Workshop at Robotics: Science and Systems (RSS) 2021 Robotics for People: Perspectives on Interaction, Learning, and Safety

Dialogue Object Search Extended Abstract

Monica Roy*, Kaiyu Zheng*, Jason Liu, Stefanie Tellex

Department of Computer Science
Brown University

*Equal contribution

Motivation

We envision robots that can collaborate and communicate seamlessly with humans. It is necessary for such robots to decide both what to say and how to act, while interacting with humans. This involves combining task-oriented dialogue systems with decision making under uncertainty for embodied agents. We believe a task that captures the sequential nature of both the dialogue and physical decision making is necessary towards this goal.

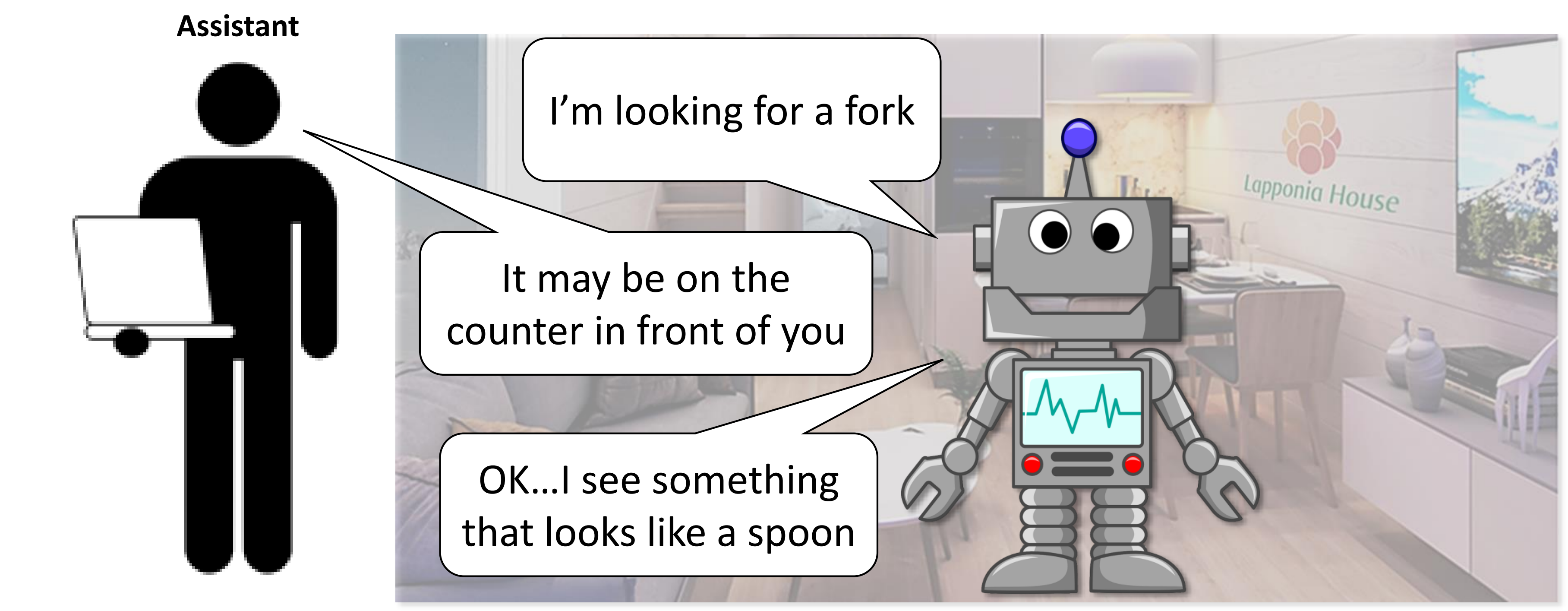
Motivation from Literature: Dialogue systems have traditionally involved users interacting with a *virtual agent* [1] for personal assistance (e.g. Siri), reservation booking, etc. Recent works that combine dialogue and dynamic, embodied decision making [2, 3] are limited to prediction tasks that bypass the challenges of evaluating a conversational embodied agent.

Dialogue Object Search: A New Task

Overall task: A robot is tasked to search for a target object in a human environment (e.g., kitchen) while engaging in an audio dialogue with a remote human assistant, who possesses inexact prior knowledge about the target object's location (e.g. 2D scatter plot).

Inputs: (1) a speech-based dialogue, (2) a mounted RGB-D camera, and shares its view with the human assistant. (3) sequences of RGB-D images of the scene, representing prior. Target objects are excluded from these images.

Outputs: The robot must decide what to say in the dialogue, and how to act (actions include moving (navigation) and opening/closing containers) in order to efficiently find the target while naturally interacting and collaborating with the human assistant.


Assistant
I'm looking for a fork
It may be on the counter in front of you
OK…I see something that looks like a spoon
Lapponia House
Controller

Note: figure is for illustration purpose only

References

[1] T.-H. Wen, D. Vandyke, N. Mrksiˇ c, M. Ga ´ siˇ c,´ L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young. (2017) "A network-based end-to-end trainable taskoriented dialogue system," in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Valencia, Spain: Association for Computational Linguistics

[2] Vries, H.D., Shuster, K., Batra, D., Parikh, D., Weston, J., & Kiela, D. (2018). "Talk the Walk: Navigating New York City through Grounded Dialogue." *ArXiv, abs/1807.03367*.

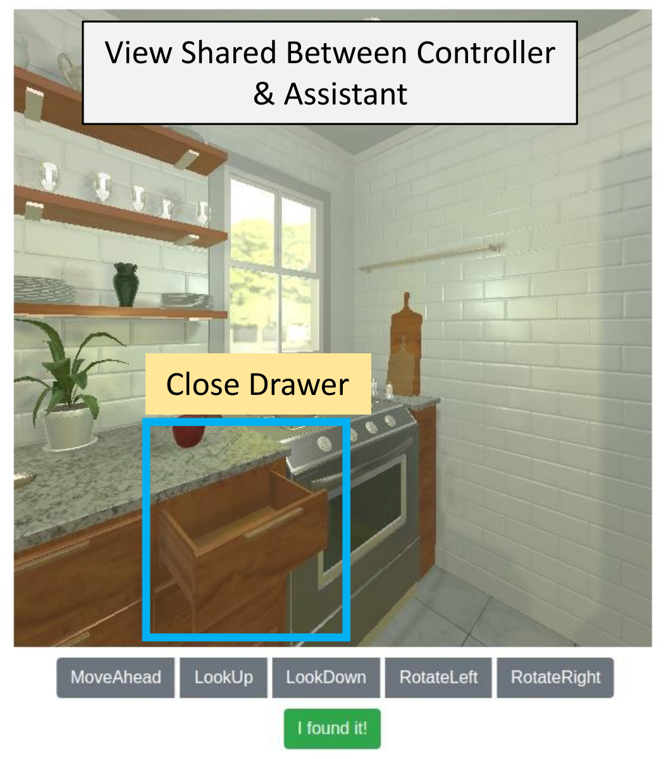
[3] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer. (2020) "Vision-and-dialog navigation," in Conference on Robot Learning. PMLR, pp. 394–406.

[4] Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., ... & Farhadi, A. (2017). Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Pilot Study

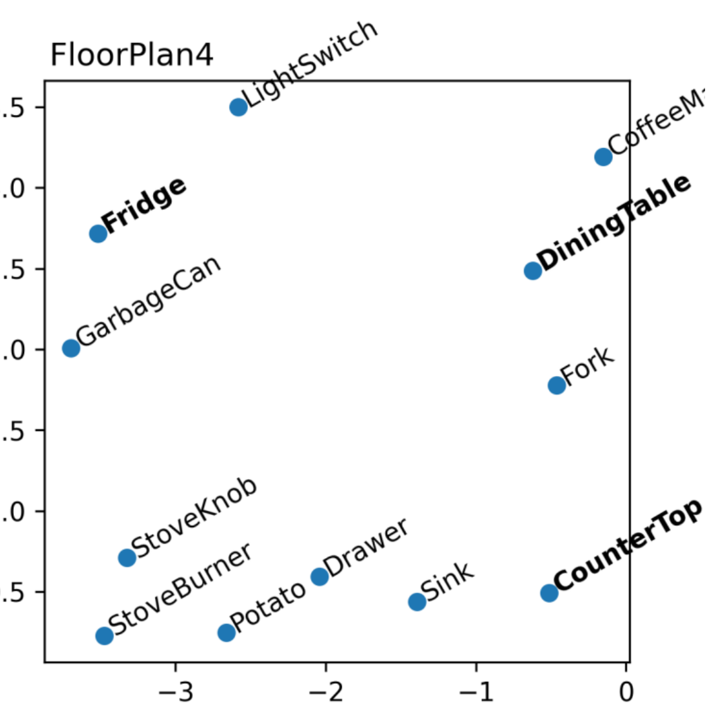
We designed and conducted a pilot study among three pairs of people (authors' lab members)
Objective: Understand how a human would behave if they are in the robot's position.
Study Design:

- We designate two roles according to the above problem setting. The ***Assistant*** is the person assisting in the process as the robot searches for a given target object. The ***Controller*** is the person who is taking on the role of the robot.
- We used AI2- THOR [4] as the simulated home environments.
- We used Zoom to record the audio and create transcripts of the dialogue.
- We implemented a web-based data collection tool where the Controller controls the agent in AI2-THOR through the web interface, and the Assistant has access to a 2D scatter plot of a subset of objects in the (see right)



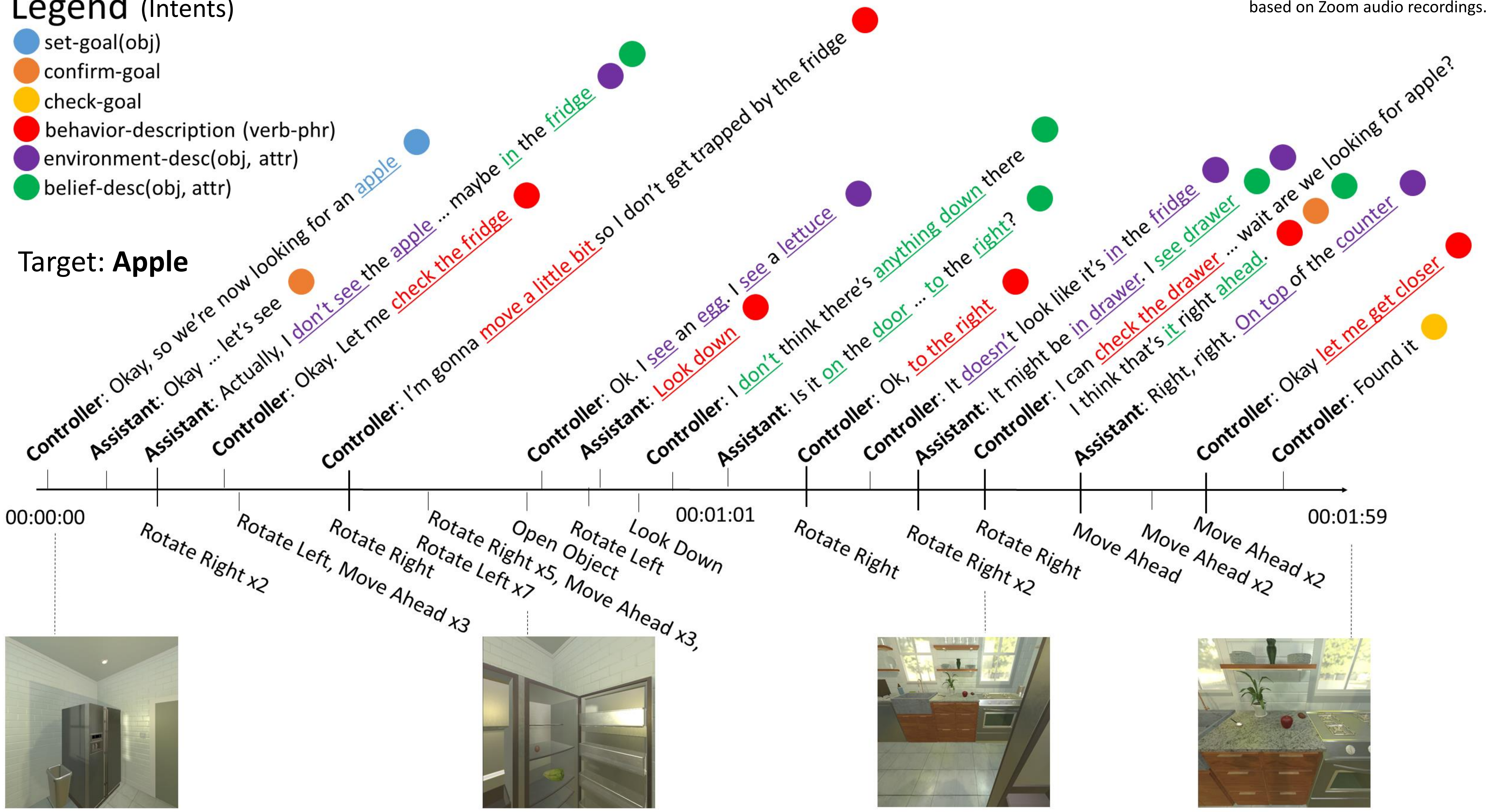
Controller interface

Scatter Plot Provided to the Assitant



Example Trial Collected from Pilot Study

Note: language in the dialogue example below is manually transcribed based on Zoom audio recordings.



Legend (Intents)

- ● set-goal(obj)
- ● confirm-goal
- ● check-goal
- ● behavior-description (verb-phr)
- ● environment-desc(obj, attr)
- ● belief-desc(obj, attr)

Target: **Apple**

Controller: Okay, so we're now looking for an apple
Assistant: Okay … let's see
Assistant: Actually, I don't see the apple … maybe in the fridge
Controller: Okay. Let me check the fridge
Controller: I'm gonna move a little bit so I don't get trapped by the fridge
Controller: Ok. I see an egg. I see a lettuce
Assistant: Look down
Controller: I don't think there's anything down there
Assistant: Is it on the door … to the right?
Controller: Ok, to the right
Controller: It doesn't look like it's in the fridge
Assistant: It might be in drawer. I see drawer
Controller: I can check the drawer … wait are we looking for apple?
I think that's it right ahead.
Assistant: Right, right. On top of the counter
Controller: Okay let me get closer
Controller: Found it

00:00:00
Rotate Right x2
Rotate Left, Move Ahead x3
Rotate Right
Rotate Left x7
Rotate Right x5, Move Ahead x3,
Open Object
Rotate Left
Look Down
00:01:01
Rotate Right
Rotate Right x2
Rotate Right
Move Ahead
Move Ahead x2
Move Ahead x2
00:01:59

Findings & Next Steps

We experimented with both speech-based dialogue and text-based dialogue. Our observations:
Speech: Participants typically engage in frequent back-and-forth
Text: the Controller must decide between controlling the agent in AI2-THOR versus typing in the chat, resulting in (1) search without interaction (2) hard to tell if assistant's input is considered

Common dialogue behaviors across trials:
(1) Specify and confirm target object; (2) Describe what is observed in the view
(3) Describe belief about target location (4) Describe intended or suggested behaviors
We codified these into preliminary **intents** (see above)

Next steps & Challenges:
(1) Scalability of data collection procedure (how to?)

- scalable and accurate transcription of the collected audio as well as intent labeling.
- We seek suggestions for strategies to collect such data at scale.

(2) Evaluation design: Should involve both experiment with simulated assistants and real human assistants