Language-guided Semantic Mapping and Mobile Manipulation in Partially Observable Environments

Siddharth Patki
University of Rochester
spatki@ur.rochester.edu

Ethan Fahnestock
University of Rochester
efahnest@u.rochester.edu

Thomas M. Howard
University of Rochester
thomas.howard@rochester.edu

Matthew R. Walter
Toyota Technological Institute at Chicago
mwalter@ttic.edu

Abstract: Recent advances in data-driven models for grounded language understanding have enabled robots to interpret increasingly complex instructions. Two fundamental limitations of these methods are that most require a full model of the environment to be known a priori, and they attempt to reason over a world representation that is flat and unnecessarily detailed, which limits scalability. Recent semantic mapping methods address partial observability by exploiting language as a sensor to infer a distribution over topological, metric and semantic properties of the environment. However, maintaining a distribution over highly detailed maps that can support grounding of diverse instructions is computationally expensive and hinders real-time human-robot collaboration. We propose a novel framework that learns to adapt perception according to the task in order to maintain compact distributions over semantic maps. Experiments with a mobile manipulator demonstrate more efficient instruction following in a priori unknown environments.

Keywords: Natural Language, Perception, Human-Robot Interaction

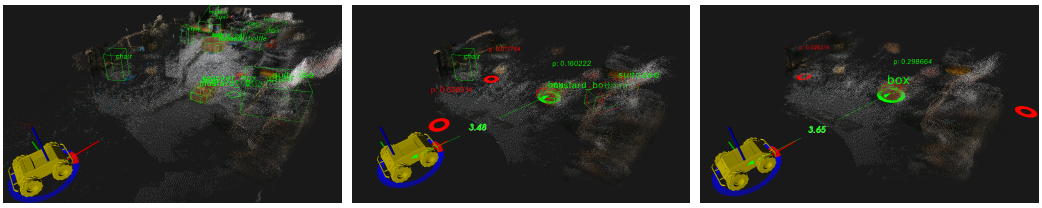
1 Introduction

Realizing robots that can work effectively alongside people in cluttered, unstructured environments (Fig. 1) requires command and control mechanisms that are both intuitive and efficient. Natural language provides a flexible medium through which users can communicate with robots without the need for specialized interfaces or significant training. For example, a voice-controllable wheelchair [1] permits people with limited mobility to independently navigate their environments without using sip-and-puff arrays or head-actuated switches.

Figure 1: A user commands a robot to "retrieve the ball inside the box" in an a priori unknown environment.

Significant progress in data-driven approaches to language understanding have enabled robots to both interpret and generate complex free-form utterances in a variety of domains [2, 3, 4, 5, 6, 7]. Symbol grounding-based methods formulate language understanding as a problem of associating linguistic phrases with their corresponding referents in the robot's model of its state and action space. This places two fundamental limitations on grounding-based approaches to language understanding. First, most contemporary solutions require a priori knowledge of the robot's environment in the form of a "world model" that expresses the metric and semantic properties of every object and location in the robot's environment. This model is typically created by augmenting a SLAM-generated metric map with manually and/or automatically inferred semantic information. Critically, this prevents the robot from interpreting commands in unobserved or partially observed environments. Second, advances in sensor technology and computer vision algorithms have give rise to a wealth of information that can be infused

3rd Conference on Robot Learning (CoRL 2019), Osaka, Japan.



(a) exhaustive world map (b) distribution of exhaustive maps (c) distribution of compact maps

Figure 2: Our framework learns to exploit environment and task-related information implicit in a given utterance to infer a distribution over compact task-relevant maps in a priori unknown environments. Consider the command "retrieve the ball inside the box". Traditional approaches to language grounding involve reasoning over (a) a highly detailed model of the environment that is computationally expensive to maintain and assumed to be known a priori. To enable grounding in unknown or partially observed environments, recent methods consider maintaining a distribution over (b) highly detailed maps that include all observed objects as well as the hypothesized location of unknown objects referenced in the utterance. In contrast, our proposed approach learns to reason over (c) a distribution of compact maps that model only task-relevant objects by adapting perception based on the utterance. In the above figures, circles denote the hypothesized locations for a box that contains a ball from different maps in the distribution.

into these world models. This results in exhaustively detailed representations of the environment (Fig. 2a), such as those that could model all of the spoons on a table, or door handles on the doors and their affordances. While these models are sufficient for language grounding, the computational cost of perceiving all objects and performing symbol grounding in their context [8] precludes real-time language grounding. Conversely, a poorly detailed model of the environment that assumes a coarse, static representation of objects limits the diversity of the instructions that can be grounded.

Towards addressing the problem of interpreting instructions in a priori unknown environments, recent work by Hemachandra et al. [9] presents an approach that exploits information communicated by the human that may indirectly inform a robot about its target environment. Specifically, their approach extracts spatial-semantic properties of objects and regions conveyed as part of an instruction in order to infer a distribution over possible maps, permitting language understanding in novel environments. However, their method maintains a distribution over unnecessarily detailed semantic maps that are generated with the help of fiducials [10]. In practice, building such highly detailed world models without using fiducials is computationally expensive, and hinders smooth human-robot collaboration. A recent line of work [11, 12] proposes a model that learns to dynamically adapt the configuration of the robot's perception pipeline by inferring the classifiers needed to express the symbols that would later be needed by the symbol grounding model. In this work, we present an efficient approach to collaborative mobile manipulation that jointly learns the models for map inference and adaptive perception. The proposed framework infers the subset of perceptual classifiers needed to efficiently update environment models in an online fashion for the execution of multiple tasks in novel or partially observed worlds. Experimental results on a Clearpath Robotics Husky A200 Unmanned Ground Vehicle outfitted with a Universal Robotics UR5 manipulator and Robotiq gripper demonstrate faster task execution in partially observed worlds compared to a fixed-perception baseline for mobile manipulation tasks.

2 Related Work

Statistical approaches to language understanding have enabled robots to follow complex free-form instructions involving object manipulation [8, 2, 3], navigation [4, 5, 6, 7] and mobile manipulation [13, 14]. A common approach to language understanding is to treat it as a symbol grounding problem [13, 8], whereby one learns a model that associates (i.e., "grounds") each word in an utterance to its corresponding referent in the robot's model of its state and action space. Such approaches typically require a "world model" to be known a priori in the form of a map that expresses the location, geometry, semantic type, and colloquial name of all objects and regions in the environment. In practice, these maps are often generated by first using a state-of-the-art SLAM algorithm [15, 16, 17], which produces flat representations that only model spatial information. Semantic

and topological properties are then manually added to realize a representation sufficient for language understanding. Notable exceptions include the work of Duvallet et al. [18], which learns to follow navigational instructions in unknown environments based upon human demonstrations, as well as recent work on language-based visual navigation in novel environments [19, 20]. The latter differ from our work in that they map language directly to actions, and do not (explicitly) infer a compact world model from language. Meanwhile, statistical parsing-based methods [5, 21, 6, 7] associate natural language utterances to a meaning representation that typically takes the form of a lambda calculus. Such an approach avoids the need for an explicit world model, typically at the expense of requiring a down-stream controller capable of executing inferred plans in unknown environments.

Also relevant is recent work that focuses on grounding unknown or ambiguous utterances. One approach to dealing with ambiguous utterances is to utilize inverse grounding [22, 23] to generate targeted questions for the user that are deemed to be most informative, e.g., in terms of the reduction in entropy for the grounding distribution [24]. Meanwhile, several methods learn a priori unknown grounding models by exploring the relationship between novel linguistic predicates and the robot's world model and/or its percepts [2, 25, 26, 27]. Our work differs in that we assume that the concepts are known, but that the instantiations of these concepts in the robot's environment are unknown.

Similar to how our framework performs map inference, state-of-the-art semantic mapping frameworks build rich representations of the world from the robot's multimodal sensor streams [28, 29], including linguistic descriptions [30, 31]. The latter methods attempt to reason over all perceptual cues irrespective of the utterance. In contrast, our framework uses natural language as another sensor to maintain a distribution over the metric, topological, and semantic properties of the unknown environment. This distribution is then used for language grounding and planning.

Most language grounding methods perform inference over the entire power set of objects, regions, actions, and other constituents in the search space. The Distributed Correspondence Graph (DCG) [8] reduces the complexity of grounding from exponential to linear by performing inference separately across conditionally independent constituents in a graphical model of language grounding. Recent variations of the DCG [8] further improve computational efficiency by performing inference in a multi-stage, coarse-to-fine manner. We leverage DCG in this work to learn the proposed models for adapting perception, map inference, and symbol grounding.

3 Technical Approach

Many contemporary approaches frame natural language understanding as inference over a learned distribution that associates linguistic elements to their corresponding referents in a symbolic representation of the robot's state and action space. The space of symbols $\mathbf{\Gamma} = \{\gamma_1, \gamma_2 ... \gamma_n\}$ includes concepts derived from the robot's environment model, such as objects and locations, and includes the viable robot behaviors, such as navigating to a desired location or manipulating a specific object. The distribution over symbols is conditioned on the parse of the utterance $\mathbf{\Lambda} = \{\lambda_1, \lambda_2 ... \lambda_n\}$, and a model of the world Υ that expresses environment knowledge extracted from sensor measurements $\boldsymbol{z}_{1:t}$ using a set of perceptual classifiers $\boldsymbol{P} = \{p_1, p_2 ... p_n\}$. Natural language understanding framed as a symbol grounding problem then follows as maximum a posteriori inference over the power set of referent symbols $\mathcal{P}(\mathbf{\Gamma})$.

$$\mathbf{\Gamma}^* = \underset{\mathcal{P}(\mathbf{\Gamma})}{\arg\max} \; p(\mathbf{\Gamma} | \mathbf{\Lambda}, \Upsilon) \tag{1}$$

This approach reasons in the context of a known model of the world Υ that is assumed to express all information necessary to ground the given utterance. This precludes language understanding in unobserved (i.e., novel) or partially observed environments for which the world model is incomplete, thereby making accurate inference (1) infeasible. To address this problem, we instead treat symbol grounding as inference conditioned on a latent model of the robot's environment $\overline{\Upsilon}$. Specifically, we learn a model that exploits environmental information implicit in an utterance to build a distribution over the topological, metric, and semantic properties of the environment

$$p(\overline{\Upsilon}_t | \mathbf{\Lambda}_{1:t}, \boldsymbol{z}_{1:t}, \boldsymbol{u}_{1:t}, \boldsymbol{P}), \tag{2}$$

where $\mathbf{\Lambda}_{1:t}$, $\boldsymbol{z}_{1:t}$, and $\boldsymbol{u}_{1:t}$ denote the history of utterances, sensor observations and odometry, respectively, and $\boldsymbol{P}$ is the set of classifiers in the robot's perception pipeline. This allows maintaining a world model distribution that not only embeds the perceived entities from sensor data but also

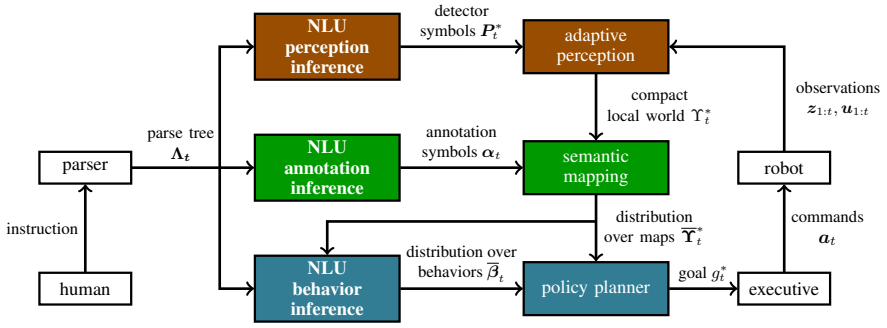


Figure 3: The system architecture for language understanding in unknown environments using adaptive perception, semantic mapping, and natural language symbol grounding. The three language understanding models learned from the corpus of annotated instructions are highlighted in bold.

models the unperceived information about the environment expressed in the utterance. This enables symbol grounding in unknown or partially observed environments. As we describe shortly, we maintain this distribution using a Rao-blackwellized particle filter, whereby each particle effectively denotes a hypothesized world model $\Upsilon_t^i \in \overline{\boldsymbol{\Upsilon}}_t$.

Treating the environment model as a latent random variable, we formulate symbol grounding as a problem of inferring a distribution over robot behaviors $\overline{\boldsymbol{\beta}}_t$. A behavior β_t is a representation of the intended robot actions expressed by the symbols in the inferred groundings $\boldsymbol{\Gamma}_t^*$. Each behavior $\beta_t^i \in \overline{\boldsymbol{\beta}}_t$ is inferred in the context of the corresponding world $\Upsilon_t^i \in \overline{\boldsymbol{\Upsilon}}_t$ and the instruction $\boldsymbol{\Lambda}_t$. The optimal trajectory $\boldsymbol{x}_t^*$ that the robot should take in the context of a distribution of behaviors then amounts to a planning under uncertainty problem formulated as inference

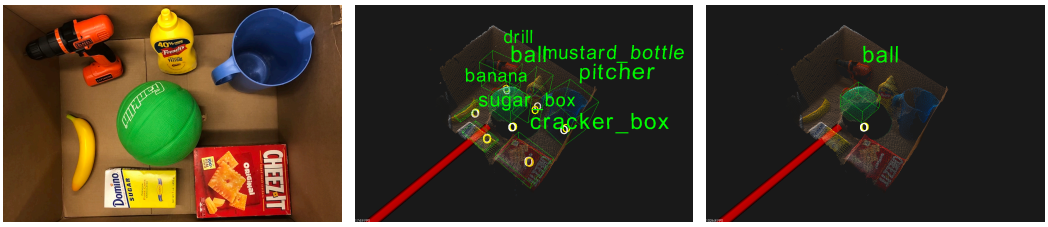
$$\boldsymbol{x}_t^* = \underset{\boldsymbol{x}_t \in \boldsymbol{X}_t}{\arg\max} \sum_{\Upsilon_t^i \in \overline{\boldsymbol{\Upsilon}}_t} \underbrace{p(\boldsymbol{x}_t | \beta_t^i, \Upsilon_t^i)}_{\text{path planning}} \times \underbrace{p(\beta_t^i | \boldsymbol{\Lambda}_t, \Upsilon_t^i)}_{\text{behavior inference}} \times \underbrace{p(\Upsilon_t^i | \boldsymbol{\Lambda}_{1:t}, \boldsymbol{z}_{1:t}, \boldsymbol{u}_{1:t}, \boldsymbol{P})}_{\text{semantic mapping}} \tag{3}$$

As the robot explores its environment, the distribution over world models is updated by incorporating new detections from the robot's perception pipeline and by incorporating information contained in any instructions that follow. Every time an update is made to the world distribution, the optimal trajectory is recomputed.

The ability to ground diverse natural language instructions is inherently linked to the richness of the robot's representation of the environment. However, building highly detailed models of unstructured environments and performing symbol grounding in their context is computationally expensive and places a runtime bottleneck on the model described by Equation 3. A recent line of research [11, 12] has shown that perception can be adapted by leveraging language to build task-specific, compact world models for efficient symbol grounding. Figure 3 shows the architecture of the proposed model that uses adaptive perception for the task of exploratory mobile manipulation in a priori unknown environments. We leverage adaptive perception online to build compact maps of the environment as the robot explores it. We hypothesize that using adaptive perception will improve the computational efficiency of semantic mapping (Eqn. 2) and behavior inference (Eqn. 1), and thus improve the runtime language understanding (Eqn. 3). In the following sections, we describe each of the individual learned models of the proposed architecture.

3.1 Adaptive Perception

In practice, a large fraction of the objects and the corresponding symbols are inconsequential to inferring the meaning of an utterance. In such cases, there exists a compact environment representation that is sufficient to interpret the utterance. A recent line of work [11, 12] proposes adapting the robot's perception pipeline according to the demands of the language utterance. The goal is to quickly provide minimal task-relevant world models that are sufficiently expressive to permit accurate and fast language grounding. Following [11] we learn a Distributed Correspondence Graph (DCG) [8] based probabilistic model that exploits natural language in order to infer a small, succinct subset of perceptual classifiers $\boldsymbol{P}_t^* = f(\boldsymbol{P}, \boldsymbol{\Lambda}_t)$ as conditioned on the utterance $\boldsymbol{\Lambda}_t$. This allows



(a) cluttered box (b) exaustive perception (c) adaptive perception

Figure 4: A comparison of world models generated by (b) exhaustive and (c) adaptive perception after first observing a cluttered box for the instruction "pick up the ball inside the box".

dynamic adaptation of the robot's perceptual capabilities according to the current task resulting in compact models of the world Υ_t^*.

Previously, adaptive perception has been shown to be effective in building compact models of the world from single observations [11] or a log of past observations [12]. The proposed architecture leverages adaptive perception online in the context of SLAM to build compact maps of the novel environment during exploration.

3.2 Semantic Mapping with Adaptive Perception

We model the robot's environment as a semantic graph [30] $\boldsymbol{\Upsilon}_t = \{\boldsymbol{G}_t, \boldsymbol{X}_t, \boldsymbol{L}_t\}$. The topology $\boldsymbol{G}_t$ is comprised of nodes n_i that represent distinct objects and locations and edges e_{ij} that express spatial relationships between pairs of nodes (e.g., as inferred from language and the robot's motion). The metric map $\boldsymbol{X}_t$ associates a pose x_i with each node n_i in the graph in similar fashion to pose graph SLAM [17]. The layer $\boldsymbol{L}_t$ expresses semantic attributes of each node (e.g., the type of each region/object and its colloquial name).

In order to follow instructions in a priori unknown environments, we leverage information about the environment implicit in a given utterance to maintain an informed distribution $\overline{\boldsymbol{\Upsilon}}_t$ over possible world models. We learn a DCG-based [8] probabilistic model that exploits natural language to infer a distribution over the available "annotations" $\boldsymbol{\alpha}_t$ (of which there may be none). These annotations include the type and relative location of different objects and regions in the environment. As an example, consider the instruction "get the drill from the box". DCG inference yields a distribution that assigns a high likelihood to annotations that suggest the existence of one or more objects of type "box" and "drill". High likelihood is associated with spatial relations that express a drill object as being "inside" a box object. The instruction can then be grounded in the context of a distribution of hypothesized worlds that incorporate the inferred annotations.

We maintain this distribution via a Rao-Blackwellized particle filter (RBPF) [32, 33], using a sample-based distribution over topologies, a Gaussian distribution over the metric map, and a Dirichlet over semantic information. We sample changes to the topology (as represented by a collection of particles) according to sensor-based observations and language-based annotations. We then update the resulting distribution over the metric map using an extended information filter. Maintaining a distribution over highly detailed world models and grounding instructions in their context is computationally non-trivial and places a bottleneck on the runtime efficiency of instruction following. We depart from previous work [9] by integrating adaptive perception to maintain a distribution over compact maps $\overline{\boldsymbol{\Upsilon}}_t^*$ that afford more efficient behavior inference

$$p(\overline{\boldsymbol{\Upsilon}}_t^* | \boldsymbol{\Lambda}_{1:t}, \boldsymbol{z}_{1:t}, \boldsymbol{u}_{1:t}, \boldsymbol{P}_t^*). \tag{4}$$

3.3 Behavior Inference with Adaptive Perception

We frame the problem of behavior inference as one of inferring a distribution over grounded behaviors $\overline{\boldsymbol{\beta}}_t$ for a given utterance in the context of the distribution over hypothesized worlds $\overline{\boldsymbol{\Upsilon}}_t$. When the space of symbols (groundings) $\boldsymbol{\Gamma}$ is large, the environment Υ is unstructured, and the free-form utterance $\boldsymbol{\Lambda}$ is complex and, making exact inference (1) becomes computationally intractable. Distributed Correspondence Graphs (DCG) [8] employ an approximate factorization of the grounding

distribution in Equation 1 that assumes conditional independence across the linguistic and symbolic constituents according to the hierarchical structure of language. DCG frames language understanding as an association problem by introducing the notion of correspondence variables $\phi_{ij} \in \mathbf{\Phi}$ that associate linguistic elements $\lambda_i \in \mathbf{\Lambda}$ (e.g., words and phrases) with symbols $\gamma_{ij} \in \mathbf{\Gamma}$. In practice, a large fraction of the object and region symbols are irrelevant to inferring the meaning of an utterance. In such cases, there exists a compact environment representation Υ^* that is sufficient to interpret the utterance. Reasoning over compact world models reduces the size of the search space, improving the complexity of inference. Following our earlier work [11, 12], we use adaptive perception to build these compact world representations. DCG inference then follows as a search for the correspondence variables $\mathbf{\Phi}^*$ that maximize the following factored distribution. Note that the grounding for a phrase depends on the child phrase groundings as contained in the true correspondence $\mathbf{\Phi}_{ci}$

$$\mathbf{\Phi}^* = \underset{\phi_{ij} \in \mathbf{\Phi}}{\arg\max} \prod_{i=1}^{|\mathbf{\Lambda}|} \prod_{j=1}^{|\mathbf{\Gamma}|} p(\phi_{ij} | \gamma_{ij}, \lambda_i, \mathbf{\Phi}_{ci}, \Upsilon^*). \tag{5}$$

In a priori unknown environments, our framework performs grounding inference (Eqn. 5) for each hypothesized world in the distribution. Thus the cost of behavior inference tends to be linear in the number of particles used to maintain the world distribution. Due to the computational advantage of performing inference over compact world models, our framework allows to reason over a larger distribution of possible worlds while being time efficient. Behavior inference finally yields a distribution over behaviors $\overline{\boldsymbol{\beta}}_t$. Each behavior $\beta_t^i \in \overline{\boldsymbol{\beta}}_t$ is parameterized by a type (navigate, retrieve or pickup) and a goal pose $g_t^i \in \boldsymbol{G}_t$.

3.4 Planning Under Uncertainty

We hypothesize that leveraging adaptive perception for semantic mapping and behavior inference will improve the runtime of the proposed system (Eqn. 6) compared to our baseline (Eqn. 3).

$$\boldsymbol{x}_t^* = \underset{\boldsymbol{x}_t \in \boldsymbol{X}_t}{\arg\max} \sum_{\Upsilon_t^{*i} \in \overline{\mathbf{\Upsilon}}_t^*} \underbrace{p(\boldsymbol{x}_t | \boldsymbol{\beta}_t^i, \Upsilon_t^{*i})}_{\text{path planning}} \times \underbrace{p(\boldsymbol{\beta}_t^i | \mathbf{\Lambda}_t, \Upsilon_t^{*i})}_{\substack{\text{behavior inference} \\ \text{with adaptive perception}}} \times \underbrace{p(\Upsilon_t^{*i} | \mathbf{\Lambda}_{1:t}, \boldsymbol{z}_{1:t}, \boldsymbol{u}_{1:t}, \boldsymbol{P}_t^*)}_{\substack{\text{semantic mapping} \\ \text{with adaptive perception}}} \tag{6}$$

Given a distribution over behaviors, identifying a suitable trajectory $\boldsymbol{x}_t^*$ amounts to a planning under uncertainty problem. We solve this problem with a policy that greedily chooses the best behavior β_t^* that maximizes the following optimization function

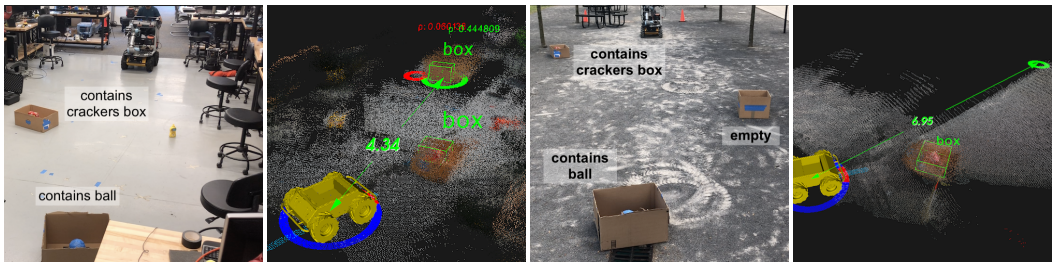
$$\beta_t^* = \underset{\beta_t^i \in \overline{\boldsymbol{\beta}}_t}{\arg\max} \underbrace{\psi(\beta_t^i)}_{\substack{\text{decaying gaussian} \\ \text{cost function}}} \times \underbrace{p(\beta_t^i | \mathbf{\Lambda}_t, \Upsilon_t^{*i})}_{\substack{\text{behavior inference} \\ \text{with adaptive perception}}} \times \underbrace{p(\Upsilon_t^{*i} | \mathbf{\Lambda}_{1:t}, \boldsymbol{z}_{1:t}, \boldsymbol{u}_{1:t}, \boldsymbol{P}_t^*)}_{\substack{\text{semantic mapping} \\ \text{with adaptive perception}}}, \tag{7}$$

where $\psi(\beta_t^i) = e^{-d(r_t, g_t^i)^2/10}$ is a decaying Gaussian value function with $d(r_t, g_t^i)$ being the Euclidean distance between the robot and the goal location g_t^i corresponding to the behavior β_t^i. This allows the robot to favor exploring goals that might be less likely but are closer to the robot. This process is repeated until the robot completes the instructed task.

4 Experimental Setup

To evaluate the scalability of this framework, we performed experiments on a Clearpath Husky A200 Unmanned Ground Vehicle outfitted with a Universal Robotics UR5 arm and Robotiq 3-finger Adaptive Robot Gripper (Fig. 2). An Intel RealSense D435 RGB-D camera was mounted on the UR5 wrist was for object detection, while a Hokuyo UTM-30LX LIDAR was used for localization. The object detection pipeline consisted of YOLO V3-based [34] object detector trained on the COCO dataset as well as 15 tiny YOLO V3 detectors each trained on individual classes from the Open Images V4 [35] and YCB [36] datasets. The sensing range was limited to 4.5 m indoors and 7 m outdoors. The natural language understanding models were trained on a data-augmented corpus of approximately 115 instructions annotated separately for perception, behavior, and map inferences. The primary contribution of this work is an efficient approach to instruction following in unknown environments, and not the underlying grounding model itself, which has previously been shown to handle a large diversity of utterances [8, 9, 11, 12, 37].

While we motivate the problem of understanding complex instructions in the context of large-scale unstructured environments, we focus on a pair of targeted experiments to better analyze the behavior of the proposed framework. We designed our experiments to test the impact of adaptive perception (AP) on the runtime of instruction following in two different environmental settings. One of the experiments was designed as a controlled experiment in indoor settings to allow a direct time comparison of task execution times when using adaptive perception against exhaustive perception (EP). The other experiment was designed to demonstrate the exploratory capacity of the architecture. Figure 5 illustrates the experimental setups. In both environments, the first command given was "retrieve the ball inside the box". The controlled experiment followed with a second command of "pick up the crackers box inside the box", whereas the exploratory experiment followed with "go to the crackers box". All commands in these experiments were provided as constituent parse trees. The box containing the ball was was not observable from the starting location, while the box with crackers box was observable.



(a) indoor environment (b) indoor world model (c) outdoor environment (d) outdoor world model

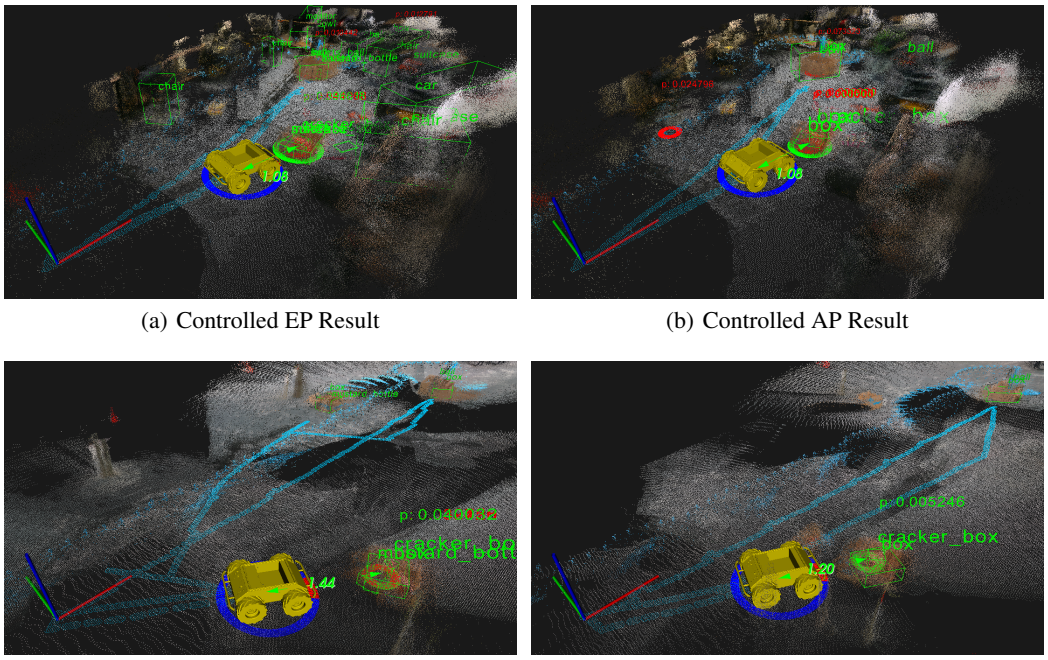
Figure 5: Experiments were conducted in (a) indoor and (c) outdoor environments, and involved instructing the robot to retrieve an object from a box in a priori unknown environments. In both cases, the robot first explores the nearest box, which does not contain the target object. At this point, the robot either (b) explores the target box that comes within the robot's field-of-view (for the indoor experiments), or (d) further explores the environment (for the outdoor experiments).

To allow the robot to build accurate world models of its surroundings the robot's speed was limited to an average of 0.3 m per perception cycle. Ten particles were used to maintain the world distribution for indoor experiments, while we used twenty particles for the outdoor experiments to account for the larger experimental workspace. A spatially sampled subset of past observations was stored during the execution of each behavior. When a second instruction was received during an adaptive perception experiment, a new set of perceptual classifiers was inferred and the semantic map was updated using those classifiers and stored observations.

5 Results and Discussion

We evaluated the effect of adaptive perception on the runtime of various aspects of task execution as outlined in Table 1. World models generated at the end of each trial are depicted in Figure 6. The impact of adaptive perception on the compactness of the generated world models is emphasized more for the indoor experiment as the indoor environment contained a higher number of objects (24 vs. 9). As behavior inference is performed on each hypothesized world model in the distribution, the efficiency gains provided by adaptive perception would enable reasoning over more number of environment hypotheses in the same amount of time. This is important as it allows for maintaining more particles and thus more efficient exploration.

The difference in behavior inference time is less prominent in the outdoor trials due to the the sparsity of the environment. The noticeable reduction in the perception run-time enables our framework to operate more efficiently while processing the same number of observations. This reduces the time required for task execution. However, an advantage of using exhaustive perception is that it does not require re-analyzing past observations when interpreting new instructions. Such is necessary with adaptive perception when subsequent instructions involve new detectors. This resulted in a shorter second task completion time for exhaustive perception in the outdoors environment.



(a) Controlled EP Result (b) Controlled AP Result

(c) Exploratory EP Result (d) Exploratory AP Result

Figure 6: The robot's visualization at the end of the experiment for each of the four trials run. The blue path illustrates the path that the robot took. The collection of green bounding boxes and their respective labels make up the world models generated during each trial. The first command issued in all trials was "retrieve the ball inside the box" and the second command issued was "drive to the crackers box in the box".

	AP Controlled	EP Controlled	AP Exploratory	EP Exploratory
avg. behavior inf. time per world (s)	0.020	0.035	0.016	0.019
avg. perception loop period (s)	0.700	4.141	0.655	4.099
time spent analyzing past obs. (s)	19.6	0	13.5	0
first task time (s)	186.6	351.2	214.4	593.5
second task time (s)	90.5	149.5	22.0	20.4
total detected objects	9	24	8	11

Table 1: Computational efficiency with adaptive (AP, ours) and exhaustive perception (EP).

6 Conclusion

We proposed a novel framework that improves the efficiency of grounding natural language instructions in a priori unknown environments. Integral to our approach is the coupling of three learned language understanding models with distinct symbolic representations for adaptive perception, map inference, and behavior inference. Physical experiments on a mobile manipulator demonstrate higher language grounding efficiency over a contemporary baseline that employs exhaustive perception. In ongoing work, we are exploring hierarchical spatial-semantic representations and more complex mobile manipulation tasks that consider affordances and dynamics of perceived objects.

Acknowledgments

This work was supported in part by the National Science Foundation under grants IIS-1638072 and IIS-1637813 and by ARO grant W911NF-15-1-0402.

References

[1] S. Hemachandra, T. Kollar, N. Roy, and S. Teller. Following and interpreting narrated guided tours. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2011.

[2] J. Thomason, J. Sinapov, M. Svetlik, P. Stone, and R. J. Mooney. Learning multi-modal grounded linguistic semantics by playing "I spy". In *Proc. Int'l Joint Conf. on Artificial Intelligence (IJCAI)*, 2016.

[3] M. Shridhar and D. Hsu. Interactive visual grounding of referring expressions for human-robot interaction. *Proceedings of Robotics: Science and Systems*, 2018.

[4] T. Kollar, S. Tellex, D. Roy, and N. Roy. Toward understanding natural language directions. In *Proc. ACM/IEEE Int'l. Conf. on Human-Robot Interaction (HRI)*, 2010.

[5] C. Matuszek, D. Fox, and K. Koscher. Following directions using statistical machine translation. In *Proc. ACM/IEEE Int'l. Conf. on Human-Robot Interaction (HRI)*, 2010.

[6] C. Matuszek, E. Herbst, L. Zettlemoyer, and D. Fox. Learning to parse natural language commands to a robot control system. In *Proc. Int'l. Symp. on Experimental Robotics (ISER)*, 2012.

[7] J. Thomason, S. Zhang, R. J. Mooney, and P. Stone. Learning to interpret natural language commands through human-robot dialog. In *Proc. Int'l Joint Conf. on Artificial Intelligence (IJCAI)*, 2015.

[8] R. Paul, J. Arkin, D. Aksaray, N. Roy, and T. M. Howard. Efficient grounding of abstract spatial concepts for natural language interaction with robot platforms. *Int'l J. of Robotics Research*, 37(10):1269–1299, 2018.

[9] S. Hemachandra, F. Duvallet, T. M. Howard, N. Roy, A. Stentz, and M. R. Walter. Learning models for following natural language directions in unknown environments. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2015.

[10] E. Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE International Conference on Robotics and Automation*, pages 3400–3407. IEEE, 2011.

[11] S. Patki and T. M. Howard. Language-guided adaptive perception for efficient grounded communication with robotic manipulators in cluttered environments. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2018.

[12] S. Patki, A. Daniele, M. Walter, and T. Howard. Inferring compact representations for efficient natural language understanding of robot instructions. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2019.

[13] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proc. Nat'l Conf. on Artificial Intelligence (AAAI)*, 2011.

[14] M. Walter, M. Antone, E. Chuangsuwanich, A. Correa, R. Davis, L. Fletcher, E. Frazzoli, Y. Friedman, J. Glass, J. How, J. Jeon, S. Karaman, B. Luders, N. Roy, S. Tellex, and S. Teller. A situationally aware voice-commandable robotic forklift working alongside people in unstructured outdoor environments. *J. of Field Robotics*, 2014.

[15] M. R. Walter, R. M. Eustice, and J. J. Leonard. Exactly sparse extended information filters for feature-based SLAM. *Int'l J. of Robotics Research*, 26(4), 2007.

[16] E. Olson, J. Leonard, and S. Teller. Fast iterative optimization of pose graphs with poor initial estimates. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2006.

[17] M. Kaess, A. Ranganathan, and F. Dellaert. iSAM: Incremental smoothing and mapping. *Trans. on Robotics*, 24(6):1365–1378, 2008.

[18] F. Duvallet, T. Kollar, and A. Stentz. Imitation learning for natural language direction following through unknown environments. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, pages 1047–1053, 2013.

[19] H. Mei, M. Bansal, and M. R. Walter. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *Proc. Nat'l Conf. on Artificial Intelligence (AAAI)*, 2016.

[20] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[21] D. L. Chen and R. J. Mooney. Learning to interpret natural language navigation instructions from observations. In *Proc. Nat'l Conf. on Artificial Intelligence (AAAI)*, 2011.

[22] S. Tellex, R. Knepper, A. Li, D. Rus, and N. Roy. Asking for help using inverse semantics. In *Proc. Robotics: Science and Systems (RSS)*, 2014.

[23] Z. Gong and Y. Zhang. Temporal spatial inverse semantics for robots communicating with humans. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2018.

[24] S. Tellex, P. Thaker, R. Deits, T. Kollar, and N. Roy. Toward information theoretic human-robot dialog. In *Proc. Robotics: Science and Systems (RSS)*, 2012.

[25] L. She and J. Chai. Interactive learning of grounded verb semantics towards human-robot communication. In *Proc. Association for Computational Linguistics (ACL)*, 2017.

[26] M. Tucker, D. Aksaray, R. Paul, G. J. Stein, and N. Roy. Learning unknown groundings for natural language interaction with mobile robots. In *Int'l Symp. on Robotics Research*, 2017.

[27] J. Thomason, J. Sinapov, R. J. Mooney, and P. Stone. Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In *Proc. Nat'l Conf. on Artificial Intelligence (AAAI)*, 2018.

[28] H. Zender, O. Martínez Mozos, P. Jensfelt, G. Kruijff, and W. Burgard. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, 56(6), 2008.

[29] A. Pronobis, O. Martínez Mozos, B. Caputo, and P. Jensfelt. Multi-modal semantic place classification. *Int'l J. of Robotics Research*, 29(2–3), 2010.

[30] M. R. Walter, S. Hemachandra, B. Homberg, S. Tellex, and S. Teller. Learning semantic maps from natural language descriptions. In *Proc. Robotics: Science and Systems (RSS)*, 2013.

[31] S. Hemachandra, M. R. Walter, S. Tellex, and S. Teller. Learning spatial-semantic representations from natural language descriptions and scene classifications. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2014.

[32] A. Doucet, N. de Freitas, K. Murphy, and S. Russell. Rao-Blackwellised particle filtering for dynamic Bayesian networks. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 176–183, Stanford, CA, June–July 2000.

[33] D. Hähnel, D. Fox, W. Burgard, and S. Thrun. A highly efficient FastSLAM algorithm for generating cyclic maps of large-scale environments from raw laser range measurements. In *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, 2003.

[34] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[35] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018.

[36] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in manipulation research: The YCB object and model set and benchmarking protocols. *IEEE Robotics and Automation Magazine*, 2015.

[37] J. Arkin, R. Paul, D. Park, S. Roy, N. Roy, and T. M. Howard. Real-time human-robot communication for manipulation tasks in partially observed environments.