1 Abstract

Due to both the lack of perfect of knowledge from the robots' sides, and the impreciseness of language and communication from humans, we must provide novel frameworks for robots to overcome ambiguity. Given previous work by Whitney et al. [2017] towards asking simple questions to clarify intents to pick up objects with a POMDP framework, we describe a further extension to allow for clarifying dialogue with perceived objects' attributes. Specifically, we define a model using predicates over attributes and their values as potential questions alongside Whitney et al. [2017]'s object pointing question. In order to address the combinatorial explosion in the POMDP model introduced by an enlarged state space, we **hope to** define a greedy entropy minimizing heuristic to determine what question, and at what level of abstraction, is most helpful towards identifying the human operator's desired object [Note: we have not worked with this yet as of May 13,2018]. We will demonstrate this new **AQ+FETCH-POMDP**, or Attribute Question + FETCH-POMDP, model improves both the accuracy and speed of command understanding over previous works.

2 Intro

In just the past ten years, applications of artificial intelligence and robotics have spread to a ever-growing diverse set of industries: robots are homecleaning agents, medical assistants, self-driving cars, and more. However, in order to highlight the unique strengths of both people in their creativity and innovation and robots with their precision and ability to comprehensively process massive amounts of data, robots must be able to successfully collaborate with human peers.

While traditional voice assistants such as Siri, Google Home, and Alexa have found their way into the regular life of millions, such agents typically follow a one-way communication dialogue, with strictly the human operator prompting the autonomous agent to act. Furthermore, these often manually specificied dialogue systems often function following a one-directional tree of decisions, without room for backtracking or error correction. In order to improve the ability for robots to tackle more a complex set of tasks and ultimately act in collaboration, we look to provide a framework for the robot agent to prompt the operator as well, in order to intelligently clarify the desired task in the face of uncertainty. Furthermore, object picking tasks are found in almost all industries, from washing dishes, to picking olives or assembling a car. Starting with a voice-command specific input space and the limited action space of picking objects, we are able to provide a novel dialogue construction framework focusing on optimizing the clarification capabilities of a robotic agent like Rethink Robotic's Baxter.

The POMDP model establishes a natural framework for handling uncertainty, to both encourage question asking when uncertain, and also take action when the task is clear. Following previous work by Whitney et al. [2017], we examine FETCH-POMDP as a jumping-off point for using natural language to ask for objects and using clarifying questions as social feedback for object picking. One of the limitations of FETCH-POMDP, however, is its limitations to a single question for the robot, point, which points to an object to ask the operator which object is desired. Besides asking about an individual object, we want to allow for clarifying questions about an ambiguously referenced object with regards to its specific attributes. We are motivated by our

personal experience with using language to eliminate uncertainty for object picking tasks. When there are multiple spoons in a specific scene, when referring to a specific spoon, people are likely to communicate a specific spoon using its characteristics, such as "the blue spoon". With this in mind, our framework aims to include attribute-specifying questions to the robots' arsenal. In order to deal with this significantly larger state space, we **hope to** introduce a novel entropy heuristic that we minimize rather than fully-solving the POMDP. This will allow us to support both this newly introduced attribute-value question along with additional future question types while minimizing computation costs.

We **hope to eventually** evaluate this modified POMDP (not a true POMDP) through a real-world user study. When presenting users with a scenario to request a variety of objects through multiple arrangements, we compare the effectiveness of these additionally added questions along with our entropy heuristic to previous baselines established in FETCH-POMDP. With regards to both accuracy and speed, we hope to significantly improve users' ability to perform object picking tasks and successfully deal with ambiguity.

3 Related Works

Whitney et al. [2017] provides the most comprehensive and similar social feedback dialogue model for object picking. Based on this preliminary research, question asking alongside the POMDP formulation for dialogue system has huge potential towards navigating the ambiguity of language in human to robot interactions. This work serves as the starting point for expanding the POMDP formulation towards navigating ambiguity with a much more powerful set of questions.

Whitney et al. [2017] presents a POMDP model that models state as D, the desired object, and LR, last referred to object. This matches its action space, which are simply wait, point(x), and pick(x), where x is a given object. Furthermore this model observes both speech and human gesture (a person pointing). The FETCH-POMDP model provided rewards designed for a 6 object scenario such that the model would would consider the actions with wait being the least costly, pointing being more costly, picking the wrong object extremely costly, and picking the correct object as most rewarding. This model was able to demonstrate, that as opposed to a model not having a pointing action/social feedback capabilities, in two different 6 object circumstances, at least as good performance in the speed and accuracy of this model. The first circumstance was with 6 objects where human pointing would clearly refer to a specific object, referred to as the nonambiguous situation, and the second was where human pointing was unclear, referred to as the ambiguous situation.

Whitney et al. [2017] outlined that gesture in combination with a simple bag-of-words language model was generally sufficient for referencing objects. With this in mind, we decided to occlude gesture, in order to allow for more ambiguous references during a given interaction, and specifically highlight social feedback via attribute question asking.

Deits et al. [2013] provides a base reference for creating an entropy based heuristic for measuring uncertainty and determining the appropriate action a robotic agent should take.

4 POMDP Definition

We define a POMDP by $\langle S, A, T, R, \Omega, O, \gamma \rangle$.

4.1 State

Our states are $(D, Q_t, Q_v) \in S$, where D is the desired object, Q_t is the question type, and Q_v is the question value.

$Q_t \in \{\texttt{none}, \texttt{point}, \texttt{attr}\}$. If no question has been asked yet, then $Q_t = \texttt{none}$. If a question has been asked, then Q_t is the type of the last question asked, where `point` is the same as Whitney et al. [2017] and `attr` is our new type of question.

If $Q_t = \texttt{none}$, then we assign $Q_v = \texttt{none}$. If $Q_t = \texttt{point}$, then Q_v is the object that was asked about. If $Q_t = \texttt{attr}$, then Q_v is a predicate that represents the attribute-value pair that Baxter asked about. For example, if Baxter asked if the color of the desired object was blue, then $Q_v = \lambda x\ \texttt{color-blue}(x)$.

4.2 Actions

The actions are $A = \{\texttt{wait}\} \cup \{\texttt{point}_i\}_{i \in I} \cup \{\texttt{attr}_{k,v}\}_{k \in K, v \in V_k} \cup \{\texttt{pick}_i\}_{i \in I}$. I is the set of objects. K is the set of attributes. V_k is the set of possible values for attribute k.

$\texttt{attr}_{k,v}$ means that Baxter asks: "Does attribute k have value v?"

4.3 Transitions

D never changes. If a question is asked, Q_t and Q_v will change deterministically to reflect that.

Note It might make sense to have a small chance that D changes. This means that in the absence of input, the belief will converge back to uniform, which is pretty natural. (See earlier ICRA paper with Miles Eldon.)

4.4 Reward

As in the original paper, we assign constant cost to all actions based on the time each action takes.

4.5 Observations

Our observations are $(l_b, l_r) \in \Omega$. l_r contains all affirmative/negative words, and l_b contains all words that appear in $V = \cup_{k \in K} V_k$, the set of all values for all attributes.

Note: Ask why truncating l_b is not mentioned in Whitney et al. [2017].

We define our observation function by $O(o \mid s)$, which is the probability of observing $o \in \Omega$ given state $s \in S$. We assume that the observation is conditionally independent of the last action taken given s.

$$O(o \mid s) = \mathbb{P}(l_b, l_r \mid D, Q_t, Q_v) \tag{1}$$

Note: Maybe add reasoning for our conditional independence assumptions.

We assume that l_b and l_r are conditionally independent given the state.

$$O(o \mid s) = \mathbb{P}(l_b \mid D, Q_t, Q_v)\mathbb{P}(l_r \mid D, Q_t, Q_v) \tag{2}$$

4.5.1 Base Utterance

We assume that l_b is conditionally independent from Q_t and Q_v given D.

$$\mathbb{P}(l_b \mid D, Q_t, Q_v) = \mathbb{P}(l_b \mid D) \tag{3}$$

We then define the distribution in the same way as the original paper. p_l is a pre-defined parameter representing the probability of observing an utterance. Recall that l_b contains only the words that are in our vocabulary, V.

$$\mathbb{P}(l_b \mid D) = \begin{cases} p_l \prod_{w \in l_b} P(w \mid D) & l_b \text{ is not empty} \\ 1 - p_l & l_b \text{ is empty} \end{cases} \tag{4}$$

We then define the probability of seeing individual words. Here, V_D is the set of the desired object's values for each attribute, $V = \cup_{k \in K} V_k$ is the set of all values for all attributes, and α is the smoothing parameter.

$$\mathbb{P}(w \mid D) = \frac{\mathbb{1}_{V_D}(w) + \alpha}{|V_D| + \alpha|V|} \tag{5}$$

Note We decided to keep this the same as Whitney et al. [2017]. However, we don't think it makes sense for the base utterance to influence things after Baxter has asked a question, given that all our our questions are yes/no questions (so we should only pay attention to the response utterance). If we follow this train of thought, then we would instead define l_b as uniform for the cases where $Q_t \in \{\texttt{point}, \texttt{attr}\}$. However, we have to consider the case where the user says "no, the red one". We would want to account for "red" in the update. But what about the case where the user says "no, not the blue one"? In this case, we wouldn't want to account for "blue".

4.5.2 Potential New Model for Base Utterance

Instead of $P(w \mid D)$, we could define a join distribution over the words and attributes:

$$\mathbb{P}(w, k \mid D) = \mathbb{P}(w \mid k, D)\mathbb{P}(k \mid D) \tag{6}$$

So instead of $p_l \prod_{w \in l_b} P(w \mid D)$ we would use

$$p_l \prod_{w \in l_b} \mathbb{P}(w \mid k, D)\mathbb{P}(k \mid D) \tag{7}$$

where each $w \in l_b$ is in V_k for exactly one attribute $k \in K$. More formally, we would write this as:

$$p_l \prod_{w \in l_b} \sum_{k \in K} \mathbb{1}_{w \in V_k} \mathbb{P}(w \mid k, D)\mathbb{P}(k \mid D) \tag{8}$$

We define $\mathbb{P}(w \mid k, D)$ similarly to $\mathbb{P}(w \mid D)$ above, but with an attribute-specific distribution – $\mathbb{P}(w \mid k, D)$ is defined over V_k. Here, $D.k \in V_k$ is the value the desired object has for attribute k.

$$\mathbb{P}(w \mid k, D) = \frac{\mathbb{1}_{w=D.k} + \alpha}{1 + \alpha|V_k|} \tag{9}$$

We have a few ideas of how to define the distribution over attributes, $\mathbb{P}(k \mid D)$:

- Uniform over K. We suspect that under this definition, this new model would work exactly the same as the old model.
- Make it uniform if $Q_t \in \{\texttt{none}, \texttt{point}\}$, but if $Q_t = \texttt{attr}$ we assign a higher probability to the attribute that was just asked about. Note that while this would probably be a more accurate model of the user, we would no longer be able to assume that l_b is conditionally independent of Q_t given D.
- $\mathbb{P}(k \mid D) = \frac{|V_k|}{|V|}$, potentially with smoothing. This would reflect the fact that the user could be more specific by saying the value of the attribute with the most potential values, because this provides the most information (reduces entropy the most).

4.5.3 Response Utterance

We use a simplified model for l_r: we say it has to be positive, negative, or neither. If we observe an affirmative word, then $l_r = \texttt{yes}$. If we observe a negative word, then $l_r = \texttt{no}$. If we observe neither or both, then $l_r = \texttt{none}$.

Note: We will specifically define which words are affirmative and negative later. (We will start by using the same sets of words defined in Whitney et al. [2017])

When $Q_t = \texttt{none}$, we define $\mathbb{P}(l_r \mid D, Q_t, Q_v)$ to be uniform over $\{\texttt{yes}, \texttt{no}, \texttt{none}\}$.

For the following cases, we define a parameter ϵ that represents the small probability that the person doesn't answer a question in the way we assume they would. We also use p_l from Whitney et al. [2017] (the probability of observing an utterance).

For the case when $Q_t = \texttt{point}$, we define two different distributions for $\mathbb{P}(l_r \mid D, Q_t, Q_v)$ for the

two cases when $Q_v = D$ and $Q_v \neq D$:

$$P(l_r \mid D, Q_t, Q_v) = \begin{cases} \begin{cases} p_l(1-\epsilon) & Q_v = D \\ p_l\epsilon & Q_v \neq D \end{cases} & l_r = \texttt{yes} \\ \begin{cases} p_l\epsilon & Q_v = D \\ p_l(1-\epsilon) & Q_v \neq D \end{cases} & l_r = \texttt{no} \\ 1 - p_l & l_r = \texttt{none} \end{cases} \tag{10}$$

Now we consider the case where $Q_t = \texttt{attr}$ and $Q_v = \lambda x\ \texttt{attr-val}(x)$ is the predicate that represents the attribute and value Baxter asked about. We define two different distributions for $\mathbb{P}(l_r \mid D, Q_t, Q_v)$ for the two cases when $\texttt{attr-val}(D)$ is true and when it's false:

$$P(l_r \mid D, Q_t, Q_v) = \begin{cases} \begin{cases} p_l(1-\epsilon) & \texttt{attr-val}(D) \\ p_l\epsilon & \neg\texttt{attr-val}(D) \end{cases} & l_r = \texttt{yes} \\ \begin{cases} p_l\epsilon & \texttt{attr-val}(D) \\ p_l(1-\epsilon) & \neg\texttt{attr-val}(D) \end{cases} & l_r = \texttt{no} \\ 1 - p_l & l_r = \texttt{none} \end{cases} \tag{11}$$

5 Belief Update Tests

We want to test our model in a variety of scenarios and see how the belief distribution is updated. Here are some ways ways to vary our tests:

- Vary number of attributes
- Vary the question type
- Vary the utterance (observation)

For all tests, we assume the initial belief distribution is uniform. We use $\epsilon = 0.01$, $p_l = 0.95$, and $\alpha = 0.2$.

5.1 Two objects and one attribute

We have two objects, one red and one blue. Color is the only attribute.

$I = \{\texttt{ML}_R, \texttt{ML}_B\}$ $K = \{\texttt{color}\}$ $V_{\texttt{color}} = \{\texttt{red}, \texttt{blue}\}$ $s_0 = (D, \texttt{none}, \texttt{none})$

$z_0 =$"blue" $l_b =$["blue"] $l_r = \texttt{none}$

Result $b(\texttt{ML}_B) = 0.857$ $b(\texttt{ML}_R) = 0.143$

5.2 Two objects and two attributes

The two attributes are color and orientation.

$K = \{\texttt{color}, \texttt{orientation}\}$ $V_{\texttt{color}} = \{\texttt{red}, \texttt{blue}\}$ $V_{\texttt{orientation}} = \{\texttt{x-aligned}, \texttt{y-aligned}\}$

$s_0 = (D, \texttt{none}, \texttt{none})$

5.2.1 "blue" with no overlap

We have two objects, one blue x-aligned and one red y-aligned.

$I = \{\texttt{ML}_{BX}, \texttt{ML}_{RY}\}$ z_0 = "blue" l_b =["blue"] $l_r = \texttt{none}$

Result $b(\texttt{ML}_{BX}) = 0.857$ $b(\texttt{ML}_{RY}) = 0.143$

5.2.2 "blue" with overlapping characteristic

We have two objects, one blue x-aligned and one red x-aligned.

$I = \{\texttt{ML}_{BX}, \texttt{ML}_{RX}\}$ z_0 = "blue" l_b =["blue"] $l_r = \texttt{none}$

Result $b(\texttt{ML}_{BX}) = 0.857$ $b(\texttt{ML}_{RX}) = 0.143$

5.2.3 "blue x-aligned" with no overlap

We have two objects, one blue x-aligned and one red y-aligned.

$I = \{\texttt{ML}_{BX}, \texttt{ML}_{RY}\}$ z_0 = "blue x-aligned" l_b =["blue", "x-aligned"] $l_r = \texttt{none}$

Result $b(\texttt{ML}_{BX}) = 0.973$ $b(\texttt{ML}_{RY}) = 0.027$

5.2.4 "blue x-aligned" with overlapping characteristic

We have two objects, one blue x-aligned and one red x-aligned.

$I = \{\texttt{ML}_{BX}, \texttt{ML}_{RX}\}$ z_0 = "blue x-aligned" l_b =["blue", "x-aligned"] $l_r = \texttt{none}$

Result $b(\texttt{ML}_{BX}) = 0.857$ $b(\texttt{ML}_{RX}) = 0.143$

5.3 Point question

We have two objects, one red and one blue. Baxter has just pointed at $\texttt{ML}_B$.

$I = \{\texttt{ML}_R, \texttt{ML}_B\}$ $K = \{\texttt{color}\}$ $V_{\texttt{color}} = \{\texttt{red}, \texttt{blue}\}$ $s_0 = (D, \texttt{point}, \texttt{ML}_B)$

5.3.1 "yes"

z_0 = "yes" l_b = [] $l_r = \texttt{yes}$

Result $b(\texttt{ML}_B) = 0.99$ $b(\texttt{ML}_R) = 0.01$

5.3.2 "yes the blue one"

z_0 = "yes the blue one" l_b = ["blue"] $l_r = \texttt{yes}$

Result $b(\texttt{ML}_B) = 0.998$ $b(\texttt{ML}_R) = 0.002$

5.3.3 "no"

z_0 = "no" l_b = [] $l_r = \texttt{no}$

Result $b(\texttt{ML}_B) = 0.01$ $b(\texttt{ML}_R) = 0.99$

5.3.4 "no not the blue one"

z_0 = "no not the blue one" l_b = ["blue"] $l_r = \texttt{no}$

Result $b(\texttt{ML}_B) = 0.057$ $b(\texttt{ML}_R) = 0.943$

5.3.5 "no the red one"

z_0 = "no the red one" l_b = ["red"] $l_r = \texttt{no}$

Result $b(\texttt{ML}_B) = 0.002$ $b(\texttt{ML}_R) = 0.998$

5.4 Attribute question

If we use only two objects, this case is the same as above. So we will only test it on the case where there are three objects, and two of them have the same color.

$I = \{\texttt{ML}_{R1}, \texttt{ML}_{R2}, \texttt{ML}_B\}$ $K = \{\texttt{color}\}$ $V_{\texttt{color}} = \{\texttt{red}, \texttt{blue}\}$ $s_0 = (D, \texttt{attr}, \texttt{color-blue(x)})$

5.4.1 "yes"

z_0 ="yes" l_b = [] $l_r = \texttt{yes}$

Result $b(\texttt{ML}_{R1}) = 0.010 \quad b(\texttt{ML}_{R2}) = 0.010 \quad b(\texttt{ML}_B) = 0.980$

5.4.2 "yes the blue one"

$z_0 =$"yes the blue one" $\quad l_b =$ ["blue"] $\quad l_r = \texttt{yes}$

Result $b(\texttt{ML}_{R1}) = 0.002 \quad b(\texttt{ML}_{R2}) = 0.002 \quad b(\texttt{ML}_B) = 0.997$

5.4.3 "no"

$z_0 =$"no". $\quad l_b =$ [] $\quad l_r = \texttt{no}$

Result $b(\texttt{ML}_{R1}) = 0.497 \quad b(\texttt{ML}_{R2}) = 0.497 \quad b(\texttt{ML}_B) = 0.005$

5.4.4 "no not the blue one"

$z_0 =$"no not the blue one". $\quad l_b =$ ["blue"] $\quad l_r = \texttt{no}$

Result $b(\texttt{ML}_{R1}) = 0.485 \quad b(\texttt{ML}_{R2}) = 0.485 \quad b(\texttt{ML}_B) = 0.029$

5.4.5 "no the red one"

$z_0 =$"no the red one". $\quad l_b =$ ["red"] $\quad l_r = \texttt{no}$

Result $b(\texttt{ML}_{R1}) = 0.500 \quad b(\texttt{ML}_{R2}) = 0.500 \quad b(\texttt{ML}_B) = 0.001$

6 Evaluation

In evaluating the performance of our model and various existing models, we specifically look at the ability for each model to from an initial confused state, where the robotic agent initiates the interaction with clarifying action, to hand off the desired object to the human user agent.

Given that robotic agents will almost inevitably encounter situations in which the user agent provided information is incomplete, not well defined, or otherwise not understood perfectly by the robotic agent, we believe that measuring performance of models given an initially confused state will highlight the differences in each's ability to provide meaningful social feedback. This way, we specifically highlight and examine social feedback capabilities.

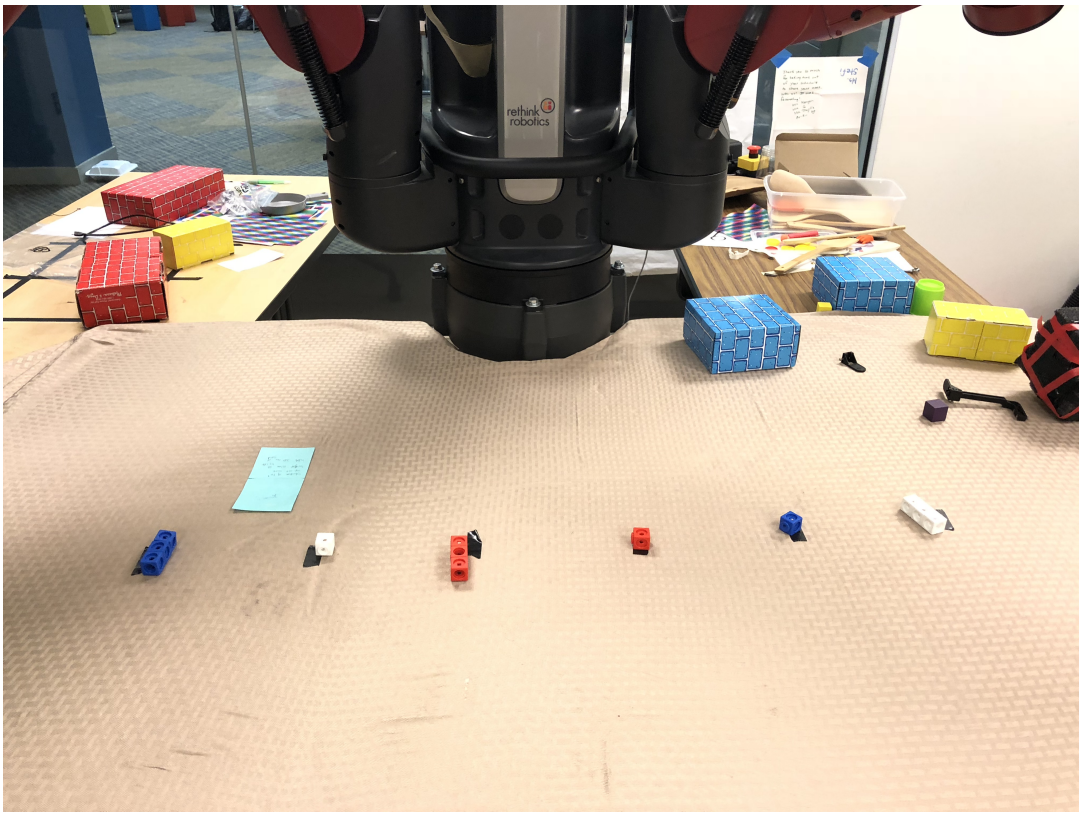
With this in mind, we will define the start of an interaction as the time when Baxter begins its first action and the end of the interaction as the time when Baxter finishes its handoff during

Table 1: Results from different models

	AQ	FETCH
User 1	63	29
User 2	57	33
User 3	62	91

its pick action. Furthermore, we define a correct pick as a handoff of an item that is the user's desired object, and all other picks as an incorrect pick.

We evaluated Baxter's performance with our proposed AQ+FETCH-POMDP and the Whitney et al. [2017]'s FETCH-POMDP with gesture occluded on one scenario:



From left to right: 1 Blue Rod, 1 White Block, 1 Red Rod, 1 Red Block, 1 Blue Block, 1 White Rod.

Here are the results from preliminary testing with 3 users are in Table 1 for speed.

Furthermore, we see that in both models the desired object was always correctly identified.

From these preliminary results, we can see that there is no guaranteed improvement from having attribute question asking in this given interaction. Likely due to the small sample size, we not yet able to see conclusive evidence of what the differences between the two models are.

We can clearly note though, given a user that answers correctly to all prompts, the upper bounds for questions needed in the AQ-FETCH model is an improvement over that of FETCH in this scenario. Under the given circumstances, AQ-FETCH would need to ask at most 3 questions to correctly identify a desired object, while FETCH would need to ask at most 5. With further user studies with more individuals, we hope to see this difference highlighted more clearly.

7 Conclusion

In this paper we present a new model for reducing confusion by asking clarifying questions during object fetching. Just like in the original FETCH-POMDP paper, the robot has the ability to point to an object to ask if that's the object the user wants. Now, it also has the ability to ask questions about the attributes of the desired object. We find that the POMDP is still solvable with this expanded state-action space.

There are many ways to build on this model. We could give the robot the ability to ask what the value is of a certain attribute. We could also integrate features to allow the robot to understand relative position (so the user can refer to objects that are to the left/right of other objects). There are many ways to improve the language model, since the one we use here is simple.

This model can be applied to a variety of other decision making tasks. Any home robot will need to perform complex tasks in order to be useful, and even with the current state of the art in natural language processing and world modeling, the robot can easily become confused. Even if there's no combination of words that a human can say that will convey the task that they want, robots can still use their knowledge to clarify meaning. This will lead to more successful interactions

8 References

David Whitney, Eric Rosen, James MacGlashan, Lawson L.S. Wong, and Stefanie Tellex. Reducing errors in object-fetching interactions through social feedback. *International Conference on Robotics and Automation*, 2017.

Robin Deits, Stefanie Tellex, Pratiksha Thaker, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction*, 2013.