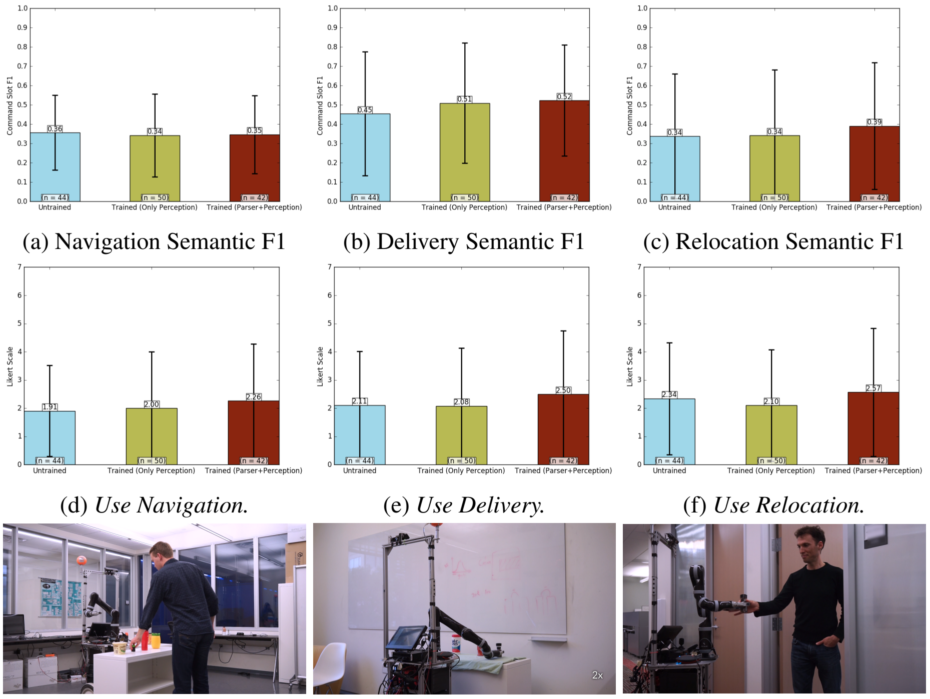
Figure 3: The perceptual concept model learned for *can* after training from conversations with human users.

Acknowledgements

This work was supported by a National Science Foundation Graduate Research Fellowship to the first author, an NSF EAGER grant (IIS-1548567), and an NSF NRI grant (IIS-1637736). Portions of this work have taken place in the Learning Agents Research Group (LARG) at UT Austin. LARG research is supported in part by NSF (CNS-1305287, IIS-1637736, IIS-1651089, IIS-1724157), TxDOT, Intel, Raytheon, and Lockheed Martin. Peter Stone serves on the Board of Directors of Cogitai, Inc. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research. We would like to thank Subhashini Venugopalan for her help in engineering deep visual feature extraction for language grounding.

[3] https://cloud.google.com/speech/

[4] http://www.cstr.ed.ac.uk/projects/festival/



(a) Navigation Semantic F1 (b) Delivery Semantic F1 (c) Relocation Semantic F1

(d) *Use Navigation.* (e) *Use Delivery.* (f) *Use Relocation.*

(g) The robot asks questions about items to learn *rattling*. (h) The robot decides grasps a *rattling container*. (i) The robot hands over the item at the destination.

Figure 4: **Top:** The average semantic slot f scores between the semantic roles in the target task and the task confirmed by the user. **Middle:** Survey prompt responses about usability. **Bottom:** The agent learns a new word, *rattling*, which requires perception using the auditory sensing modality, and uses this new concept model to correctly identify and move the target item

References

Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395.

Piyush Khandelwal, Fangkai Yang, Matteo Leonetti, Vladimir Lifschitz, and Peter Stone. 2014. Planning in Action Language $\mathcal{BC}$ while Learning Action Costs for Mobile Robots. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*.

Piyush Khandelwal, Shiqi Zhang, Jivko Sinapov, Matteo Leonetti, Jesse Thomason, Fangkai Yang, Ilaria Gori, Maxwell Svetlik, Priyanka Khante, Vladimir Lifschitz, J. K. Aggarwal, Raymond Mooney, and Peter Stone. 2017. Bwibots: A platform for bridging the gap between ai and human–robot interaction research. *The International Journal of Robotics Research (IJRR)*, 36.

Radu Bogdan Rusu and Steve Cousins. 2011. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China.

Jivko Sinapov, Priyanka Khante, Maxwell Svetlik, and Peter Stone. 2016. Learning to order objects using haptic and proprioceptive exploratory behaviors. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*.

Jivko Sinapov, Connor Schenck, and Alexander Stoytchev. 2014. Learning relational object categories using behavioral exploration and multimodal perception. In *IEEE International Conference on Robotics and Automation*.

Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Justin Hart, Peter Stone, and Raymond J. Mooney. 2017. Opportunistic active learning for grounding natural language descriptions. In *Proceedings of the 1st Annual Conference on Robot Learning (CoRL-17)*, volume 78, pages 67–76. Proceedings of Machine Learning Research.

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond Mooney. 2016. Learning multi-modal grounded linguistic semantics by playing "I spy". In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3477–3483.

Jesse Thomason, Shiqi Zhang, Raymond Mooney, and Peter Stone. 2015. Learning to interpret natural language commands through human-robot dialog. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1923–1929.