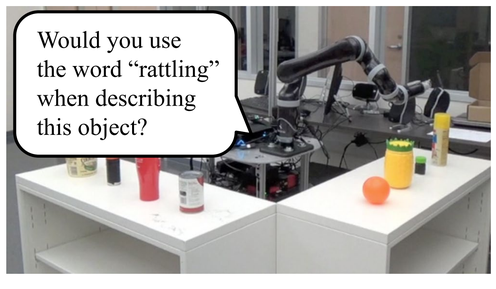
Jointly Improving Parsing and Perception for Natural Language Commands through Human-Robot Dialog

Jesse Thomason,* Aishwarya Padmakumar,† Jivko Sinapov,‡ Nick Walker,* Yuqian Jiang,† Harel Yedidsion,† Justin Hart,† Peter Stone,† and Raymond J. Mooney†

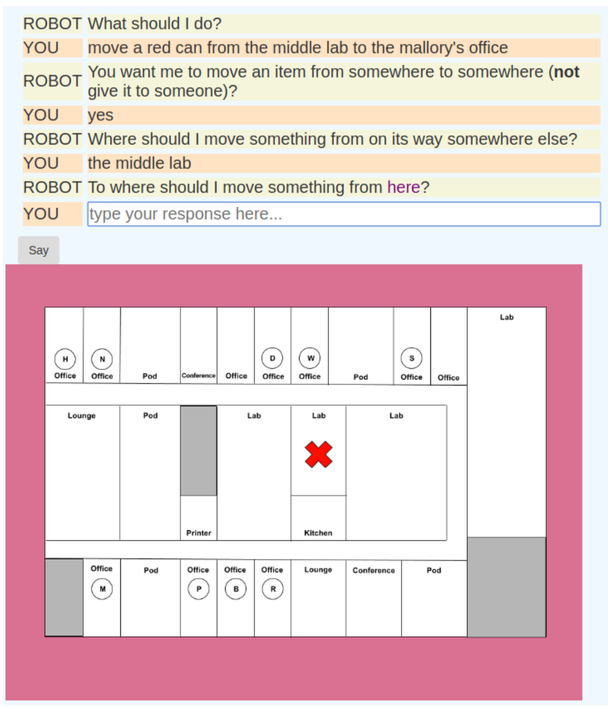
*University of Washington;
†University of Texas at Austin; ‡Tufts University

Learning via Human-Robot Dialog

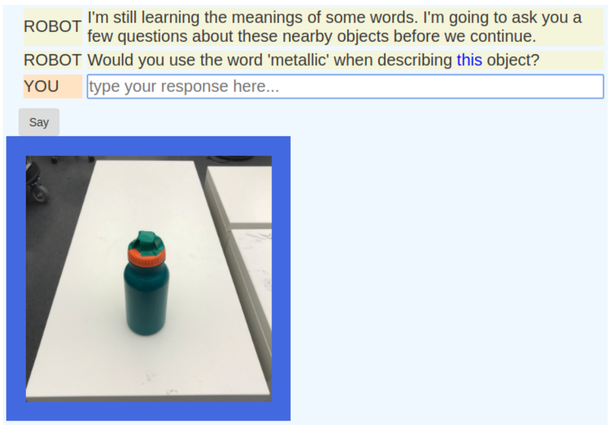
Natural language understanding for robotics can require substantial domain- and platform-specific engineering. One way to alleviate engineering for a new domain is to enable robots in human environments to adapt dynamically—continually learning new language constructions and perceptual concepts. We present an end-to-end pipeline for translating natural language commands to robot actions, and use clarification dialogs to jointly improve parsing and concept grounding.



Through dialog, a robot agent can acquire task-relevant information from a human on the fly. Here, *rattling* is a new concept the agent learns with human guidance in order to pick out a remote target object later on.



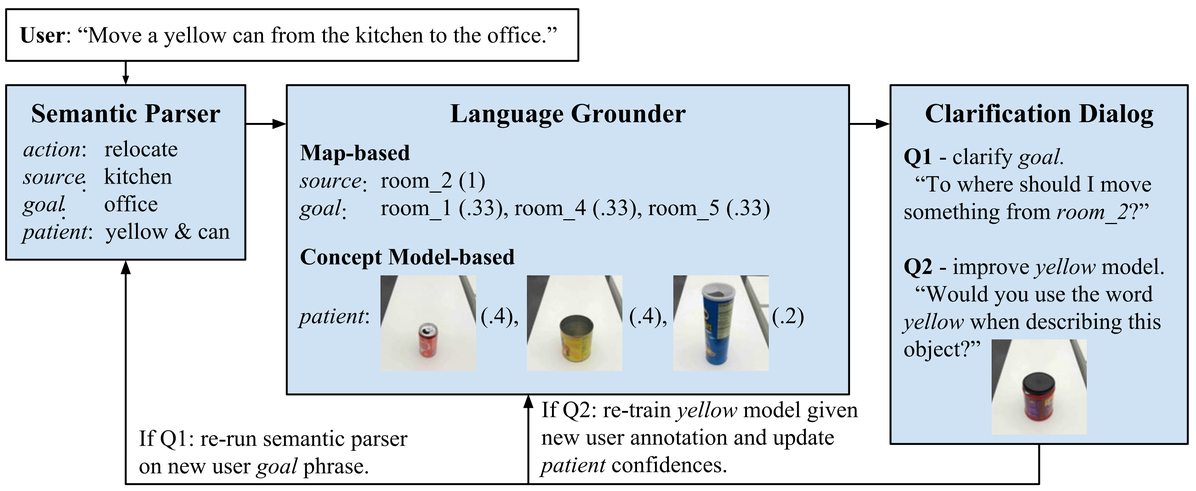
We train the agent in simulation with real human users on Mechanical Turk. The user gives a command to the learning agent, and the agent replies with questions to clarify the command. Each clarification generates a new training example for the agent's semantic parser, enabling retraining from conversational data.



In addition to asking for semantic clarifications, the agent can ask whether a concept word applies to a specific object, and for examples. Learning these models while retraining the semantic parser connects words to concept models corresponding to physical properties of objects like color, weight, and sound.

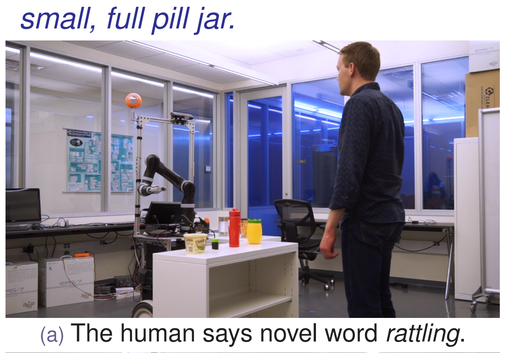
Asking questions can help a robot understand compositional language and grounded word meanings during human-robot dialogs.

Talk to the agent! `https://bit.ly/2W3jiJP`.



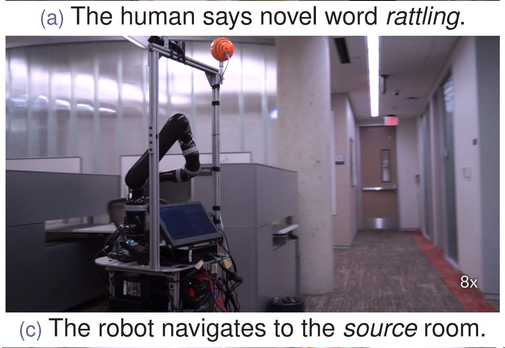
User commands are parsed into semantic slots (**left**), which are grounded (**center**) using either a known map (for rooms and people) or learned concept models (for objects) to a distribution over possible satisfying constants (e.g., all rooms that can be described as an "office"). A clarification dialog (**right**) is used to recover from ambiguous or misunderstood slots (e.g., *Q1*), and to improve concept models on the fly (e.g., *Q2*).

R: What should I do?
H: Move a rattling container from the lounge by the conference room to Bob's office.
R: Does rattling refer to a property of an object?
H: Yes.
R: Does rattling mean the same thing as empty?
H: No.
R: I am still learning the meanings of some words.
R: Show me an object you could use the word rattling when describing, or say "none of them."
H: Watch.
R: I am getting the objects on the table into focus... Okay, I see them.
H: This one. *The human picks up a small, full pill jar.*
R: I see.
R: Show me an object you could not use the word rattling when describing, or say "all of them."
H: Watch. This one. *The human picks up a red water bottle.*
R: I see.
R: Show me an object you could use the word rattling when describing, or say "none of them."
H: Watch. This one. *The human picks up a container half-full of beans.*
R: I see.
R: Thanks. Now back to business.
R: You want me to move an item from 3.514 to somewhere?
H: Yes.
R: I should move something from 3.514 to 3.510?
H: Yes.

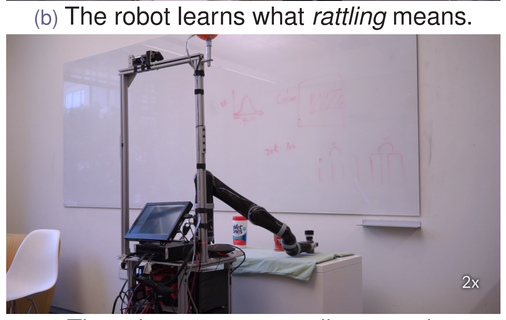


(a) The human says novel word *rattling*.

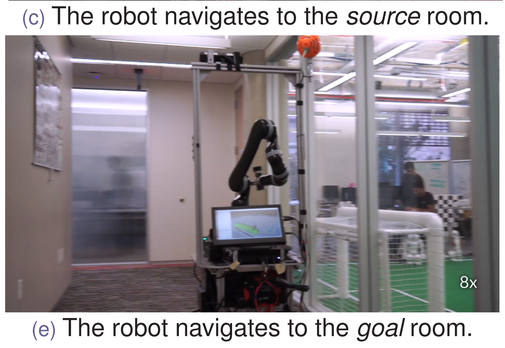
(b) The robot learns what *rattling* means.



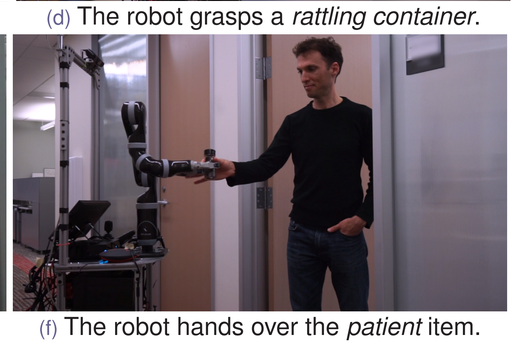
(c) The robot navigates to the *source* room.



(d) The robot grasps a *rattling container*.



(e) The robot navigates to the *goal* room.



(f) The robot hands over the *patient* item.

The robot learns a new word, *rattling*, which requires auditory perception, and is then able to navigate to the specified room, identify a *rattling container* item, and deliver the item to the specified destination.

Mechanical Turk Evaluation

After training on batches of dialogs with users on a set of training tasks, we test agents against unseen test tasks. We compare an *Initial* agent against one with a *Trained** *Perception* module, and one with *Trained Parsing* and *Perception* modules. We measure the number of clarification questions asked during the dialog. This metric should decrease as the agent refines its parsing and perception modules, needing to ask fewer questions about the unseen locations and objects in the test tasks. We also compare users' answers to usability questions answered on a 7-point Likert scale: from *Strongly Disagree* (1) to *Strongly Agree* (7).

$\mathcal{A}$	Clarification Questions ↓ Navigation (p)	Delivery (p)	Relocation (p)
In	3.02 ± 6.48	6.81 ± 8.69	22.3 ± 9.15
Tr*	4.05 ± 8.81(.46)	8.16 ± 13.8(.53)	23.5 ± 6.07(.67)
Tr	1.35 ± 4.44(.11)	7.50 ± 9.93(.72)	19.6 ± 7.89(.47)

The average number of clarification questions agents asked among successful dialogs. Also given are the p-values of a Welch's t-test between the **Tr***ained** (*Perception*) and **Tr***ained* (*Parsing+Perception*) model ratings against the **In***itial* model ratings.

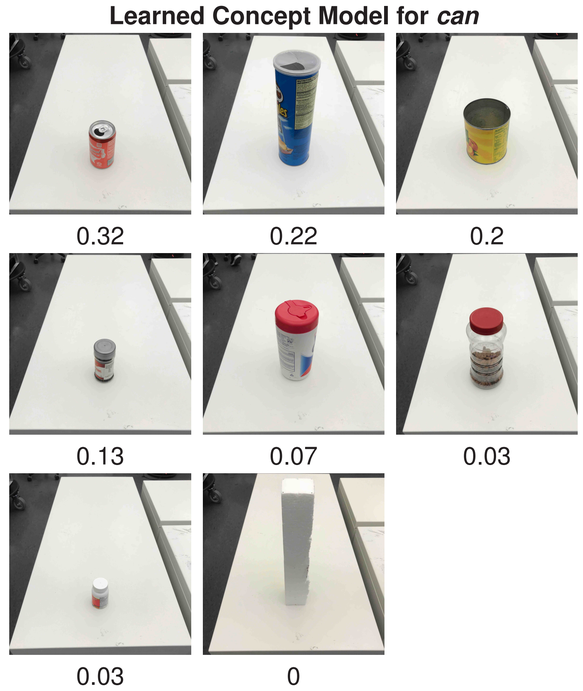
For both *navigation* and *relocation*, there is a slight decrease in the number of questions between the *Initial* agent and the *Trained (Parsing+Perception)* agent. The *Trained (Perception)* agent sees slightly worse performance across tasks, possibly due to a larger lexicon of adjectives and nouns (e.g., *can* as a descriptive noun now polysemous with *can* as a verb—*can you...*) without corresponding parsing updates.

$\mathcal{A}$	Usability Survey (Likert 1-7) ↑ Navigation (p)	Delivery (p)	Relocation (p)
In	3.09 ± 2.04	3.20 ± 2.12	3.37 ± 2.17
Tr*	3.51 ± 2.05(.09)	3.60 ± 2.09(.12)	3.60 ± 2.08(.37)
Tr	$\mathbf{3.76} \pm 2.07$(.01)	$\mathbf{3.87} \pm 2.10$(.01)	$\mathbf{3.93} \pm 2.16$(.04)

The average Likert rating given on usability survey prompts for each task across the agents. **Bold** indicates an average **Tr***ained** (*Perception*) and **Tr***ained* (*Parsing+Perception*) model ratings significantly higher than the **In***itial* model ($p < 0.05$) under a Welch's t-test.

Across tasks, the *Trained (Parsing+Perception)* agent is rated as more usable than both the *Initial* agent and the *Trained (Perception)* agent that updated only its concept models from training conversations.

Learned Concept Model for *can*



Confidence distribution for the *can* concept model on the unseen test set objects after training. In total, the agent learned 25 new concept models, as well as synonym words for existing concepts, during training.

jdtho@cs.washington.edu