Real-Time Human-Robot Communication for Manipulation Tasks in Partially Observed Environments

Jacob Arkin, Rohan Paul, Daehyung Park, Subhro Roy,
Nicholas Roy and Thomas M. Howard

1 Introduction

Effective human-robot teams coordinate activities via communication to optimally perform tasks. In human teams, visual and auditory cues are often used to communicate information about the task and/or environment that may not otherwise be directly observable. Analogously, robots that primarily rely on visual sensors cannot directly observe some attributes of objects that may be necessary for reference resolution or task execution. Algorithms that fuse information communicated by humans about hidden states of objects with information obtained from physical interactions with those objects, such as force measurements and/or joint torques, enable robots to more robustly perform tasks when posed with inaccurate or insufficient information. Also, minimizing latency in understanding language is crucial for human-robot teams because latency in communication reduces mission tempo and limits the effectiveness of collaboration. A robot with a language model that can anticipate a subset of utterances that a human collaborator will likely communicate can leverage idle system time to proactively generate and ground those expressions in the context of its current world model, thus minimizing the latency between receiving an utterance and understanding its meaning.

The experiments in this paper address natural language interaction in human-robot teams for tasks where multi-modal (e.g. visual, auditory, haptic, etc) observations are necessary for robust execution. Contemporary approaches to probabilistic language understanding in partially known environments deal with uncertainty in both physical locations and semantic labels that is resolved upon observation with visual sensors [19, 5, 4]. These works target the navigation domain and focus on the alignment of metric and semantic maps wherein the unobserved aspect of the environment is the locations of hidden objects or locations of landmarks that extend beyond the current map. This is in contrast to the work presented here that targets hidden semantic attributes of objects necessary for successful task execution. Approaches that incorporate declarative knowledge [11, 14, 9] assume that such

information is correct and sufficient for task execution, and thus are not robust to situations in which the declared knowledge is incorrectly understood by the robot or factually inaccurate.

We address the problem of realtime interpretation of instructions in scenarios that require knowledge of objects' states involved in task execution that may not be fully observed from the visual modality alone. We incorporate two information sources for estimating the latent aspects of the workspace. First, we use declarative knowledge from the human embedded in descriptions such as, "the gas can on the left is empty" that provide information about one or more of an object's hidden semantic states. Second, while manipulating the object, the robot's force/torque measurements can also inform the belief over the hidden state of the object. We estimate and propagate a belief in the presence of erroneous observations where language and interaction measurements may indicate different states. Further, reasoning with temporally increasing context adds a computational burden, hence mechanisms are needed for online language grounding.

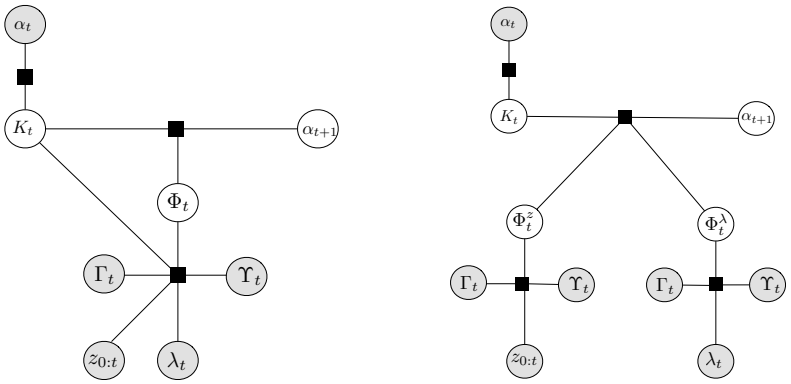
This paper presents a probabilistic model, verified through physical experiments, that allows robots to acquire knowledge about the latent aspects of the workspace through language and physical interaction in an efficient manner. The proposed model builds on three lines of work: (i) efficient language grounding in large symbol spaces [13], where approximation of the complete model is fundamental to efficient inference, (ii) acquiring factual knowledge [14] over a temporally extended visual and linguistic interaction, and (iii) proactively searching for and inferring the meaning of likely phrases given the interaction history and current state of the world [1]. Our approach is robust to noisy or incorrectly asserted semantic information via the ability to incorporate the physical interaction measurements into the internal belief state. We demonstrate the model's effectiveness on a mobile and a stationary manipulator in real-world scenarios following instructions under partial knowledge of object states in the environment.

2 Technical Approach

We introduce a probabilistic graphical model, illustrated in Figure 1, that allows a robot to follow instructions while remaining resilient to incomplete or inaccurate workspace knowledge. Consider a robot manipulator at a time step t has received sensory observations such as declarative language utterances $\lambda_{0:t}$ communicated by a human that express facts about the world (e.g. "the cup on the left is empty"), observations from a visual sensor $x_{0:t}$, and force/torque measurements $z_{0:t}$ during a physical interaction with an object. The robot's world representation consists of two components: (i) a metric model Υ_t derived from visual observations resulting in object locations and (ii) a symbolic model Γ_t comprised of logical predicates that convey semantic attributes/relationships that may be true for a possible world. A visual sensor can inform the symbolic states by providing observations of entities (blocks, cups, etc.) and spatial relationships (on, inside, front, etc.). Task execution

in real-world domains often requires knowledge of the intrinsic object attributes that cannot be determined from visual observations alone, however. For example, executing the instruction "clear away the cups on the table" requires knowledge of the internal states of the cups (full or empty) to decide whether the cup should be trashed or put away in a store.

In order to estimate a belief over the unobserved object states, we incorporate language descriptions from a human (e.g. "the cup on the table is empty") as well as force measurements during the robot's physical interaction (e.g. end-effector force profiles during manipulation). This belief over unobserved object states is included in the probabilistic graphical model as a random variable K_t that represents the state of knowledge about semantic object attributes in the world. For example, if the workspace contains a cup and a box, the knowledge state K_t would include a set of binary variables such as $IsFull(cup)$, $IsMovable(box)$ etc. Due to the diverse and large space of both language and percepts, joint inference about the knowledge state is challenging. Below, we describe our approach for making the learning problem computationally tractable.



(a) Joint model **(b)** Factored model

Fig. 1: Probabilistic models for knowledge acquisition over latent object attributes from descriptive language utterances and force measurements during physical interaction. (a) The joint model postulates a single grounding factor for both percepts $z_{0:t}$ and language λ_t resulting in likely correspondences Φ_t propagated to the next time step α_{t+1}. Learning in this joint model is challenging due to the large combined space of possible language utterances and interaction observations. (b) The factored model postulates separate grounding factors that relate percepts $z_{0:t}$ and language λ_t with likely correspondences Φ_t^z and Φ_t^λ respectively. The estimated groundings from both visual and linguistic modalities are fused to inform a posterior belief over the latent semantic state α_t propagated to the next time step.

2.1 Inferring Latent Semantic States

Our goal is to infer the latent semantic states of the objects (e.g., empty/full etc.) from language utterances and force measurements during physical interaction. Each semantic object state is symbolic in nature and represented as a discrete binary random variable; we assume that object properties are independent. We assume a

Bernoulli distribution over the latent semantic states of objects and impose a conjugate Beta distribution prior with hyper-parameter α_t [2]. The distribution over the state K_t is parameterized by α_t and models the current belief over the true likelihood of a symbolic attribute (e.g., $p(IsFull(cup) = 1)$). Formally, the true likelihood over the predicate is $p(IsFull(cup) = 1) \sim Ber(\theta)$ with $\theta \sim Beta(a_t, b_t)$ where $\alpha_t = \{a_t, b_t\}$. The model fuses the factual knowledge provided via language from the human and force/torque measurements acquired from direct physical interaction to estimate a belief over the latent symbolic state of the object. Formally, we pose the problem as inference on a dynamic Bayesian network, incrementally fusing observations to yield a posterior estimate over the latent knowledge state as:

$$p(\alpha_{t+1} | \lambda_{0:t}, z_{0:t}, \Upsilon_t, \Gamma_t, \alpha_t) \tag{1}$$

where α_{t+1} is the hyper-prior over the knowledge predicate at time $t+1$.

The task of inferring symbolic knowledge from language and force/torque measurements can be formulated as estimating correspondences between measurements and probable semantic concepts. Following prior work on language grounding [7, 13, 14], we introduce a correspondence variable Φ_t to indicate the association between language or percepts and hidden object states. Additionally, the diverse nature of the observations makes direct estimation of Equation 1 computationally challenging; so, we introduce a factorization wherein the estimation of the knowledge update is conditioned on the current hyper-prior and the state of the observation correspondence variable. As illustrated in Figure 1a, we can update the formulation of the desired distribution accordingly:

$$p(\alpha_{t+1} | \lambda_{0:t}, z_{0:t}, \Upsilon_t, \Gamma_t, \alpha_t) = \sum_{\Phi_t} \overbrace{p(\alpha_{t+1} | \Phi_t, \alpha_t)}^{\text{Knowledge Update}} \overbrace{p(\Phi_t | \lambda_{0:t}, z_{0:t}, \Upsilon_t, \Gamma_t)}^{\text{Observation Correspondences}} . \tag{2}$$

However, this formulation suffers from the challenge of jointly inferring a correspondence for the language and force/torque observations. We assume conditional independence between visual and linguistic modalities and factorize the conditional probability of the observation correspondence variable Φ_t; we represent the correspondence variables for language and force/torque observations as Φ_t^λ and Φ_t^z respectively. This factorized model, illustrated in Figure 1b, is equivalent to Equation 1 under the stated assumptions and can be expressed as:

$$\sum_{\Phi_t^\lambda} \sum_{\Phi_t^z} \overbrace{p(\alpha_{t+1} | \Phi_t^\lambda, \Phi_t^z, \alpha_t)}^{\text{Knowledge Update}} \overbrace{p(\Phi_t^\lambda | \lambda_{0:t}, \Upsilon_t, \Gamma_t)}^{\text{Lang. Corresp.}} \overbrace{p(\Phi_t^z | z_{0:t}, \Upsilon_t, \Gamma_t)}^{\text{Percept Corresp.}} . \tag{3}$$

The language grounding factor involves estimating probable correspondences Φ_t^λ between the language utterance and semantic concepts that the robot can interpret (e.g., a grounding $IsEmpty(cup)$ inferred from "the cup on the table is empty"). This factor is trained using a structured log-linear model with features extracting linguistic cues and spatial attributes of the current world state [7]. The estimation

of semantic attributes from the human's utterance can be viewed as a declarative top-down inference over symbolic knowledge. A second source of symbolic knowledge comes from signatures derived from the robot's physical interaction with the object. The model infers latent object states (*IsFull*, *IsMovable* etc.) using a hidden Markov model (HMM) with time series observations of the 3-axis end-effector force and arm pose measurements recorded during the physical interaction with the object. The HMM is trained using 30 interaction traces with varied object states and configurations. The inference results in likely correspondences Φ_t^z, a process that can be viewed as a bottom-up source of symbolic knowledge derived from grounding raw force measurements from physical object manipulation.

As shown above in Equation 3, the two sources of symbolic knowledge, (i) the top-down groundings from descriptive language and (ii) the bottom-up groundings from interaction measurements, can then be fused to inform a posterior belief over the latent knowledge state. The conjugacy property of the Beta-Bernoulli distributions allows closed-form posterior updates marginalizing out the likelihood over the object states. The updated Beta distribution parameters are estimated as:

$$\alpha_{t+1} = \{a_t + n_\lambda^1 + n_z^1, b_t + n_\lambda^0 + n_z^0\}, \tag{4}$$

where $\{n_\lambda^1, n_\lambda^0\}$ and $\{n_z^1, n_z^0\}$ indicate the number of true and false observations derived from language $\Phi_{0:t}^\lambda$ and haptic groundings $\Phi_{0:t}^z$ respectively. The Bayesian treatment can be viewed as state estimation over logical symbolic variables and offers resilience to noisy or contradictory observations from language and physical interactions.

2.2 Proactive Instruction Following & Deliberative Interaction

Given an input language instruction, the robot must select actions μ_{t+1} as per the human's instructions and the robot's current belief over the world state as maintained as per the model described in Section 2.1. We extend the Temporal Grounding Graphs [14] for estimating intended manipulation goals from an input instruction. As discussed, the task of interpreting an instruction requires knowledge of hidden attributes for objects involved in the intended plan execution. For example, the interpretation of the instruction "clear the cups on the table" requires knowledge of the latent full/empty binary states of the cups to decide their appropriate destinations in the clearing task. We determine the expected state likelihoods by sampling the Bernoulli distribution from the current Beta prior. The model selects actions based on a confidence measure, computed as normalized entropy of the distribution, over the states of task-relevant objects. An uninformed belief necessitates information gathering actions such as repeatedly manipulating the cup until the latent belief is sufficiently informed to execute the intended action. The model formulation for inferring μ_{t+1} can be seen below:

$$p(\mu_{t+1}|\lambda_t,\Upsilon_t,\Gamma_t,\alpha_t) = \int_{K_t}\sum_{\lambda_t}\overbrace{p(\mu_{t+1}|K_t,\Phi_t^{\lambda})}^{\text{Estimating Actions}}\overbrace{p(K_t|\alpha_t)}^{\text{Belief}}\overbrace{p(\Phi_t^{\lambda}|\lambda_t,\Upsilon_t,\Gamma_t)}^{\text{Language Grounding}} \quad (5)$$

.

The language grounding factor acts as a computational bottleneck as it involves a search over a large space of interpretations for an input instruction. Rather than reactively interpreting a full instruction, we proactively compute groundings that are likely to be relevant for future instructions. This results in real-time inference by boot-strapping a novel utterance with estimated groundings (true correspondences) from the set of proactively grounded phrases possessing similar parse structure. Formally, the set of proactive groundings Φ_t^{psg} is determined as a function of the current environment state Υ_t and the space of instructions determined by a grammar G. Since conditional independence is assumed across both individual phrases within the parse tree and individual groundings within the full space of semantic concepts, any given phrase with the same Υ_t will always ground to the same set of symbols regardless of parent phrases in the parse tree. Once the symbols that correspond to a simple phrase have been found, they can be reused within more complex phrases for as long as changes in the environment do not alter their meaning. Further, the set of candidate groundings are estimated by constraining the space of possible sentences from a vocabulary by imposing linguistic production rules from a phrase grammar. We leverage the hierarchical and compositional structure of language to construct proactive grounding sets in a bottom-up manner. The pre-computation of likely groundings reduces the runtime for interpreting a novel instruction by restricting the online computation to salient phrases λ_t^s which are typically smaller than the full set of phrases in the instruction ($|\lambda_t^s| \leq |\lambda_t|$). The resulting model formulation is equivalent to Equation 5 and can be mathematically described as:

$$\int_{K_t}\sum_{\lambda_t^s}\overbrace{p(\mu_{t+1}|K_t,\Phi_t^{\lambda})}^{\text{Estimating Actions}}\overbrace{p(K_t|\alpha_t)}^{\text{Belief}}\overbrace{p(\Phi_t^{\lambda}|\lambda_t^s,\Upsilon_t,\Gamma_t)p(\lambda_t^s|\lambda_t,\Phi_t^{psg})}^{\text{Proactive Language Gnd.}}. \quad (6)$$

3 Experiments & Results

In order to validate the performance of the proposed system and its components, we designed three independent experiments. The first experimental evaluation targets the proactive symbol grounding component in simulation and quantitatively compares the inference runtime to a reactive baseline. For this experiment, we assumed a sufficiently expressive symbolic representation (see [13]), a grammar, and a corpus of annotated examples used for training. To quantify performance, we trialed different durations of proactive grounding time, increasing from 0 seconds to 8 seconds in 2 second intervals, during which the process grounded candidate phases, illustrated in Table 1 as "PSG Duration" (Proactive Symbol Grounding Duration) and "Number of Grounded Phrases" respectively. The row "NLSG Inference Time" (Natural

Language Symbol Grounding Time) reports the runtime for a novel utterance; as expected, the runtime decreases as a function of the PSG Duration due to the process generating more matches to phrases in the novel utterance's parse tree and thus reducing the number of phrases to be computed at inference time. We include a trial with 0 seconds of proactive grounding time to establish a baseline of performance for the natural language symbol grounding process without any boot-strapping by the proactive grounding module.

PSG Duration (sec)	0.0	2.0	4.0	6.0	8.0
Number of Grounded Phrases	0	31	62	102	146
NLSG Inference Time (sec)	0.21	0.18	0.14	0.13	0.09

Table 1: The proactive symbol grounding runtime (PSG) versus natural language symbol grounding (NLSG) runtime for a single instruction.

The second experimental evaluation used a Baxter Research Robot in a table top setup populated with household objects as shown in Figure 3. In the first scenario, the robot's workspace contained two coffee cups (with closed lids), a tray and a trash can; the internal state of the cups was hidden with one cup being empty and the other full. We assume that the robot possesses learned background knowledge that empty cups are to be discarded in the trash and full cups are to be placed on the tray. As discussed in Section 2.1, the robot also possesses a trained hidden Markov model for classifying signatures from physical interaction with the cups. A plot of the different z-axis force measurements for a full and an empty cup can be seen in Figure 2. The robot did not have access to the internal state of the cups. The robot was instructed to, "clear away the cups on the table," resulting in a grounded solution referencing the two coffee cups. The grounding model estimated the probable grounding of the sentence as the two cups on the table. The robot picked up each, updating the belief over the internal states according to force/torque sensing. This knowledge allows the robot to estimate the correct location to discard the empty cup in the trash or to place the filled cup on the tray.

In a subsequent scenario, the human declared, "the cups on the table are empty," before instructing the robot to "clear away the cups." Contradictory to the initial statement, the actual state of one of the cups is filled and should not be discarded. The robot determined the true state of the cups during interaction, correctly updating its prior belief from force/torque sensing and choosing the correct actions.

Figure 4 shows the resulting changes to both the Beta distribution and the expected likelihood of the expressed fact as the robot interacts with one of the cups in the second scenario. The robot first receives a declarative fact from language expressed as "the cups on the table are empty", leading to a posterior update to the Beta hyper-prior for the likelihood using the estimated grounding ($IsFull(cup)$ = True). Upon engaging in a time-series of physically interactions with the cup whose hidden state is actually $IsFull(cup)$ = False, the robot successively updates the latent belief over the symbolic state. The robot interacts with the object until the normalized entropy of the latent distribution is sufficiently informative (set via a threshold). The

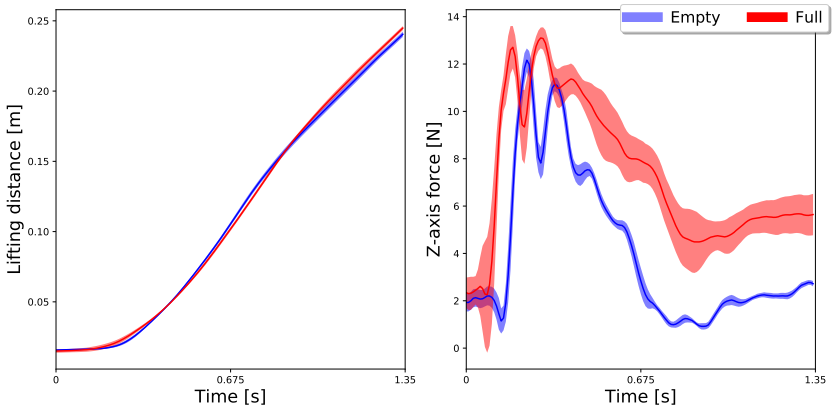


Fig. 2: Lifting distance and z-axis force measurements over time for both full (red) and empty (blue) cups. The patterns of force measurements over lifting distances are modeled by two HMMs which are then leveraged during log-likelihood-based binary classification to infer an object's state.

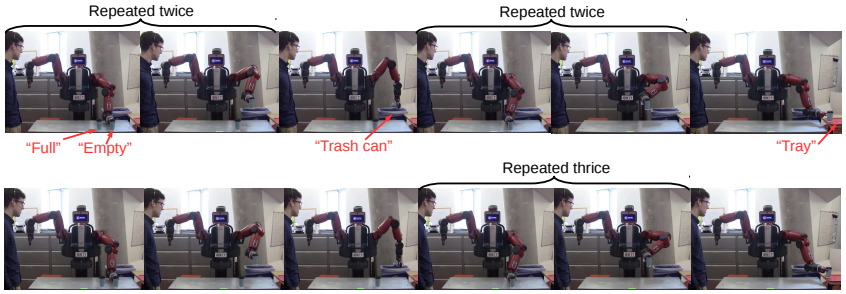
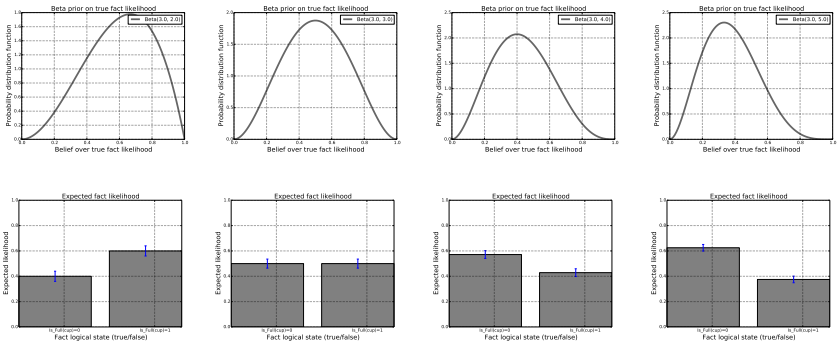


Fig. 3: The second experiment evaluating knowledge acquisition over latent object states from declarative knowledge and physical interaction. The Baxter robot was instructed to "clear away the cups on the table". **Top:** the robot attempts to pick up each cup in turn and infers the latent state of the cups from the time series of interactions. Once the belief is sufficiently informed the robot discards the empty cup in the trash bin and the filled cup on the tray. **Bottom:** the human informs the robot that "the cups on the table are empty" a fact that is true only for only one of the cups. The robot's physical interaction results in a posterior belief correcting the prior resulting from the incorrectly stated fact and correctly accomplished the task of clearing in correct locations.

estimation of the correct belief allows the robot to correctly follow the instruction of clearing the empty cups despite initially receiving an incorrect fact from the human.

In the third experimental evaluation, we tested an integrated system that incorporates both the proactive symbol grounding process for fast inference and the joint use of declarative knowledge and force sensing for updating beliefs about objects' states. We used a Clearpath Husky A200 mounted with a Universal Robots UR5 manipulator in a mobile manipulation setting composed of two cases, as shown in Figure 5; the internal state of the cases was hidden. The case on the robot's left was full and heavy, and the case on the right was empty and light. We executed three different types of scenarios in this experiment: (i) no declarative knowledge, (ii) accurate declarative knowledge, and (iii) inaccurate declarative knowledge. In



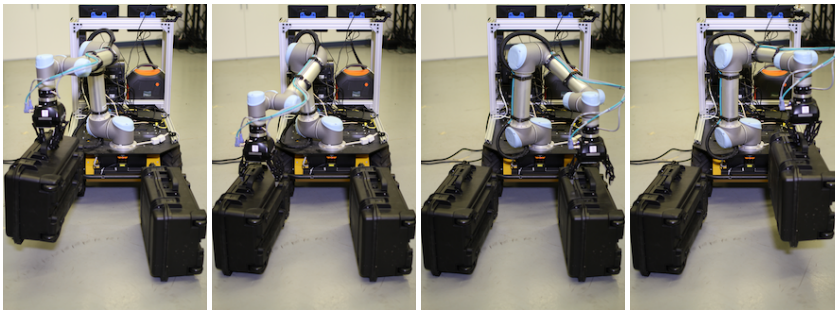
(a) Lang. $IsFull(cup) = 1$ **(b)** Obs. $IsFull(cup) = 0$ **(c)** Obs. $IsFull(cup) = 0$ **(d)** Obs. $IsFull(cup) = 0$

Fig. 4: Temporal evolution of belief over factual knowledge informed by language and interaction. The Beta distribution at time t for the Bernoulli likelihood over factual groundings is plotted in top row. The maximum likelihood true likelihoods for a predicate state appears below. Temporal evaluation from left to right. The initials "Lang." and "Obs." denote estimated groundings obtained from language and time series interaction data respectively. The estimation of the correct belief allows the robot to correctly follow the instruction of clearing the empty cups to the trash and placing the fill cup on the tray.

one case of (i), the husky was instructed to, "pick up the heavy case" resulting in an ambiguous grounded reference solution. The robot picked up the left case, updating the belief that it was heavy; a second interaction made the robot confident enough to complete the action. In one case of (ii), the human accurately declared, "the case on the left is heavy" followed by, "pick up the heavy case." The robot picked up the left case, updating its belief which reinforced the human's provided fact. A single force/torque interaction and the accurate declared fact made the robot sufficiently confident to complete the action; the fact reduced the number of required interactions. In one case of (iii), the human declared, "the case on the right is heavy" followed by, "pick up the heavy case." The robot picked up the case on the right, updating its belief in contradiction to the human's provided fact. The robot then lifted the left case twice to be sufficiently confident and complete the action. A video demonstrating the experiments with both the Baxter and Husky robots can be found here: *https://www.youtube.com/watch?v=JTVJkJavU6g*

4 Related Work

The approach described in this paper and the associated experiments address human-robot collaborative tasks in which multi-modal observations (language, vision, and haptic) are required to sufficiently inform the robot about latent object attributes necessary for task completion. Some contemporary approaches to probabilistic language understanding in partially known environments dealt with uncertainty in the metric location of objects or landmarks that was resolved upon observation with the visual sensors [19, 5, 4, 12]. We address a different element of "partial observabil-



Initial state of the right case is heavy. | Updated belief is uncertain about heavy case. | Interaction with the other case. | Updated belief that the left case is heavy.

Fig. 5: A third experiment incorporating both proactive symbol grounding and updates to beliefs about objects' states via declarative knowledge and force/torque sensing. The Husky robot with a mounted robot arm was inaccurately told, "the case on the right is heavy" before receiving the instruction "pick up the heavy case."

ity" by inferring the state of object attributes as opposed to hypothesized locations of objects or landmarks that exist beyond the robot's current visual field or internal map of the explored world. We also incorporate a novel knowledge state variable in our graphical model and incrementally update a distribution over that knowledge state rather than reason over a distribution of maps.

There are other recent approaches for leveraging multi-modal observations to learn object attributes. In some work, human gesture was incorporated as a modality to learn object and relation classifiers, as well as attributes such as color [8, 10, 20]. Others have incorporated audio and haptic as modalities to learn visually-hidden attributes [3], such as whether a container is full or not based on the sounds produced while picking up and shaking [17]. There is also work to learn behaviorally-grounded or sensorimotor-grounded classifications [6], such as the work by Sinopav et al that uses vision, proprioception, and audio to learn semantic labels for objects via the robot's interaction with them [16].

Our contribution leverages language as a source of information about latent object states by grounding declarative statements from user utterances. Other natural language symbol grounding approaches that incorporate declarative knowledge [11, 18, 14, 9, 15] assumed that such information is correct and sufficient for task execution. In contrast, our model incrementally fuses information from language and force/torque interactions making task execution more robust to inaccurate or incorrectly understood declarations.

5 Discussion & Conclusion

This work demonstrated the ability to proactively infer a context for a future instruction from past interactions and world state. The ability to proactively predict the set of likely groundings has a significant improvement in the runtime of ground-

ing a command, contributing towards real-time online symbol grounding in complex workspaces. Further, we observe that the task of interpreting an instruction is crucially tied to the semantic knowledge the robot has acquired about its workspace. In real-world scenarios, the robot only has partial knowledge about its workspace. We demonstrated how both linguistic descriptions from a human and signatures derived from the robot's physical interaction can be used to infer the latent semantic properties of the environment. We showed how a probabilistic model can update and maintain a belief over the latent knowledge state of the world from noisy linguistic and interaction observations. The experiments in this work contribute towards bridging the gap between higher-order inputs from the human such as language and the low-level representation of the robot such as cost functions, interaction forces etc. via grounded learning of semantic concepts.

As part of future work, we intend to demonstrate the effectiveness of the system on additional robot platforms in field experiment settings as part of an effort to both further test the reproducibility of our results and explore how our approach can be integrated into a larger system architecture for complex human-robot teaming scenarios. In addition, we seek to expand this framework to acquiring world knowledge in the context of multi-step plans common in complex missions rather than functioning as a multi-modal strategy for resolving reference ambiguity. To support this goal, we intend to increase the number of concepts expressed in training and consider more complex correlations and background knowledge in state estimation.

6 Acknowledgements

Authors gratefully acknowledge funding support in part by the Robotics Consortium of the U.S. Army Research Laboratory under the Collaborative Technology Alliance Program (RCTA) and the Toyota Research Institute (TRI) Award Number LP-C000765-SR. We thank our colleagues in the lab for helpful feedback on this paper.

References

1. J. Arkin and T. M. Howard. Experiments in proactive symbol grounding for efficient physically situated human-robot dialogue. In *Late-breaking Track at the SIGDIAL Special Session on Physically Situated Dialogue (RoboDIAL-18)*, July 2018.
2. C. M. Bishop. Probability distributions. In M. Jordan, J. Kleinberg, and B. Schölkopf, editors, *Pattern Recognition and Machine Learning*, chapter 2. Springer-Verlag New York, 2006.
3. V. Chu, I. McMahon, L. Riano, C. G McDonald, Q. He, J. M. Perez-Tejada, M. Arrigo, T. Darrell, and K. J. Kuchenbecker. Robotic learning of haptic adjectives through physical interaction. *Robotics and Autonomous Systems*, 63:279–292, 2015.
4. F. Duvallet, M. R. Walter, T. M. Howard, S. Hemachandra, J. Oh, S. Teller, N. Roy, and A. Stentz. Inferring maps and behaviors from natural language instructions. In *Experimental Robotics*, pages 373–388. Springer, 2016.

5. 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)*, Seattle, WA, May 2015.
6. V. Hogman, M. Bjorkman, and D. Kragic. Interactive object classification using sensorimotor contingencies. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 2799–2805. IEEE, 2013.
7. T. M. Howard, S. Tellex, and N. Roy. A natural language planner interface for mobile manipulators. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, pages 6652–6659. IEEE, 2014.
8. T. Kollar, J. Krishnamurthy, and G. P. Strimel. Toward interactive grounded language acqusition. In *Robotics: Science and Systems*, volume 1, pages 721–732, 2013.
9. T. Kollar, V. Perera, D. Nardi, and M. Veloso. Learning environmental knowledge from task-based human-robot dialog. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, pages 4304–4309. IEEE, 2013.
10. C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proc. Nat'l Conf. on Artificial Intelligence (AAAI)*, 2014.
11. C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proc. of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland, June 2012.
12. J. Oh, T. M. Howard, M. R. Walter, D. Barber, M. Zhu, S. Park, A. Suppe, L. Navarro-Serment, F. Duvallet, A. Boularias, O. Romero, J. Vinkkrov, T. Keegan, R. Dean, C. Lennon, B. Bodt, M. Childers, J. Shi, K. Daniilidis, N. Roy, C. Lebiere, M. Hebert, and A. Stentz. Integrated intelligence for human-robot teams. In *Proceedings of the 2016 International Symposium on Experimental Robotics*, October 2016.
13. 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. *International Journal of Robotics Research*, June 2018.
14. R. Paul, A. Barbu, S. Felshin, B. Katz, and N. Roy. Temporal grounding graphs for language understanding with accrued visual-linguistic context. In *Proc. of the Twenty-Sixth Int. Joint Conf. on Artificial Intelligence*, pages 4506–4514, 2017.
15. I. E. Perera and J. F. Allen. Sall-e: Situated agent for language learning. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, page 1241–1247, 2013.
16. J. Sinapov, C. Schenck, K. Staley, V. Sukhoy, and A. Stoytchev. Grounding semantic categories in behavioral interactions: Experiments with 100 objects. *Robotics and Autonomous Systems*, 62(5):632–645, 2014.
17. J. Sinapov and A. Stoytchev. From acoustic object recognition to object categorization by a humanoid robot. In *Proc. of the RSS 2009 Workshop on Mobile Manipulation, Seattle, WA*, 2009.
18. J. Thomason, J. Sinapov, M. Svetlik, P. Stone, and R. J. Mooney. Learning multi-modal grounded linguistic semantics by playing "i spy". In *IJCAI*, pages 3477–3483, 2016.
19. M. R. Walter, S. Hemachandra, B. Homberg, S. Tellex, and S. Teller. A framework for learning semantic maps from grounded natural language descriptions. *Int'l J. of Robotics Research*, 33(9):1167–1190, 2014.
20. D. Whitney, M. Eldon, J. Oberlin, and S. Tellex. Interpreting multimodal referring expressions in real time. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, pages 3331–3338. IEEE, 2016.