

Multimodal estimation and communication of latent semantic knowledge for robust execution of robot instructions

The International Journal of
Robotics Research
1–26
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0278364920917755
journals.sagepub.com/home/ijr

Jacob Arkin^{1*} , Daehyung Park^{2*} , Subhro Roy², Matthew R Walter³,
Nicholas Roy², Thomas M Howard¹ and Rohan Paul^{2,4,#}

Abstract

The goal of this article is to enable robots to perform robust task execution following human instructions in partially observable environments. A robot's ability to interpret and execute commands is fundamentally tied to its semantic world knowledge. Commonly, robots use exteroceptive sensors, such as cameras or LiDAR, to detect entities in the workspace and infer their visual properties and spatial relationships. However, semantic world properties are often visually imperceptible. We posit the use of non-exteroceptive modalities including physical proprioception, factual descriptions, and domain knowledge as mechanisms for inferring semantic properties of objects. We introduce a probabilistic model that fuses linguistic knowledge with visual and haptic observations into a cumulative belief over latent world attributes to infer the meaning of instructions and execute the instructed tasks in a manner robust to erroneous, noisy, or contradictory evidence. In addition, we provide a method that allows the robot to communicate knowledge dissonance back to the human as a means of correcting errors in the operator's world model. Finally, we propose an efficient framework that anticipates possible linguistic interactions and infers the associated groundings for the current world state, thereby bootstrapping both language understanding and generation. We present experiments on manipulators for tasks that require inference over partially observed semantic properties, and evaluate our framework's ability to exploit expressed information and knowledge bases to facilitate convergence, and generate statements to correct declared facts that were observed to be inconsistent with the robot's estimate of object properties.

Keywords

Human–robot collaboration, semantic state estimation, Bayesian modeling, multimodal interaction, natural language understanding

1. Introduction

Our goal is to enable a robot to understand and robustly execute high-level commands from a human in partially known workspaces. Communication is integral to effective coordination and collaboration among human–robot teams. In human teams, perceptual and auditory descriptions are often used to understand the environment and communicate intent about the task and/or environment that may not otherwise be directly observable. Similarly, robots that primarily rely on visual sensors cannot directly observe all attributes of objects in which some attributes may be necessary for reference resolution or task execution. For example, as shown in Figure 1, the knowledge of whether an object can be pushed or moved by a robot manipulator, or whether it is heavier in comparison with another object, may be relevant for manipulation tasks but difficult to estimate from vision

¹Robotics and Artificial Intelligence Laboratory, University of Rochester, USA

²Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA

³Robot Intelligence through Perception Laboratory, Toyota Technological Institute at Chicago, USA

⁴Department of Computer Science and Engineering, Indian Institute of Technology Delhi, India

*Jacob Arkin and Daehyung Park contributed equally to this work.

#Rohan Paul is currently at IIT Delhi. This project was initiated while he was a postdoc at MIT.

Corresponding author:

Jacob Arkin, University of Rochester, Hajim School of Engineering and Applied Sciences, Electrical and Computer Engineering Department, Robotics and Artificial Intelligence Lab (RAIL), Rochester, NY 14627, USA.

Email: jarkin@ur.rochester.edu

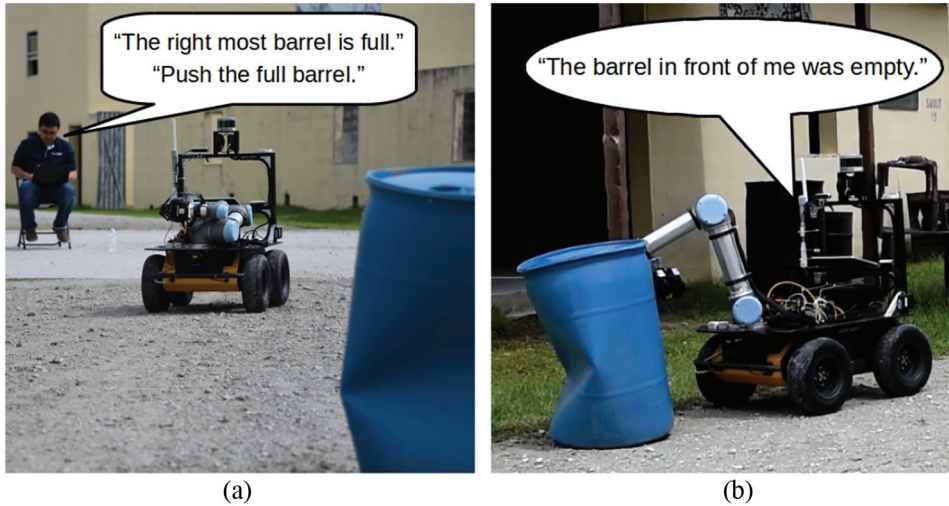


Fig. 1. Bi-directional communication for human-robot teams: (a) a Husky with a UR5 arm, understanding language utterances in a partially known environment; (b) multimodal semantic knowledge estimation followed by linguistic feedback generation to the human operator. A human operator can share his mental model of an object with a robot by stating declaratively that “the barrel on the right is full.” However, the shared world knowledge can be inaccurate in partially observable environments. Upon updating the world knowledge state via physical estimation, the robot reports back a declarative statement in order to correct the operator’s mental model.

alone. The lack of knowledge of non-visual properties may make it impossible to synthesize plans or lead to unanticipated failures during plan execution.

In this work, we address the problem of inferring semantic properties of the world that may not be observable from exteroceptive modalities such as visual or LiDAR sensors. We incorporate three information sources for estimating the latent world properties. First, we use factual, task-relevant knowledge that is implicit or explicit in the natural language communication between the robot and its human partner. For example, the utterance “the nearest barrel is empty” provides factual knowledge about a property of the indicated object. Second, we leverage the robot’s ability to directly interact with the world to inform its belief over the latent attributes of the environment. Force and torque observations and other end-effector measurements provide cues about physical properties of an object, such as whether it can be pushed or lifted, or whether it is pliable. Third, we utilize commonsense knowledge about particular object types (e.g., that plastic containers are typically lighter and less rigid than their metal counterparts) present in crowdsourced corpora, such as the VerbPhysics dataset (Forbes and Choi, 2017), derived from human judgement annotations.

We present a probabilistic model and inference algorithm that estimate semantic knowledge about the workspace through natural language communication, physical interaction measurements, and background knowledge sources. This is a challenging estimation problem as it involves distilling high-level semantic knowledge from low-level measurements arising from physical interactions or highly complex and varied sources such as human language utterances and relational data stores. We present a probabilistic model that fuses measurements from multiple modalities into a

probabilistic belief over the latent semantic knowledge about world entities. We factor the inference task into one of estimating the presence of semantic properties from each modality and of temporally fusing the semantic observations into a probabilistic belief that is robust to erroneous or contradictory evidence. We show how the robot can use this model to plan exploratory actions to improve its belief over latent semantic properties of its world model. The ability to infer missing semantic aspects of the world allows robots to follow instructions while remaining resilient to incomplete or inaccurate workspace knowledge.

Further, we observe that effective human-robot teaming requires seamless communication as well as transparent ways to provide feedback in case of observed discrepancies between the mental model of the human and that of the robot. We describe how a robot can learn to synthesize linguistic feedback to the human operator when the robot’s direct observations differ from the inferred model of the human. Finally, we address the problem of reducing latency in instruction interpretation and feedback generation that arises while evaluating possible associations between language utterances and semantic entities in the world, particularly in large environments. We propose an approach that anticipates future language interactions based on changes in the environmental context and the robot’s environmental knowledge. This allows the robot to pre-compute associations, thereby reducing the latency of future command interpretation and language generation tasks.

We demonstrate the model’s effectiveness in real-world scenarios in which fixed or mobile manipulation platforms follow natural language instructions in environments that are only partially known. By fusing declarative knowledge provided by natural language with observations made

during physical interactions, our method successfully infers the latent object attributes necessary for task execution. We show that the proactive approach to language understanding and feedback generation improves the run-time performance. The proposed model builds on the following lines of work: (i) efficient language grounding in large semantic spaces (Paul et al., 2018), where the approximation of the complete model is fundamental to efficient inference; (ii) acquiring factual knowledge (Paul et al., 2017) over a temporally extended visual and linguistic interaction; (iii) learning an informed belief from background knowledge corpora; and (iv) improved efficient communication by proactively searching for and inferring the meaning of likely phrases given the interaction history and current state of the world (Arkin and Howard, 2018).

Contemporary approaches that incorporate declarative knowledge (Kollar et al., 2013b; Matuszek et al., 2012a; Paul et al., 2017) 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. Approaches such as those of Walter et al. (2013), Walter et al. (2014b), Hemachandra et al. (2015), and Duvallet et al. (2014) incorporate language in semantic mapping in partially known environments in order to simultaneously infer a metric map and semantic labels for regions from visual or range-based observations. Similarly, Daniele et al. (2017a) used language to learn kinematic models of articulated objects. Our work expands the scope of semantic properties from region types alone to fine-grained physical and abstract properties of objects and further incorporates active interaction and high-level commonsense knowledge for making predictions.

This article expands significantly on an earlier conference paper describing this framework (Arkin et al., 2018). We present a thorough exposition of the proposed model with additional technical details, an expanded background and problem formulation, and a more thorough description of related work. We extend the core technical contributions in the following ways. First, we incorporate a data-driven model to estimate an informed prior over object attributes derived from background commonsense knowledge corpora. Second, we extend the model to provide linguistic feedback to the human in the event that there is disagreement between the human’s inferred model of the environment and the robot’s internal estimate derived from physical interaction. Third, we include new experimental results and additional field demonstrations.

This article is organized as follows. We present the background material and problem formulation in Section 2. Section 3 presents the model for representing semantic knowledge and details the process of fusing multiple modalities into a probabilistic belief over the correctness of semantic aspects of the world model. Section 4 approaches the problem of command following in a manner that takes into account uncertainty in the acquired knowledge of entities in the scene. In Section 5, we present an approach

for providing feedback to the human operator when discrepancies are detected between the human’s inferred model of the environment and that of the robot. Section 6 tackles the crucial issue of reducing latency in command understanding as well as linguistic feedback generation. The experimental evaluation and results are described in detail in Section 7. Section 8 is devoted to reviewing related efforts. Finally, Section 9 concludes the article and lays out avenues for future research.

2. Problem formulation

2.1. Robot and workspace model

We consider a robot manipulator operating in a workspace populated with a set of rigid bodies \mathcal{O} . Let Y_t denote the metric state of the world at time t that includes the pose of the robot and other entities in the scene, typically populated by a perception system. A human operator communicates with the robot through a natural language interface. Let Λ_t denote the language utterance received by the robot at time t . We assume that the human either instructs the robot to perform high-level tasks, such as “clearing,” “packing,” “inspection,” etc. or provides factual descriptions, such as “the barrel on the left is empty.”

The robot’s goal is to derive a plan that affects the world state in order to satisfy the human’s command. We model the plan μ_t as a sequence of actions that change the state of the world, such as “grasping,” “moving,” “placing,” “pushing,” or “poking” an object. We assume that the robot makes proprioceptive measurements of the world through physical interaction with its surroundings. Let $Z_t = \{z_{t_0}, z_{t_1}, \dots, z_{t_n}\}$ denote a proprioceptive observation recorded at time t that consists of a sequence of force/torque measurements and manipulator poses observed during interaction.

The robot’s decision-making and planning requires semantic knowledge about the world. We present a representation and a framework for estimating semantic aspects of the world in the next section.

2.2. Semantic attributes and knowledge

Let Γ denote the space of concepts or “groundings” that express semantic properties of the world. Groundings model semantic *attributes* associated with entities (e.g., class types and factual knowledge) as well as *relationships* between entities (e.g., spatial relations and relative orientations). We represent concepts as a set of discrete symbols using the predicate-role representation (Russell and Norvig, 2016). Each predicate represents a semantic property or a relationship $\sigma \in \Sigma$ that is expressed for a certain set of entities in the robot’s world model $\mathbf{o} \subseteq \mathcal{O}$. The space of grounding symbols Γ can be expressed as

$$\Gamma = \{(\sigma, \mathbf{o}) | \sigma \in \Sigma, \mathbf{o} \subseteq \mathcal{O}\} \quad (1)$$

A class of grounding symbols models Boolean object categories such as $\text{IsBlock}(\mathbf{o})$, $\text{IsBarrel}(\mathbf{o})$, and $\text{IsBox}(\mathbf{o})$,

where $o \in \mathcal{O}$ is an object instance in the world model. A second category of symbols expresses physical object properties, such as $\text{IsMovable}(o)$, $\text{IsHeavy}(o)$, or $\text{IsPushable}(o)$. A third class of symbols models spatial relationships, such as $\text{Front}(o_i, o_j)$, $\text{Left}(o_i, o_j)$ or $\text{Inside}(o_i, o_j)$, between object instances o_i and o_j in \mathcal{O} . In this work, we assume the predicates Σ and the class of grounding symbols are known and fixed ahead of time but that the object instances \mathcal{O} are not known. Finally, we introduce a symbolic abstraction over the continuous actions that the robot can take. Following Howard et al. (2014b), actions are modeled as a set of symbols that represent the goals or objectives of the robot's motion. For example, the symbol $\text{Grasp}(o)$ represents motions that result in a force closure of an object of interest. Similarly, we introduce other symbolic actions such as picking an object, $\text{Pick}(o)$, or moving an object o to a goal location r , $\text{Move}(o, r)$.

A robot's ability to follow commands is fundamentally tied to its knowledge about the world. The robot's semantic knowledge about the world is typically informed via sensors that are noisy and error-prone. Hence, we introduce a representation to model the robot's belief over semantic knowledge of the world. Let K_t denote the knowledge state that consists of semantic attributes (e.g., "pushable," "movable," and "rigid") associated with individual object instances, and semantic relationships (e.g., "relative strength" and "relative weight") associated with pairs of objects. Let $k_t \in K_t$ represent a single semantic attribute or a relationship. We model the uncertainty over semantic knowledge using a probabilistic belief over the knowledge state $p(K_t)$,

$$p(K_t) = \prod_{k_t \in K} p(k_t = \text{True}) \quad (2)$$

Here, we assume that the distributions over each semantic property are independent. For example, if the workspace contains a "cup" and a "box," the knowledge state K_t is represented as a set of independent binary random variables: $p(\text{IsFull}(\text{cup}) = \text{True})$, $p(\text{IsMovable}(\text{box}) = \text{True})$, etc. In this work, we focus on estimating the aforementioned physical properties of objects (restricted to unary attributes and binary relationships). The robot's belief over semantic world knowledge informs the robot's decision-making and planning process. Next, we formalize the task of interpreting and executing an instruction in the context of acquired knowledge about the world.

2.3. Following instructions under semantic knowledge uncertainty

The robot's goal is to interpret and act according to the human's instruction in the context of its current knowledge about the world. A planning model that reasons over which actions are applicable requires some knowledge about the objects the robot can potentially interact with. Note that

we consider planning domains that may only be partially known. In particular, the robot may lack relevant semantic knowledge that is required for planning manipulation interactions. For example, manipulation tasks may require knowledge of the intrinsic object attributes that cannot be determined from visual observations alone. Consider executing the instruction "clear away the cups on the table," in which empty cups should go in the trash and full cups should be put aside. This task requires knowledge of the internal states of the cups (full or empty) to decide how each cup should be treated. We consider three sources of non-exteroceptive knowledge for "filling in" knowledge about latent aspects of the world model: linguistic communication from the human, direct physical interaction by the robot, and commonsense knowledge corpora.

Formally, the robot is assumed to be primed with a background knowledge corpus \mathbf{B}_0 . The robot receives language utterances from the human $\Lambda_{0:t}$ and acquires interaction measurements $\mathbf{Z}_{0:t}$. At time $t+1$, the robot is provided a language instruction Λ_{t+1} and must synthesize a plan μ_{t+1} in the context of prior observations $\{\Lambda_{0:t}, \mathbf{Z}_{0:t}\}$, the metric world state Y_t , and background knowledge \mathbf{B}_0 . The estimation of the most likely plan $\hat{\mu}_{t+1}$ as per the human's instruction in the context of the world model can be formulated as

$$\hat{\mu}_{t+1} = \underset{\mu_{t+1}}{\operatorname{argmax}} p(\mu_{t+1} | \Lambda_{t+1}, Y_t, \Lambda_{0:t}, \mathbf{Z}_{0:t}, \mathbf{B}_0, \Gamma) \quad (3)$$

Equation (3) involves deriving actions from past linguistic and physical interaction measurements. This inference problem is intractable due to the large space of language and intrinsic force measurements. We introduce the explicit representation of semantic world knowledge K_t at time t that factors the estimation task into more tractable learning tasks:

$$\begin{aligned} & p(\mu_{t+1} | \Lambda_{t+1}, Y_t, \Lambda_{0:t}, \mathbf{Z}_{0:t}, \mathbf{B}_0, \Gamma) \\ &= \int_{K_t} \overbrace{p(\mu_{t+1} | \Lambda_{t+1}, Y_t, K_t, \Gamma)}^{\text{Instruction following}} \overbrace{p(K_t | \Lambda_{0:t}, \mathbf{Z}_{0:t}, \mathbf{B}_0, \Gamma)}^{\text{Knowledge estimation}} dK_t \quad (4) \end{aligned}$$

Here, learning the factor $p(K_t | \Lambda_{0:t}, \mathbf{Z}_{0:t}, \mathbf{B}_0, \Gamma)$ involves acquiring semantic knowledge about the world from observations and background knowledge.

Section 3 presents a probabilistic model of the belief over latent semantic properties informed by observations and prior knowledge. The factor $p(\mu_{t+1} | \Lambda_{t+1}, Y_t, K_t, \Gamma)$ in Equation (4) models plan inference conditioned on the robot's cumulative estimate of its world knowledge. We detail this factor in Section 4 and show how the robot can maintain this distribution over semantic knowledge by actively interacting with the world before synthesizing a plan. Section 6 addresses the task of providing realtime feedback when a discrepancy is observed between the robot's knowledge and the inferred model of the human operator.

3. Bayesian multimodal semantic knowledge estimation

This section addresses the problem of estimating latent semantic attributes associated with objects in the world model from multimodal observations and background knowledge corpora. We first introduce a probabilistic representation of semantic knowledge and then present a Bayesian formulation for incremental online estimation using past language descriptions, direct physical interaction, and background knowledge corpora.

3.1. Probabilistic knowledge

The knowledge state K_t consists of discrete random variables k_t , each modeling a latent object property. We model semantic attributes k_t as Bernoulli random variables with parameter θ_t^k . We introduce a conjugate beta distribution prior with hyper-parameter α_t^k over the Bernoulli distribution parameter θ_t^k as

$$p(k_t) \sim \text{Bernoulli}(\theta_t^k) \quad (5a)$$

$$\theta_t^k \sim \text{Beta}(\alpha_t^k) \quad (5b)$$

The distribution over k_t is parameterized by θ_t^k and, in turn, α_t^k and models the current belief over the true likelihood of a symbolic attribute and consists the shape parameters (a_t^k, b_t^k) characterizing the beta distribution. The likelihood over the semantic attribute variable k_t given the beta distribution parameter α_t^k can be expressed as

$$p(K_t | \Lambda_{0:t}, \mathbf{Z}_{0:t}, \mathbf{B}_0, \Gamma) = \int_{\alpha} \left(\overbrace{p(K_t | \Lambda_t, Z_t, \alpha_{t-1}, \Gamma)}^{\text{Updated knowledge state at time } t} \overbrace{p(\alpha_{t-1} | \Lambda_{0:t-1}, \mathbf{Z}_{0:t-1}, \Gamma, \alpha_0)}^{\text{Belief over knowledge state at time } t-1} \overbrace{p(\alpha_0 | \mathbf{B}_0, \Gamma)}^{\text{Prior from background knowledge}} \right) \quad (8)$$

$$p(k_t | \alpha_t^k) = \int_{\theta} p(k_t | \theta_t^k) p(\theta_t^k | \alpha_t^k) \quad (6)$$

where the beta-distributed random variable θ_t^k is marginalized out.² For a detailed exposition on conjugate distributions, we refer the reader to Bishop (2006).

Our goal is to infer the knowledge state given past observations that arise from language and physical interaction $\{\Lambda_{0:t}, \mathbf{Z}_{0:t}\}$, as well as a priori knowledge from background sources \mathbf{B}_0 , $p(K_t | \Lambda_{0:t}, \mathbf{Z}_{0:t}, \mathbf{B}_0, \Gamma)$. Following the treatment above, we assume that the likelihood over the state K_t is Bernoulli distribution with parameter α_t . We use a Bayesian filter to recursively maintain the knowledge state distribution over time given new observations Z_t ,

$$p(K_t | \Lambda_{0:t}, \mathbf{Z}_{0:t}, \mathbf{B}_0, \Gamma) = \int_{\alpha} \overbrace{p(K_t | \Lambda_t, Z_t, \alpha_{t-1}, \Gamma)}^{\text{Updated knowledge state at time } t} \overbrace{p(\alpha_{t-1} | \Lambda_{0:t-1}, \mathbf{Z}_{0:t-1}, \mathbf{B}_0, \Gamma)}^{\text{Belief over knowledge state at time } t-1} \quad (7)$$

Here, the beta distribution parameter α_{t-1} represents the belief over the knowledge state K_{t-1} at the previous time step $t-1$ as $p(K_{t-1} | \alpha_{t-1})$. This belief is informed by observations $\{\Lambda_{0:t-1}, \mathbf{Z}_{0:t-1}\}$ until time $t-1$ and background knowledge \mathbf{B}_0 . Hence, the factor $p(\alpha_{t-1} | \Lambda_{0:t-1}, \mathbf{Z}_{0:t-1}, \mathbf{B}_0, \Gamma)$ can be viewed as the predictive posterior over the knowledge state at $t-1$, i.e., the belief that integrates past evidence until time $t-1$, before incorporating the current set of observations $\{\Lambda_t, Z_t\}$. The second factor $p(K_t | \Lambda_t, Z_t, \alpha_{t-1}, \Gamma)$ updates the predictive posterior using the current set of observations $\{\Lambda_t, Z_t\}$. The result is the posterior over the knowledge state at time t , which is propagated to the next time step.

Note that the factorization in Equation (7) assumes that, given the prior and current observation, the knowledge state is independent of the previous observations and background knowledge. Formally, the belief over the knowledge state α_{t-1} at the previous time step $t-1$ decouples the estimation of the belief over the next knowledge state K_t from past observations $\Lambda_{0:t-1}, \mathbf{Z}_{0:t-1}$ and the prior knowledge corpus \mathbf{B}_0 given the current set of observations $\{\Lambda_t, Z_t\}$.

Now, we turn our attention to initializing the dynamic Bayesian network at time t_0 . The initial prior over the knowledge state be represented by the beta distribution with parameter α_0 . We assume the presence of a background commonsense corpus \mathbf{B}_0 that informs the initial belief over the knowledge state before the robot acquires any observations. We model this estimation at time t_0 by the factor $p(\alpha_0 | \mathbf{B}_0, \Gamma)$. Introducing the parameter α_0 in Equation (7) leads to the following formulation:

where the parameters α_0 and α_{t-1} are marginalized out. In practice, we approximate Equation (8) with a maximum likelihood estimate over the knowledge prior $\hat{\alpha}_0$:

$$p(K_t | \Lambda_{0:t}, \mathbf{Z}_{0:t}, \mathbf{B}_0, \Gamma) = \int_{\alpha} \overbrace{p(K_t | \Lambda_t, Z_t, \alpha_{t-1}, \Gamma)}^{\text{Knowledge update at time } t} \overbrace{p(\alpha_{t-1} | \Lambda_{0:t-1}, \mathbf{Z}_{0:t-1}, \Gamma, \hat{\alpha}_0)}^{\text{Cumulative belief until time } t-1} \quad (9)$$

Figure 2 illustrates the overall probabilistic model. The remainder of this section is organized as follows. Section 3.2 describes the inference procedure at each step in the temporal model, specifically the updates to the distribution to account for language utterances and direct physical interaction. Section 3.3 addresses the problem of learning an informed prior over semantic knowledge from background commonsense corpora. Finally, Section 3.4 shows how semantic observations from multiple modalities can be fused into a probabilistic belief over world knowledge.

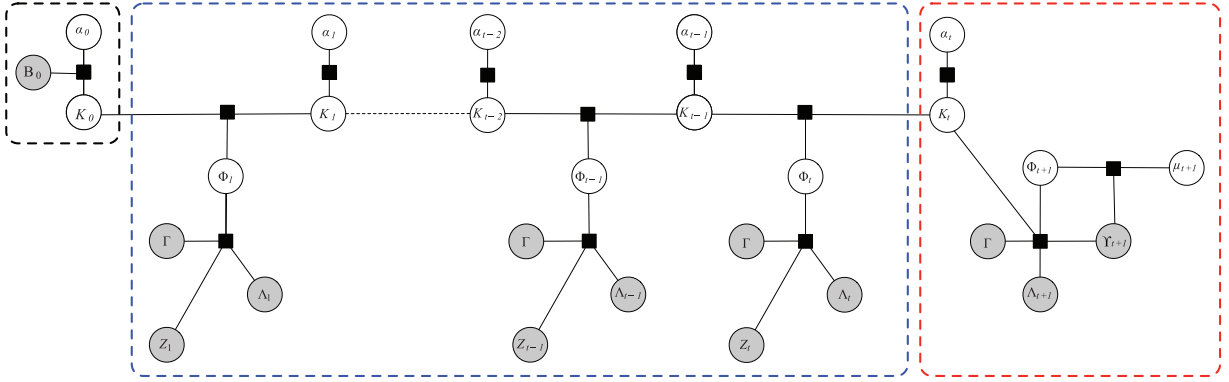


Fig. 2. A probabilistic model for robot command following with learned semantic knowledge about world model entities. The model estimates a belief over the knowledge state K_t from background knowledge B_0 and observations received until time t , which includes language utterances from the human $\Lambda_{0:t}$ and proprioceptive measurements from physical interaction $Z_{0:t}$. This estimation is posed as inference on a dynamic Bayesian network that evolves temporally with novel observations. The learned knowledge is used to follow instructions by generating an appropriate plan of actions. The model consists of three components which are indicated via gray, blue, and red boxes and are described in Sections 3, 4, and 6. Gray: The likelihood over the knowledge state is initialized as parameter α_0 learned from background commonsense knowledge sources B_0 . Blue: At each time step, a correspondence Φ_t is estimated between percepts $\{\Lambda_t, Z_t\}$ and semantic attributes contained in Γ . True correspondences indicate semantic observations that serve as evidence for updating the belief over the latent knowledge state $p(K_t|\alpha_t)$. Red: At time instant $t+1$, the robot interprets an instruction Λ_{t+1} given its current belief over the knowledge K_t state parameterized by α_t . The robot synthesizes a plan μ_{t+1} to accomplish the stated goal state or takes information gathering actions to resolve uncertainty in the semantic state. Here, Y_t denotes the metric world state. Natural language feedback is generated in case discrepancies are observed between the robot's and the human's mental model. The model shown in the illustration evolves from left to right.

3.2. Estimating semantic observations from multimodal percepts

This section details the estimation of the knowledge state K_t at time t expressed in the factor $p(K_t|\Lambda_t, Z_t, \alpha_{t-1}, \Gamma)$ in Equation (8). The knowledge estimate is derived from the input language utterance Λ_t , the physical interaction measurement Z_t , and the cumulative belief over the knowledge state, represented by α_{t-1} , until time $t-1$. This inference involves learning an association between the set of high-level semantic attributes and the language and low-level interaction observations. Learning such an association is challenging as the joint space of multimodal percepts and semantic properties can be combinatorially large. The problem can be factored by first inferring likely semantic attributes from each modality and then fusing the discrete observations into a cumulative belief over the latent knowledge state.

Following earlier work on probabilistic language grounding (Howard et al., 2014a,b; Liang et al., 2013; Paul et al., 2018, 2017; Tellex et al., 2011b), we employ a binary correspondence variable Φ_t that models the association between semantic attributes and the language and interaction measurements. For example, we express the correspondence between the language phrase “the empty cup” and the semantic grounding $\text{IsEmpty}(\text{cup})$ as the conditional likelihood $p(\Phi = \text{True}|\text{IsEmpty}(\text{cup}), \text{the empty cup})$. Fundamentally, this turns the problem of learning the joint distribution between language and percepts into a

discriminative problem of learning true or false associations between language and candidate meanings. This significantly improves the tractability of training and inference.

We extend the use of correspondence variables to associate physical interaction-based observations with the latent semantic object attributes. For example, a slowly increasing force profile while poking a barrel object is indicative of the object being pushable. Alternatively, if the force profile saturates rapidly, the robot can infer that the object is likely to be less pliable during manipulation.

The introduction of the correspondence variable allows us to factorize the distribution over the knowledge state as

$$p(K_t|\Lambda_t, Z_t, \alpha_{t-1}, \Gamma) = \sum_{\Phi_t} \overbrace{p(K_t|\Phi_t, \alpha_{t-1}, \Gamma)}^{\text{Knowledge belief update}} \overbrace{p(\Phi_t|\Lambda_t, Z_t, \Gamma)}^{\text{Language \& interaction groundings}} \quad (10)$$

Here, the factor $p(\Phi_t|\Lambda_t, Z_t, \Gamma)$ models the likelihood of the correspondences between the semantic attributes and percepts Λ_t, Z_t . We use the term *semantic observations* to denote semantic attributes indicated by the most likely set of true correspondence variables. The factor $p(K_t|\Phi_t, \alpha_{t-1}, \Gamma)$ fuses the estimated semantic observations into the belief over the latent semantic attribute.

Note that Equation (10) involves directly fusing observations derived from multiple modalities into a belief over semantic attributes. Learning in the joint space of multiple

modalities is likely to be tractable with a small number of modes. Further, we observe that language descriptions and force interactions arise from independent sources and may arrive at different instances in time. Language descriptions arrive opportunistically from the human, while force interactions are likely to arise from planned and controlled interactions by the robot. Hence, we assume conditional independence between observations arising from different modalities, which enables Equation (10) to be expressed as

$$p(K_t|\Lambda_t, Z_t, \alpha_{t-1}, \Gamma) = \sum_{\Phi_t^\Lambda, \Phi_t^Z} \underbrace{p(K_t|\{\Phi_t^\Lambda, \Phi_t^Z\}, \alpha_{t-1}, \Gamma)}_{\text{Knowledge belief update}} \underbrace{p(\Phi_t^\Lambda|\Lambda_t, \Gamma)}_{\text{Language grounding}} \underbrace{p(\Phi_t^Z|Z_t, \Gamma)}_{\text{Interaction grounding}} \quad (11)$$

where Φ_t^Λ and Φ_t^Z represent correspondence variables derived from language Λ_t and force measurements Z_t , respectively, at the current time step t . Figure 3 presents the corresponding factor graph representation.

Next, we discuss methods for deriving semantic observations from language and physical interactions. We then detail the belief update over the latent object attribute given the inferred semantic observations from each modality.

3.2.1. Estimating groundings from declarative language. We now consider the problem of interpreting factual knowledge about the world present in natural language utterances from the human. As an example, we aim to ground the declarative language utterance “the cup on

the table is empty” to the predicate $\text{IsEmpty}(\text{cup})$, where the “cup” object is located on the table.

The factor $p(\Phi_t^\Lambda|\Lambda_t, \Gamma)$ in Equation (11) models the factual knowledge inherent in declarative language utterances. Inference involves reasoning over the correspondence Φ_t^Λ between a language instruction Λ_t and semantic aspects of world entities modeled as Γ .

We incorporate a contemporary approach to grounding factual knowledge from natural language utterances (Howard et al., 2014b; Paul et al., 2017). The approach exploits the linguistic parse structure of the utterance to factor the grounding problem into separate terms for each constituent phrase. This factorization permits inference over individual phrases rather than joint inference over the entire utterance, improving scalability. For example, the model learns a grounding for the utterance “the nearest cup” as the “cup”-type object nearest to the speaker. We represent the association between individual linguistic elements and semantic concepts using a log-linear model that expresses the likelihood of the linguistic features in each parsed constituent phrase and the corresponding “grounded” attributes of the world model. We train the model using an aligned corpus of utterances and known groundings in the context of a physical world model. The model leverages the inherent compositional structure in language and learns to assign meaning to simpler constituent phrases and structure them together to infer the meaning of an instruction received at runtime (Howard et al., 2014b).

Further, the model uses linguistic structure and part-of-speech information to partition the sentence (Paul et al.,

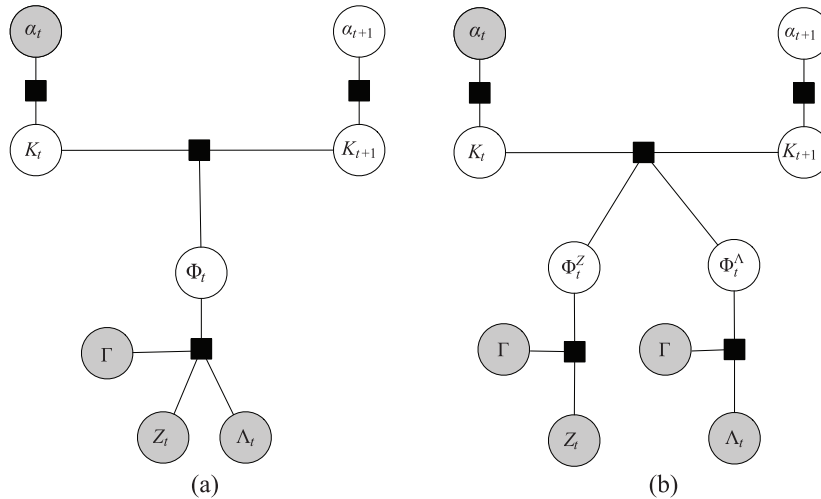


Fig. 3. Probabilistic model for knowledge acquisition over latent object attributes from descriptive language utterances and physical interaction measurements instantiated at each time step t in the dynamic Bayesian network. (a) Joint model. Semantic observations are derived jointly from physical interaction measurement Z_t and factual knowledge from language description Λ_t . The combined factor estimates true correspondences Φ_t between low-level measurements $\{Z_t, \Lambda_t\}$ and high-level semantic properties represented by Γ . The semantic property associated with true correspondences serves as a *semantic* observation. The inferred observation updates the prior belief over the latent knowledge state $p(K_t|\alpha_t)$ to a posterior belief $p(K_{t+1}|\alpha_{t+1})$ propagated to the next time step $t+1$. (b) The factored model assumes independence between semantic observations derived from language description and those derived from physical interaction. Hence, the correspondence variables are factored as Φ_t^Z and Φ_t^Λ associating physical interaction Z_t and language Λ_t with semantic concepts Γ . The estimated groundings from both visual and linguistic modalities are fused to inform a posterior distribution over the latent knowledge state.

2017) into (i) phrases that can be associated with physical aspects of the world (e.g., detected objects and spatial relations) and (ii) phrases that convey facts about the world (e.g., knowledge about the latent state of objects). The inferred factual knowledge conveyed in language provides positive or negative evidence for the underlying knowledge state of the entities described in the utterance. The ability to infer factual knowledge derived from language descriptions is particularly useful if the expressed facts relate to unobserved aspects of the world state. For example, the phrase, “the nearest cup is empty,” conveys information that is otherwise unobservable unless the robot interacts with the cup, i.e., $\text{IsEmpty}(\text{cup})$.

We assume that the user’s utterances convey factual knowledge that they believe to be true according to their internal model of the world. In practice, we store each correspondence Φ_t^A that we infer to be true along with the associated semantic properties Γ (Figure 3(b)) for future reference. This allows the robot to maintain a model of what the human believes to be true of the world and engage in bidirectional communication to correct human beliefs that are inconsistent with evidence that the robot gathers.

The estimation of semantic attributes from the human’s utterance can be viewed as a declarative top-down inference over semantic world knowledge. Next, we address the problem of deriving semantic observations from proprioceptive measurements that arise as the robot physically interacts with the world.

3.2.2. Estimating semantic properties from physical interaction. The estimation of object attributes from physical interaction is an extensively explored area (Bhattacharjee et al., 2013; Chitta et al., 2011; Chu et al., 2015). The ability to infer certain semantic properties of objects from physical interaction helps to determine an appropriate plan in visually unobservable environment. In this work, we perform offline classification of object attributes (e.g., IsFull or IsMovable) given noisy time-series physical interaction measurements during a stereotyped motion with a manipulator, such as lifting or poking. To model the noisy time-series signals, we use a hidden Markov model (HMM) that is a state-based method in which a hidden state is a latent representation of current measurements depending on the previous state. The state transition enables to model or test time-series data with variable length. In particular, we use a multivariate Gaussian HMM³ and model the emission distribution $p(Z_t|s_t)$ as a Gaussian with a full covariance matrix that models the correlation between force and pose measurements (Park et al., 2018).

The factor $p(\Phi_t^Z|Z_t, \Gamma)$ in Equation (11) relates semantic properties (i.e., object attributes) to measurements acquired through physical interaction. Each interaction-based measurement Z_t consists of a sequence of three -axis end-effector force and arm-pose measurements recorded during physical interaction with an object. We identify the

correspondence Φ_t^Z via maximum a posteriori inference. This estimation can be viewed as a bottom-up source of symbolic knowledge derived from grounding raw positional and force measurements.

We use HMMs to define an object attribute estimator f_k that is the predictive model of the factor $p(\Phi_t^Z|Z_t, \Gamma)$ given interaction experience of the semantic attribute k . Let m_{True}^k and m_{False}^k denote the HMM models trained for the True and False of an object attribute k_t . The two HMMs determine the observation likelihoods $p(Z_t|m_{\text{True}}^k)$ and $p(Z_t|m_{\text{False}}^k)$ conditioned for the presence or absence of the object attribute, respectively. The physical interaction measurement acquired online is associated with an object state by comparing the model evidence for the presence or absence of object attributes $k_t \in K_t$ as

$$f_k(Z_t, m_{\text{True}}^k, m_{\text{False}}^k) = p(Z_t|m_{\text{True}}^k)/p(Z_t|m_{\text{False}}^k) \quad (12)$$

We threshold the above likelihood ratio to arrive at a binary classification, and thus Φ_t^Z .

The HMM model m consists of state transition probability, emission probability, and initial state distribution: us from (A, B, π) , where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times d}$, and $\pi \in \mathbb{R}^n$. Let n and d denote the number of hidden states (i.e., 20 or 30 in this work) and the number of modalities (i.e., 2), respectively. In particular, we use a left-to-right HMM that does not allow backward state transitions from a higher-numbered state to a lower-numbered state in A . We also set the first element of initial state distribution π is 1 and other are zero to make the HMM always starts from the first state. These A and π help to model the temporal processes of physical measurements during the stereotyped motions. To use multivariate Gaussian HMMs, we represent the emission probability B as a set of observation mean vector and its covariance matrix. We then train each HMM by iteratively searching its model m^k that maximizes $p(\mathbf{Z}^{\text{tr}}|m^k)$ using an expectation–minimization (EM) algorithm (Rabiner, 1989). Here $\mathbf{Z}^{\text{tr}} \in \mathbb{R}^{n_d \times d \times l}$ is a set of pre-processed interaction traces with varied object states and configurations, where n_d and l are the number of training data (i.e., 20 or 30 per the presence or absence of the object property) and the length of a trace (i.e., 50–200), respectively. The pre-processing step includes smoothing, time-alignment, and scaling. However, after training, the estimation does not require smoothing and time alignment.

3.3. Learning informed priors over semantic knowledge from commonsense corpora

In this section, we focus on the problem of inferring an informed prior over world knowledge derived from a noisy commonsense corpus. In the absence of any background domain knowledge, the initial prior of the model introduced in Section 3.1 can be left uninformed and as more observations and interactions are received, the model gradually converges to the true object attributes. However, estimating the latent object attributes can be hastened if we have an

informed prior that is guided by experience. A source of experience can be found in commonsense corpora derived from human judgement tasks (Forbes and Choi, 2017; Rashkin et al., 2018; Vedantam et al., 2015; Yatskar et al., 2016). Such corpora contain crowdsourced human annotations indicating whether an attribute or a relationship is true for certain object types. For example, human judgements about the relative rigidity of plastic and metal containers would result in relational facts indicating that containers made of plastic are less rigid compared with metal containers.

Learning an informed prior over semantic knowledge K_0 from a commonsense corpus \mathcal{B}_0 at time t_0 can be posed as estimating the conditional distribution $p(K_0|\mathcal{B}_0, \alpha_0, \Gamma)$. The model is initialized at time t_0 with an uninformed beta hyper-prior α_0 .⁴ We treat the factual knowledge present in the commonsense corpus as stochastic observations of the true latent semantic attributes. Following the approach in the previous section, we introduce correspondence variables Φ_0^β that indicate the set of semantic properties associated with a true prior found before robot interaction. The introduction of the correspondence variables allows the conditional likelihood $p(K_0|\mathcal{B}_0, \alpha_0, \Gamma)$ to be expressed as

$$p(K_0|\mathcal{B}_0, \alpha_0, \Gamma) = \sum_{\Phi_0^\beta} \overbrace{p(K_0|\Phi_0^\beta, \alpha_0, \Gamma)}^{\text{Informed knowledge prior}} \overbrace{p(\Phi_0^\beta|\mathcal{B}_0, \Gamma)}^{\text{Facts from corpus}} \quad (13)$$

The factor $p(\Phi_0^\beta|\mathcal{B}_0, \Gamma)$ represents the predictive model that estimates the likelihood that a semantic attribute is true in the world given the evidence in the commonsense corpus. The predicted semantic observations are fused into the latent belief expressed by the factor $p(K_0|\Phi_0^\beta, \alpha_0, \Gamma)$, resulting in the informed prior at the start of the mission. We now discuss the model for predicting semantic properties given a background knowledge corpus and delegate the fusion of the semantic properties into a probabilistic belief to the next section.

Learning the factor $p(\Phi_0^\beta|\mathcal{B}_0, \Gamma)$ involves estimating the correctness of a semantic attribute $k_0 \in K_0$ relating object instances in \mathcal{O} . The problem of predicting attributes between semantic entities has received recent attention in the context of knowledge represented as databases, graphs, or other structured networks (Socher et al., 2013; Wang et al., 2015; Yang et al., 2014; Zhang and Chen, 2018). We adopt a contemporary approach (Yang et al., 2014) and learn a function f_B that models the association between a semantic attribute k_0 and the object types τ associated with object instances in \mathcal{O} . In this work, we restrict ourselves to binary relations and, hence, estimate the function:

$$f_B(\tau(o_i), \tau(o_j), k_0) = \begin{cases} \text{"greater than" score} \\ \text{"less than" score} \end{cases} \quad (14)$$

where $\tau(o_i)$ and $\tau(o_j)$ represent object types for object instances $\{o_i, o_j\} \in \mathcal{O}$. We use the above scores to define the factor $p(\Phi_0^\beta|\mathcal{B}_0, \Gamma)$ by normalizing it.

The aforementioned function f_B is realized using a neural architecture. We first encode the object types using Glove word embeddings (Pennington et al., 2014) that represent semantic or conceptual affinities between words, resulting in the vector embeddings $g_{\tau(o_i)} \in \mathbb{R}^{300}$ and $g_{\tau(o_j)} \in \mathbb{R}^{300}$. We introduce a single-layer feedforward neural network q with rectified linear unit (ReLU) activation functions that outputs task-specific word embeddings $y_{o_i} \in \mathbb{R}^{300}$ and $y_{o_j} \in \mathbb{R}^{300}$: $y_{o_i} = q_w(g_{\tau(o_i)})$ and $y_{o_j} = q_w(g_{\tau(o_j)})$, where w are the parameters of the network. We define a function $f_B(y_{o_i}, y_{o_j}, k_0)$ that models the association between the task-specific vector and the object attribute k_0 under consideration. We explored the following scoring functions to realize the function $f_B(y_{o_i}, y_{o_j}, k_0)$:

- **TransE** (Bordes et al., 2013)

$$- \left(2 \begin{pmatrix} V_k \\ -V_k \end{pmatrix}^T \begin{pmatrix} y_{o_i} \\ y_{o_j} \end{pmatrix} - 2y_{o_i}^T y_{o_j} + \|V_k\|_2^2 \right) \quad (15)$$

- **Bilinear**

$$y_{o_i}^T M_k y_{o_j} \quad (16)$$

- **Bilinear-diag**, same as **Bilinear** with the additional condition that M_k is constrained to be a diagonal matrix.

In the above definitions, V_k and M_k are neural network parameters learned from data. In this work, we use the VerbPhysics dataset (Forbes and Choi, 2017) that contains relative physical knowledge of object pairs encoded as relational tuples, each consisting of relationship and entity attributes. The dataset contains approximately 2500 object pairs annotated with their relative comparisons in terms of "size," "weight," "strength," and "rigidity." The model is trained to predict object attributes (e.g., "size," "weight," "strength," and "rigidity") of types (e.g., "greater than," "less than," "equal," and "unknown"). The training objective minimizes a marking-based ranking loss that encourages the scores of positively expressed semantic relationships to be higher than negatively expressed relationships (Yang et al., 2014).

The learned function provides the prior distribution over knowledge state incorporated in Equation (8). Note that the learned relational model predicts the presence of relative physical properties from abstract object-type data before fusing observations. Online, the model is conditioned on the world model to obtain a distribution over semantic attributes that are relevant for the world model. Next, we turn our attention to the problem of fusing semantic observations derived from multiple modalities into a cumulative belief over latent semantic knowledge.

3.4. Estimating belief over knowledge from multimodal semantic observations

The set of semantic observations of the world state derived from language and physical interaction must be fused into the robot's belief over semantic knowledge. The current observation Φ_t allows the robot to update its previous knowledge estimate parameterized by the beta parameter α_{t-1} to yield the updated belief over K_t . This estimation is represented by the factor $p(K_t|\Phi_t, \alpha_{t-1}, \Gamma)$ in Equation (11). The application of Bayes' rule allows the posterior distribution over K_t to be expressed as

$$\underbrace{p(K_t|\Phi_t, \alpha_{t-1}, \Gamma)}_{\text{Posterior over knowledge}} \propto \underbrace{p(\Phi_t|K_{t-1}, \Gamma)}_{\text{Observation likelihood}} \underbrace{p(K_{t-1}|\alpha_{t-1})}_{\text{Prior}} \quad (17)$$

As the beta distribution serves as a conjugate prior for the Bernoulli likelihood, the posterior distribution over the knowledge state is also beta distributed (Bishop, 2006; Blei et al., 2003). The posterior distribution parameters are obtained using closed-form updates to the prior distribution parameters informed by the current set of observations. A true correspondence variable serves as a positive observation of the associated semantic property and increment to the beta distribution parameter:

$$\begin{aligned} p(K_t|\Phi_t, \alpha_{t-1}, \Gamma) &\sim \text{Beta}(\alpha_t) \\ &\sim \text{Beta}(\alpha_{t-1} + \Phi_t) \end{aligned} \quad (18)$$

Here, the notation $\alpha_{t-1} + \Phi_t$ indicates an update of the beta distribution parameter α_{t-1} with the semantic observation indicated by the correspondence variable Φ_t . Fusing a true observation of a semantic property biases the beta distribution parameters towards favoring a Bernoulli distribution with a higher true belief over the semantic property, and vice versa for a negative observation. Partitioning the set of semantic properties into those derived from language descriptions and those derived from force interactions enables Equation (18) to be factorized as

$$p(K_t|\{\Phi_t^\Lambda, \Phi_t^Z\}, \alpha_{t-1}, \Gamma) \sim \text{Beta}(\alpha_{t-1} + \{\Phi_t^\Lambda + \Phi_t^Z\}) \quad (19)$$

The posterior distribution over the latent knowledge variable evolves incrementally with each observation. The current beta distribution parameters after fusing current observation Λ_{0t}, Z_t with the last estimate α_{t-1} can be expressed as

$$\alpha_t = \{a_t, b_t\} = \{a_{t-1} + (n_\Lambda^1 + n_Z^1), b_{t-1} + (n_\Lambda^0 + n_Z^0)\} \quad (20)$$

Here, $\{n_\Lambda^1, n_\Lambda^0\}$ and $\{n_Z^1, n_Z^0\}$ denote the number of true and false observations derived from language Φ_t^Λ and interaction groundings Φ_t^Z , respectively. Parameters $\{a_t, b_t\}$ constitute the parameter tuple for the beta parameter α_t and $\{a_{t-1}, b_{t-1}\}$ denotes the parameter tuple for the last estimate α_{t-1} . Note that Equation (20) shows that the true and false observations derived from multiple modalities bias the beta distribution parameters appropriately.

Finally, we turn our attention to representing the informed prior belief over K_0 from commonsense corpora initializing the model at time t_0 . Again, leveraging the conjugacy property of the Beta – Bernoulli distributions we can represent the belief as

$$p(K_0|\Phi_0^B, \alpha_0, \Gamma) \sim \text{Beta}(\alpha_0 + \Phi_0^B) \quad (21)$$

Recall, that prior knowledge derived from commonsense corpora serve as noisy observations of the latent semantic knowledge. As indicated in Equation (21), possibly noisy semantic assertions from background knowledge serve as pseudo-measurements and bias the beta distribution parameters before incorporating physical measurements.

Finally, we make a few remarks on the modeling choices in our probabilistic model. The model presented in this section allows the estimation and propagation of the belief over knowledge states derived from multiple and diverse sources. The ability to model uncertainty over latent state and to efficiently fuse multiple modalities provides robustness to noisy and possibly contradictory measurements. Our approach leverages conjugate priors over the likelihood over the correctness of semantic properties in the world model, enabling tractable and efficient posterior updates using observations collected online. The probabilistic formulation can be viewed as a form of *semantic state estimation*. Note that we perform inference over a restricted set of symbolic aspects of the world model. This approach can be considered a special case of more general models that represent beliefs over more complex logical rules (Zettlemoyer et al., 2008). The approach presented is also closely related to *histogram filtering*, which has been employed effectively for robot mapping and tracking applications (Thrun et al., 2005). The measurement updates in a histogram filter require empirically estimating sensor-specific detector rates. On the other hand, the Bayesian approach uses less-prescriptive uninformed priors that are updated with new evidence and is expected to be more robust to noise and erroneous measurements.

4. Instruction-following by introspecting knowledge uncertainty

Recall that our goal is to enable a robot to follow instructions in partially known domains where some object attributes necessary for synthesizing a plan are unobserved. For example, following the instruction “clear the cups on the table” requires knowledge of the internal states of the cups to decide their appropriate destinations in the clearing task (i.e., empty cups should go in the trash and full cups should be put aside). Given the probabilistic model laid out in the previous section, the robot can form a belief over the unobserved semantic properties of the world model by integrating past observations and any available prior domain knowledge. We now consider the task of synthesizing a plan as per the human's command in the context of the acquired knowledge about the world.

Formally, the robot determines a plan μ_{t+1} to satisfy the language instruction Λ_{t+1} received at time $t+1$, taking into account the metric world state Y_{t+1} and the robot’s current world knowledge $p(K_t|\alpha_{t-1})$:

$$p(\mu_{t+1}|\Lambda_{t+1}, Y_{t+1}, \alpha_t, \Gamma) = \int_{K_t} \underbrace{p(K_t|\alpha_t)}_{\text{Current knowledge belief}} \underbrace{p(\mu_{t+1}|\Lambda_{t+1}, Y_{t+1}, K_t, \Gamma)}_{\text{Instruction-following}} \quad (22)$$

The instruction-following task, represented as $p(\mu_{t+1}|\Lambda_{t+1}, Y_{t+1}, \alpha_t, \Gamma)$, can be factored as follows. First, the robot infers the goals or objectives from the natural language command based on its current knowledge about the world. This is followed by reasoning about the sequence of actions resulting in the intended goal state. This factorization allows Equation (22) to be formulated as

$$p(\mu_{t+1}|\Lambda_{t+1}, Y_{t+1}, \alpha_t, \Gamma) = \underbrace{\int_{K_t} \sum_{\Phi_{t+1}^\Lambda} p(K_t|\alpha_t)}_{\text{Action generation}} \underbrace{p(\Phi_{t+1}^\Lambda|\Lambda_{t+1}, Y_{t+1}, \Gamma)}_{\text{Instruction understanding}} \quad (23)$$

Using the maximum likelihood estimates for the knowledge state \hat{K}_t and the grounding for the input instruction Φ_{t+1}^Λ approximates Equation (23) as

$$\hat{\Phi}_{t+1}^\Lambda = \arg\max_{\Phi} p(\Phi_{t+1}^\Lambda|\Lambda_{t+1}, Y_{t+1}, \Gamma) \quad (24a)$$

$$\hat{K}_t = \arg\max_K p(K_t|\alpha_t) \quad (24b)$$

$$\hat{\mu}_{t+1} = \arg\max_{\mu} p(\mu_{t+1}|Y_{t+1}, \hat{\Phi}_{t+1}^\Lambda, \hat{K}_t) \quad (24c)$$

Here, the maximum likelihood estimate indicating the presence of a semantic property \hat{K}_t is obtained by sampling the Bernoulli distribution from the current beta prior $p(K_t|\alpha_t)$. Further, we use a contemporary language interpretation model for estimating intended manipulation goals from an input instruction (Paul et al., 2017) in the context of the robot’s current semantic knowledge. In this work, we use a set of predefined actions such as “clearing,” “packing,” “inspection,” etc. Each action is a sequence of motion primitives including “grasping,” “moving,” “placing,” “pushing,” or “poking,” etc. Each primitive is a sequence of joint values or end-effector poses. We sequence primitives by transforming and scaling each with respect to a goal.

The robot’s action generation takes into account the degree of uncertainty in the robot’s knowledge about the semantic properties of objects relevant to the input instruction. We compute the normalized entropy of the latent belief over semantic properties as a confidence measure for quantifying the robot’s uncertainty over semantic aspects of the world (Grimmett et al., 2016; Paul et al.,

2013; Triebel et al., 2016). The presence of significant uncertainty in the robot’s knowledge belief (as indicated by high entropy of the belief distribution) allows the robot to take information gathering actions such as lifting, pushing, poking, or sliding. The new set of observations are used to update the robot’s belief over the latent object states. The robot continues to interact until the latent belief is sufficiently likely that the robot can execute the final action to complete a task described in the language instruction Λ_{t+1} with high confidence of success. The robot halts plan inference and plan execution when the normalized entropy of the latent belief over semantic properties is lower than an empirically determined threshold. Finally, the estimated high-level plan is handed to a low-level motion planner that generates joint trajectories to achieve an assigned action via the decision-making process.

5. Knowledge-state feedback to the human

Humans working in teams often share world knowledge to help accomplish tasks, such as letting a teammate know that a box is exceptionally heavy. When a teammate observes that the shared knowledge is not true, it is useful to share the corrected information, improving the entire team’s world model. One limitation of the system presented in Arkin et al. (2018) is the lack of a mechanism to provide direct feedback to the human teammate. Providing robots with the capacity to generate linguistic feedback is of particular use for cases in which the robot makes proprioceptive observations during object interaction that contradict world knowledge provided by the human. If we assume that the human teammate only shares world knowledge that they believe is true, then the robot has an opportunity to provide corrective feedback regarding the contradictory observations that should be useful for the human. Such feedback can help the human make better decisions in the future and can help prevent miscommunications owing to incompatible world models.

One approach to providing such feedback via a language interface is to store both the imperative phrase used by the human to reference the object of interest and the declarative phrase used to convey the specific world knowledge. By keeping track of knowledge that was provided by the human (as opposed to other sources of knowledge, e.g., from a commonsense database), the robot can trigger a feedback response upon making a contradictory observation. The linguistic feedback can be composed of the stored imperative and declarative phrases to indicate which object and associated semantic property were different than expected. This approach has the advantage of being computationally inexpensive in that the feedback can be generated by executing a simple lookup for the phrases stored previously. However, this mechanism is brittle to changes in the world that invalidate the stored reference phrase. For example, if the robot has moved close to an object in order to interact with it, what once

may have best been described as “the barrel on the left” may now better be referred to as “the barrel directly in front” or “the nearest barrel.” As such, a declarative phrase such as “the barrel on the left is heavy” might best be corrected with linguistic feedback such as “the barrel nearest to me was not heavy.”

In order to address this brittleness, we pursue an alternative approach by inverting the learned language understanding model to generate phrases associated with the symbolic representation for both the object and hidden semantic state of interest as conditioned on the current spatial configuration of objects in the world. While this does make the feedback robust to changes in the world, it trades off the relatively low computational cost of looking up stored phrases for a significantly higher computational burden of searching over language phrases for one that sufficiently expresses the meaning intended by the symbolic representation. This section details the process for generating linguistic feedback via inverting a language understanding model.

5.1. Communicating knowledge-dissonance to the human

Consider a scenario in which a human teammate says, “the cup on the table is empty.” The robot will ground this declared knowledge and update its belief over the hidden state of the cup’s fullness. Unless the human is intentionally giving false information, the robot can also note that the human’s model of the world includes the confident belief that the cup on the table is empty. Suppose the robot then interacts with the cup and makes an observation indicating that the cup is actually full. In this case, it would be useful for the robot to be able to express this disagreement back to the human, thereby providing a correction to the human’s world model and allowing them to make more informed decisions in the future.

We are interested in a mechanism that facilitates providing this kind of feedback via a natural language interface, namely generating sentences to convey observations that contradict human-provided knowledge. By inverting the learned language understanding model used to ground declarative knowledge, the robot can effectively search for the most likely phrases that map to the set of groundings representing the object of interest and its semantic state. In related work (Tellex et al., 2014), this problem has been referred to as *inverse semantics*. Here, forward semantics refers to the process of taking language and finding associated entities or concepts in the physical world, and while inverse semantics refers to the process of taking aspects of the world and finding language to describe them. The problem formulation and subsequent factorization is inspired by Tellex et al. (2014). The main difference between their approach and what is being done in this work lies in the language understanding model. Tellex et al. (2014) used generalized grounding graphs (Tellex

et al., 2011a) as the underlying language understanding model, whereas the work presented here uses distributed correspondence graphs (Howard et al., 2014b). Using a different underlying language understanding model has important implications for the subsequent model formulation and factorization. The main advantage in this case is the improved runtime performance, the results of which are presented in Howard et al. (2014b).

The problem of inverse semantics for generating feedback can be formulated as search for the most likely sentence corresponding to the intended meaning in the context of the robot’s knowledge about its world. Formally, we estimate a feedback language utterance Λ_{t+1}^{f*} given the known set of groundings Γ , the knowledge state K_t , and metric information about entities in the world Y_t as follows:

$$\Lambda_{t+1}^{f*} = \arg \max_{\Lambda_{t+1}^f \in \Lambda} p(\Lambda_{t+1}^f | K_t, \Gamma, Y_{t+1}) \quad (25)$$

The space of possible sentences Λ is generated via a grammar G that specifies linguistic tokens and production rules for constructing the associated parse tree. This grammar is constructed by scraping the language model’s training corpus for both the tokens and rules. In order to prevent recursive construction of an infinite space of language, the generation process is constrained by the depth of a parse tree.

As we have done for language understanding, we can model this inference process as a correspondence problem wherein the value of a correspondence variable Φ_t^Λ indicates the association between language and a symbol. Because the desired groundings are already known, it is also known which correspondence variables are true. These true correspondences are indicated by Φ_t^Λ , and modify Equation (25) as follows:

$$\Lambda_{t+1}^{f*} = \arg \max_{\Lambda_{t+1}^f \in \Lambda} p(\Phi_{t+1}^\Lambda | \Lambda_{t+1}^f, K_t, \Gamma, Y_{t+1}) \quad (26)$$

In practice, the inverse semantics process is a series of forward semantics evaluations in which the choice of language is an element from Λ . The main concern with this search process is computational cost and, in turn, its impact on the real-time performance of the system. If we could further improve the runtime performance of the forward semantics model, there would necessarily be a corresponding improvement in our inverse semantics implementation. The next section describes a mechanism to effectively bootstrap the language grounding process with solutions computed in advance of an utterance expressed by a human teammate.

6. Improving runtime performance of language understanding and generation

When designing language interfaces, it is important to consider how long the system takes to react or take an action after receiving an utterance from the human. In the

proposed system, the runtime performance of the inference process is the main computational bottleneck that contributes to this latency. Interfaces to robotic systems should aim to achieve real-time responsiveness in order to maintain their effectiveness, as motivated in Section 1 with respect to mission tempo. While the work presented thus far leverages prior research on model approximations for fast inference, language grounding is treated as a reactive process. We propose further addressing this latency problem by precomputing language and grounding solutions for a given environmental context, a process we refer to as *proactive symbol grounding*. By instead proactively inferring the meaning of utterances a human teammate might say (in the context of the current state of the environment), the system has the possibility of receiving a new utterance with the solution already in-hand.

“the trash can.” Once the symbols that correspond to a simple phrase have been found, they can be reused within more complex phrases as long as changes in the environment do not alter their meaning. We leverage the hierarchical and compositional structure of language to construct proactive grounding sets in a bottom-up manner.

Recall that the command-following task can be formulated as Equation (23) defined in Section 4. Interpreting the instruction requires computing the groundings for the full instruction, i.e., for each phrase in the parse tree. A proactive approach precomputes a set of candidate correspondences for likely phrases as denoted as Φ_{t+1}^{psg} . Conditioned on these proactively grounded solutions Φ_{t+1}^{psg} , we reactively only compute correspondences Φ_{t+1}^{new} for novel phrases in the instruction Λ_{t+1} while performing a constant time retrieval for the precomputed solutions. The proactive grounding approach reformulates Equation (23) as

$$p(\mu_{t+1}|\Lambda_{t+1}, Y_{t+1}, \alpha_t, \Gamma) = \int_{K_t} \sum_{\Phi_{t+1}^{\text{new}}} \overbrace{p(K_t|\alpha_t)}^{\text{Knowledge belief}} \overbrace{p(\mu_{t+1}|Y_{t+1}, K_t, \{\Phi_{t+1}^{\text{new}}, \Phi_{t+1}^{\text{psg}}, \Gamma\})}^{\text{Generating actions}} \overbrace{p(\Phi_{t+1}^{\text{new}}|\Lambda_{t+1}, Y_{t+1}, \Phi_{t+1}^{\text{psg}}, \Gamma)}^{\text{Proactive language grounding}} \quad (27)$$

6.1. Proactive symbol grounding for language understanding

In our model, 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, which introduces an interaction latency as previously described, we instead proactively compute groundings for phrases that are likely to be relevant for future instructions. This improves the inference runtime by bootstrapping a novel utterance with estimated groundings (true correspondences) from the set of proactively grounded phrases possessing a similar parse structure. For example, consider the novel instruction “put the empty cup in the trash can.” If the robot has already proactively grounded the constituent phrase “the trash can” for the current state of the world, then the reactive inference process can simply insert the solution for “the trash can” and move on to other phrases in the parse tree.

Formally, the set of proactive correspondences Φ_{t+1}^{psg} is determined as a function of the current environment state Y_{t+1} . The space of possible language utterances is generated via a grammar G that specifies linguistic tokens and production rules and is determined by scraping the rules present within a training corpus. As 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 environment state Y_{t+1} will always ground to the same set of symbols, regardless of parent phrases in the parse tree. Relating back to the example above, “the trash can” maps to the same set of symbols whether it appears in the utterance “put the empty cup in the trash can,” “put the full cup in the trash can,” or even just the simplest form of

Note that the factor $p(\Phi_{t+1}^{\text{new}}|\Lambda_{t+1}, Y_{t+1}, \Phi_{t+1}^{\text{psg}}, \Gamma)$ only estimates the correspondences for new solutions. If we indicate the set of novel phrases in the instruction as Λ_{t+1}^s , then $|\Lambda_{t+1}^s| \leq |\Lambda_{t+1}|$. The model only reactively computes correspondences for novel phrases Φ_{t+1}^{new} , which are fewer than the full set of candidate solutions Φ_{t+1} for the instruction. As a result, the proactive approach leads to runtime improvements in online instruction interpretation.

6.2. Proactive symbol grounding for feedback generation

One of the main limitations of the approach introduced in Section 5 is the runtime performance. Finding the sentence that maximizes the probability of the known set of groundings can be thought of as a series of forward passes through the learned language understanding model. As a result, the time it takes to finish the search process depends on the runtime of each forward pass. Depending on the size of the search space, this can be prohibitively long. Fortunately, the set of proactively grounded phrases Λ^{psg} generated for addressing the latency problem of reactive language understanding can similarly bootstrap this inverse semantics process by effectively providing solutions for a subset of sentences at the cost of a constant-time lookup. As a result, the set of sentences that inverse semantics needs to compute reactively Λ^{new} is now smaller than the full set Λ . The reformulated model from Equation (26) is

$$\Lambda_{t+1}^{f*} = \arg \max_{\Lambda_{t+1}^f \in \Lambda^{\text{new}} \cup \Lambda^{\text{psg}}} p(\Phi_{t+1}^{\Lambda}|\Lambda_{t+1}^f, K_t, \Gamma, Y_{t+1}) \quad (28)$$

By effectively bootstrapping the search over language with a subset of already-grounded sentences, the reactive

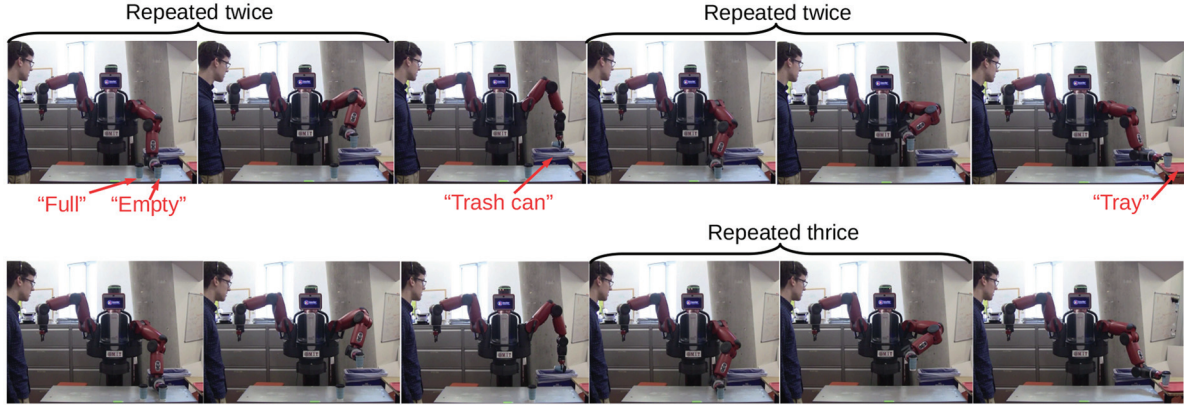


Fig. 4. Experiment evaluating knowledge acquisition over latent object attributes from declarative knowledge and physical interaction. The Baxter robot was instructed to “clear away the cups on the table.” Top: The robot attempts to pick up each cup in turn and infers the latent attribute of the cups from the time series of interactions. Once the belief is sufficiently confident, the robot discards the empty cup in the trash bin and puts 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 that resulted from the incorrectly stated fact. The posterior allows the robot to correctly accomplish the task of clearing in correct locations.

language generation process has fewer computations. In the best case, the proactive language grounding process will have already exhausted Λ and, thus, the search process consists of finding the highest value in a list. In the worst case, Λ^{psg} is empty and inverse semantics is equivalent to Equation (26). We evaluate the runtime performance both with and without the use of proactively grounded phrases and report those results in Section 7.

7. Experiments and results

In order to validate the performance of the proposed system and its components, we designed independent qualitative and quantitative experiments.

7.1. Qualitative evaluation

The first experiment aims to show knowledge acquisition over latent object attributes from declarative knowledge and physical interaction. We used a Baxter Research Robot in a tabletop setup populated with household objects as shown in Figure 4. 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 3.2, the robot also possesses trained HMMs 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 5(a). 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 latent attributes according to force/torque sensing. This knowledge allowed the robot to estimate the correct location to discard the empty cups in the trash and place the filled cups 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 6 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 first 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 $\text{IsFull}(\text{cup}) = \text{True}$. Upon engaging in a time-series of physically interactions with the cup whose hidden attribute is actually $\text{IsFull}(\text{cup}) = \text{False}$, 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 likelihood threshold). The 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 second 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

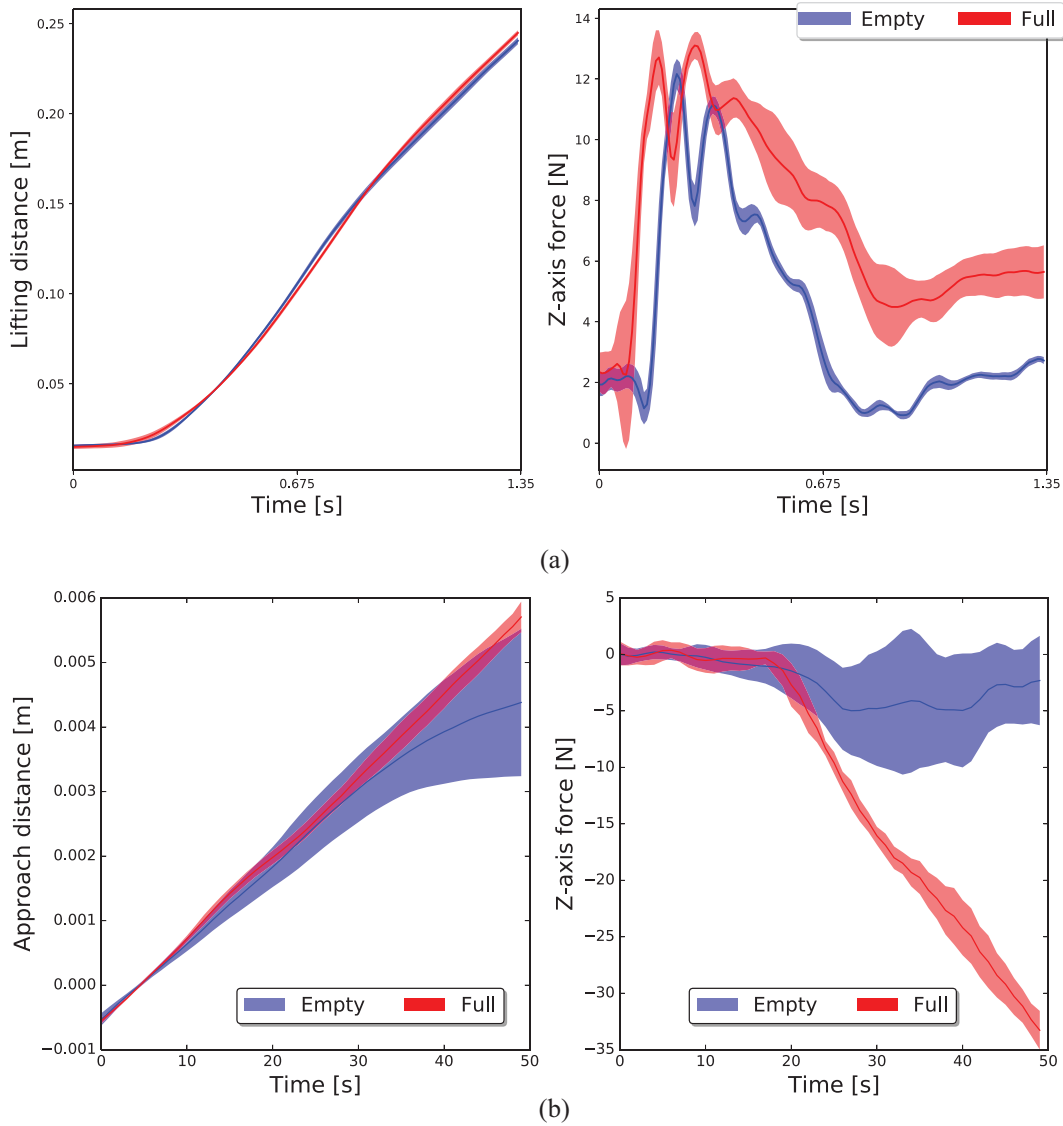


Fig. 5. Distribution of physical interaction time-series measurement during manipulation. (a) Lifting distance and z-axis force measurements over time for both full (red) and empty (blue) cups in Figure 4. (b) Approaching distance and z-axis force measurements over time for both full (red) and empty (blue) barrels in Figure 8. The time-series force measurements for the “full” and “empty” object states. The patterns of force measurements over distances are modeled by two HMMs that are then leveraged during log-likelihood-based binary classification to infer an object’s attribute.

updating beliefs about objects’ attributes. The goal of this qualitative experiment was twofold: (1) to demonstrate a scenario in which faster task completion can be achieved by incorporating human-declared knowledge about the world as compared with relying on physical interaction observations alone, and (2) to demonstrate robust task execution when provided incorrect world knowledge by a human. For this second experiment, we used a Clearpath Husky A200 mounted with a Universal Robots UR5 manipulator in a mobile manipulation setting composed of two Pelican cases, as shown in Figure 7; the internal state of the Pelican cases was hidden. The Pelican case on the robot’s left was full and heavy, and the Pelican 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 describing the state of the two Pelican cases, and (iii) inaccurate declarative knowledge. In 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

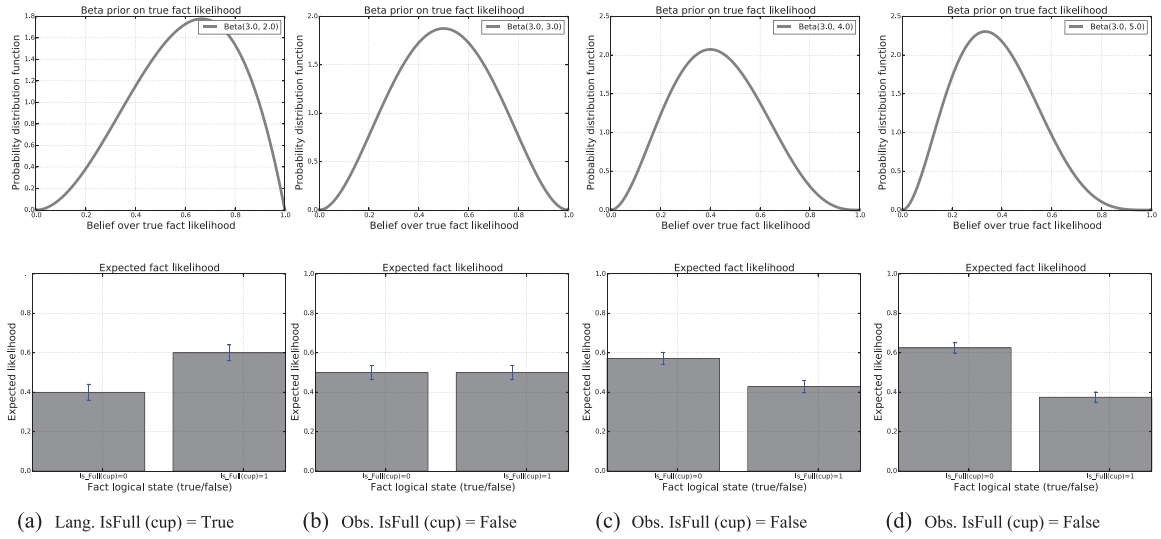


Fig. 6. The temporal evolution of belief over factual knowledge informed by language and interaction. The beta distribution at time t for the Bernoulli likelihood over factual groundings is plotted in the top row. The maximum likelihood for a predicate state appears below. Temporal evaluation from left 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.

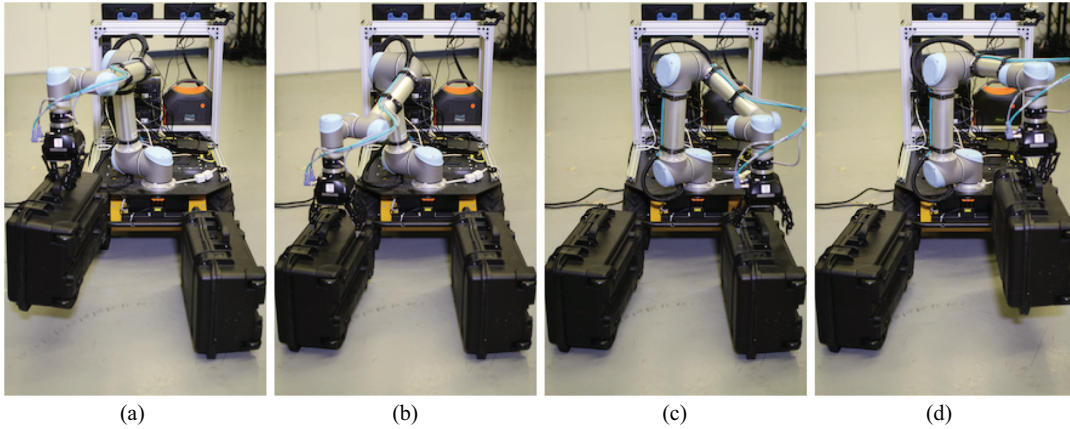


Fig. 7. An experiment incorporating both proactive symbol grounding and updates to beliefs about objects’ attributes via declarative knowledge and force/torque sensing. (a) Initial state of the right case is heavy. (b) Updated belief is uncertain about heavy case. (c) Interaction with the other case. (d) Updated belief that the left case is heavy. 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.”

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.

The third experiment, illustrated in Figure 8, was a part of a field test held in a mock village marketplace at an undisclosed testing facility. Deployed on a separate Husky with a UR5 manipulator, we demonstrated an integrated

system that incorporated both previously evaluated components and declarative knowledge feedback. Similar to previous experiments, we trained an IsFull semantic property estimator from 39 physical interaction data. In the scenario, the robot first localized itself using multimodal sensor fusion with Velodyne LiDAR, inertial measurement unit (IMU), and Intel RealSense camera data. It then constructed the world model by recognizing objects using Mask R-CNN (Massa and Girshick, 2018). Notably, the internal states of the two barrels were unobservable; in actuality, the blue barrel was empty and the other barrel was full. Via a multimodal interface (MMI) described by Barber et al. (2016), a human teammate initially shared

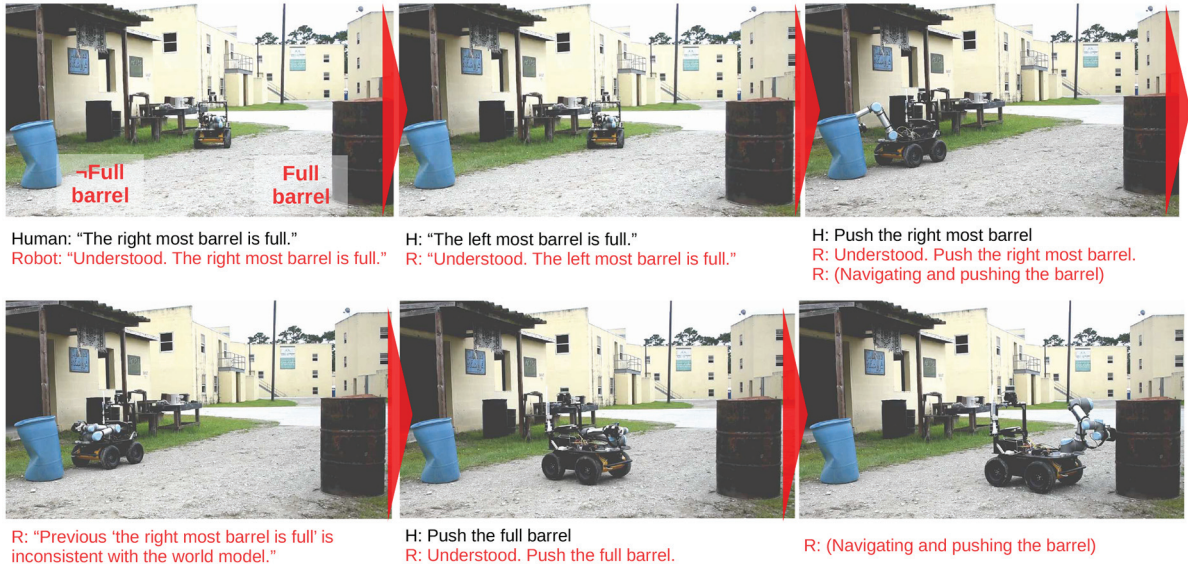


Fig. 8. Experiment demonstrating the declarative knowledge feedback and latent attribute update by declarative language utterance and physical interaction. The Husky robot with a UR5 arm is placed in a outdoor test site filled with doors, windows, barrels, bicycles, among other objects. A user verbally provided wrong and right declarative knowledge for empty and full barrels, respectively. The robot then estimates and reports the latent attribute to the user by pushing each.

their mental model of the objects by stating declaratively that both “the rightmost barrel is full” and “the leftmost barrel is full.” As mentioned, the true state of the rightmost barrel was empty, and thus the human’s shared knowledge contained an inaccuracy. The robot was then instructed to “push the rightmost barrel.” Upon doing so, it updated its belief over the internal states according to observations from force/torque sensing, which were in contradiction to the human’s shared world knowledge. As such, the robot reported back a declarative statement in order to correct the human’s mental model of the barrel. This was done by populating a template with the stored phrase that the human used to initially provide world knowledge about the barrel. With this updated information, the user then instructed the robot to “push the full barrel,” an instruction that previously would have been ambiguous. Owing to the updated shared world model, the robot was able to navigate to and push the barrel on the left as per the user’s instruction.

Videos for all qualitative evaluations are submitted as a multimedia Extensions 1–3.

7.2. Quantitative evaluation

The first statistical evaluation targets the impact of both the PSG component and use of the commonsense knowledge base informed priors on the latency of generating linguistic knowledge-state feedback. In particular, this evaluation seeks to quantify the change in feedback generation time (i.e., from the time the utterance is received to the time a response is generated) as a result of including one or both of these system components. The forward

Table 1. Language generation latency from making a contradictory observation to producing linguistic knowledge-state feedback. The results show the performance with and without both the use of proactive symbol grounding (PSG) and the informed prior. The proactive approach leads to significant reduction in latency in both cases.

	Informed Prior	No Prior
PSG	2.445 ± 2.423 s	0.169 ± 0.003 s
No PSG	94.761 ± 0.806 s	94.834 ± 0.646 s

semantics model was trained on a corpus of 807 annotated examples composed of a variety of symbolic concepts including objects in the world, object categories, physical object properties, spatial relationships, regions, and symbolic actions (see Section 2). By leveraging idle system time while the robot physically interacted with an object, the PSG process was able to precompute the solutions for a subset of 550 different language phrases that could describe the object. When the robot identifies an incorrect fact, it searches over six possible fact templates that are populated using the most likely phrase describing the object of interest, where this phrase is found via the inverse semantics process described in Section 5. The baseline case allowed no time for PSG to run, instead requiring the process to trigger reactively. In the best case, it was able to exhaust the full set of language phrases and provide fast feedback. As can be seen in Table 1, proactive symbol grounding contributed a significant reduction in the latency of feedback generation. Because the use of an informed prior can reduce the number of physical

Table 2. Runtimes showing the impact of incrementally increasing durations of proactive symbol grounding (PSG) on natural language symbol grounding (NLSG) for a single instruction. The leftmost column reports the baseline of NLSG, which is effectively 0 seconds of PSG duration. The proactive approach allows a significant reduction in latency.

PSG duration (s)	—	2.0	4.0	6.0	8.0
Number of grounded phrases	0	31	62	102	146
NLSG inference time (s)	0.21	0.18	0.14	0.13	0.09

interactions necessary for the robot to become sufficiently confident about a contradictory observation, it consequently limits the idle system time that can be used for PSG.

A second statistical evaluation targets the proactive symbol grounding component for natural language symbol grounding in simulation and quantitatively compares the inference runtime to a reactive baseline. This experiment is designed to address the question of how the amount of idle system time impacts the contribution of PSG on improved runtime performance of the inference process. For this experiment, we assumed a sufficiently expressive symbolic representation (Paul et al., 2018), 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 2 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 owing 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 bootstrapping by the proactive grounding module.

Next, we evaluated the accuracy of predicting semantic properties using the model trained from commonsense corpora. We evaluated the performance of three scoring functions that were introduced in Section 3.3. We trained the model using the aforementioned scoring functions with the VerbPhysics dataset containing 2,500 object pairs annotated with relative physical properties. The goal of the classifiers is to predict one of the four classes (greater, less, equal, or unknown) given an object pair and a physical property as input. The corpus was split into training, development and test set in the ratio 80 : 10 : 10. The classifiers were trained to minimize negative log likelihood of the data. We trained for 50 epochs with Adam optimization. The model was tested at the end of each epoch on the

Table 3. A comparison of accuracy (%) in predicting semantic physical properties from commonsense corpus. The table compares the TransE, Bilinear, and Bilinear-diag similarity functions.

Function	Size	Weight	Strength	Rigidity
TransE	92.04	92.77	85.96	83.96
Bilinear	92.96	91.97	87.39	84.07
Bilinear-diag	93.78	93.07	89.8	83.52

development set and that with the best average performance was selected to get the accuracy on the test set. Table 3 shows the performance of the models on the test set. The model based on the *Bilinear-diag* function outperforms other methods.⁶ For the rest of the experiments, we use the *Bilinear-diag* model.

We also quantitatively evaluated the impact of using an informed prior on the accuracy and rate of convergence of the belief to the correct estimate of a semantic property as the robot interacts with an object. We selected six objects for our environment: a box, a basket, a chair, a case, a fridge and a cabinet (see the images in Table 4). For each object, we focus on estimating whether each of the objects is heavy or light for the purposes of manipulation using a UR5 manipulator. The robot interacted with each object 30 times by randomly positioning the object in the manipulation region of the robot, pushing the object with the end effector and recording the force measurements and end-effector pose of the UR5 arm. The resulting physical interaction dataset was randomly permuted resulting in a total of 1,000 different manipulator interaction sequences for each object. Prior probabilities were estimated using the model laid out in Section 3.3 using the *Bilinear-diag* function, which was empirically found to be best performing (see Table 3).

Next, we estimated the heavy/light semantic property using the physical interactions alone and subsequently incorporated the informed priors along with the physical interactions. In each trial, we recorded the number of interaction attempts necessary to infer the property of the object. If we inferred the wrong attribute or we were not able to infer the correct property even after incorporating the entire sequence, the number of attempts was set to 30, the maximum length of the interaction sequence. Figure 9(b) demonstrates that the informed prior enabled faster convergence to the true estimate in comparison to using an uninformed prior represented as no prior (e.g., 0.5 for both $-Heavy$ and $Heavy$). The figure empirically demonstrates the learned priors are informative and hasten convergence to the true latent attribute. Further, the accuracy of predictions at convergence was found to be equivalent for the runs with the informed priors and the uninformed priors (see Figure 9(a)). As an example, inferring the latent attribute for the basket object required at least five interaction tries with an uninformed. The informed prior (i.e., 0.2 for $-Heavy$) decreased the

Table 4. Real objects (6) used in the third experimental evaluation for showing how an informed prior from background knowledge can assist in rapid estimation of latent semantic attributes. We recorded force/torque and end-effector positional information during 180 robot–object interaction sequences.

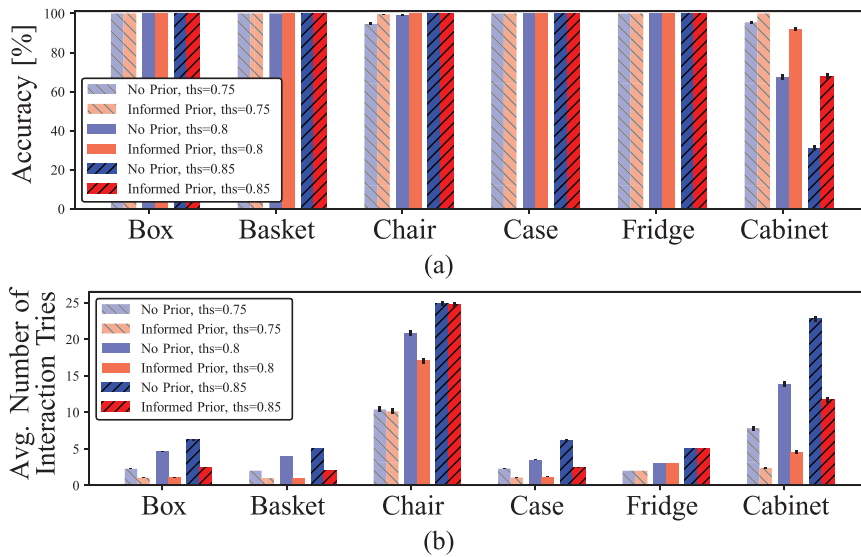
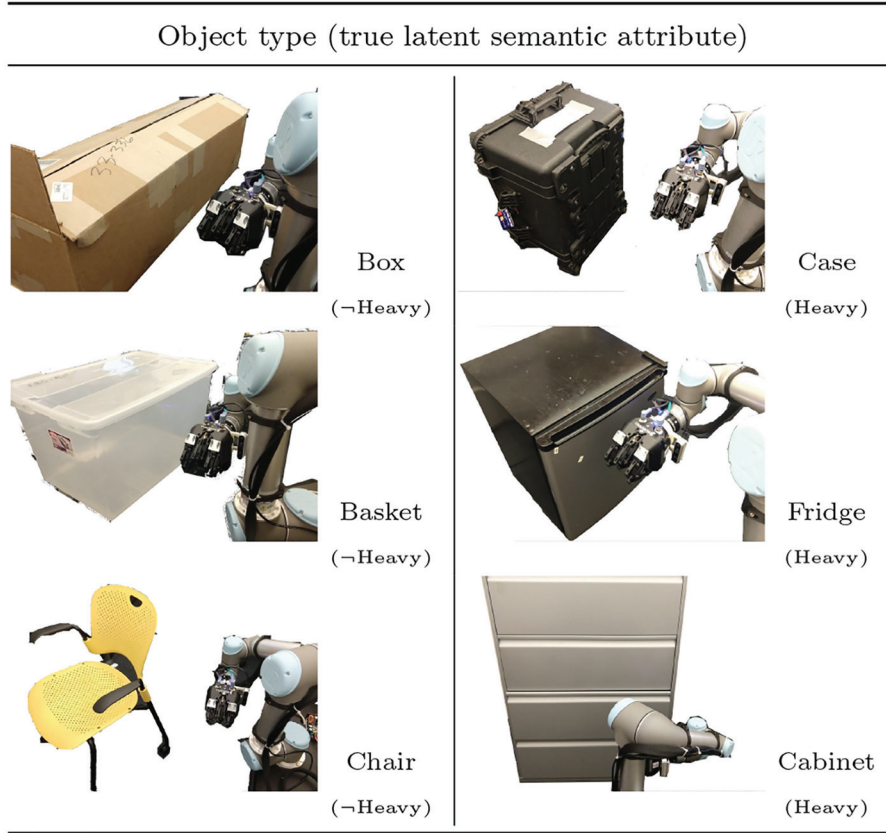


Fig. 9. Comparison of latent-attribute estimation results with or without informed prior over three likelihood thresholds, (0.75, 0.8, 0.85). (a) The accuracy of estimating the latent semantic attribute. Acc. shows the fraction of sequences in which we could infer the correct property for the object. (b) The average number of interaction tries (with standard errors) for estimating the latent semantic attribute. Avg. Tries is the average number of interaction attempts needed to estimate whether an object is heavy or not.

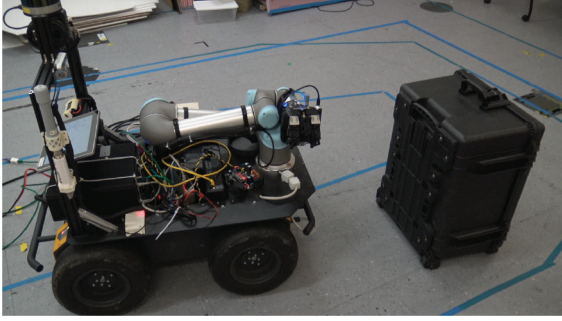


Fig. 10. Semantic latent-attribute estimation experiment. The Husky robot with a UR5 manipulator detects a Pelican case using a RealSense camera and attempts to touch it to infer a latent attribute that is not visually observable.

necessary tries by two interactions without decreasing the accuracy, where the likelihood threshold was 0.85 in this experiment.

Finally, we evaluated a fully integrated system that incorporated previously evaluated components, PSG, and linguistic feedback generation. As shown in Figure 10, we placed a Husky with a UR5 manipulator in a partially observable environment with a “full” semantic attribute of a Pelican case. In the scenario, the robot first recognized the Pelican case by using a RealSense camera mounted on the rear sensor arch. A human operator then provided a declarative fact, “the case is full” or “the case is empty.” Otherwise, the operator did not provide any fact. The robot was then commanded to infer the Pelican case’s latent attribute through physical interactions with or without informed prior. Once the belief over any latent attribute is higher than a threshold (i.e., 0.9) via Bayesian update, the robot reported the inference result. The robot performed five experiments per each scenario (total six scenarios), correctly estimated the true attribute (i.e., “full”), and recorded the number of required physical interactions with belief changes per each. Figure 11 shows both informed prior and correct factual knowledge are helpful to minimize the number of required physical interactions. It shows the Bayesian semantic knowledge estimator successfully propagated the belief over semantic world properties from multiple and diverse sources, and also presents the probabilistic model corrected inaccurate knowledge, “empty” or no prior, online.

Note that the commonsense corpora derived from human annotations might contain erroneous facts resulting in incorrectly informed priors. Either incorrect utterance or incorrectly informed priors may lead to incorrect linguistic feedback, which is not observed in our experiments.

8. Related work

Significant attention has been paid to the problem of endowing robots to interpret natural language instructions. Contemporary statistical approaches to language understanding have been developed that enable robots to follow

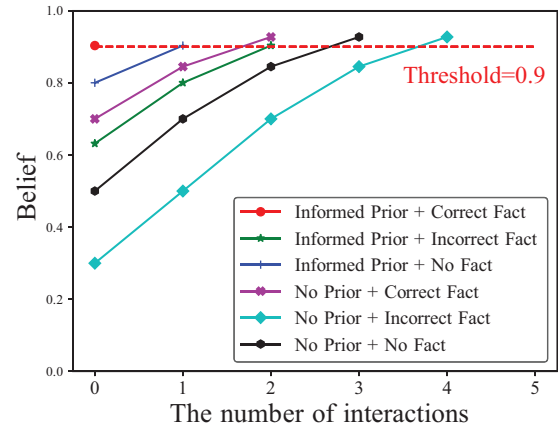


Fig. 11. Comparison of semantic latent-attribute estimation with or without informed prior over declarative knowledge. The Husky robot with a UR5 manipulator attempted to touch a Pelican case (see Figure 10) and infers its latent attribute (i.e., full). Once the belief over any attribute is higher than the 0.9 threshold via Bayesian update, the manipulator finishes the estimation.

complex free-form instructions that involving object manipulation (Misra et al., 2016; Paul et al., 2018; Shridhar and Hsu, 2018; Thomason et al., 2016), navigation (Howard et al., 2014b; Kollar et al., 2010; Matuszek et al., 2010, 2012b; Thomason et al., 2015), and mobile manipulation (Tellex et al., 2011b; Walter et al., 2014a). Such approaches commonly formulate language understanding as a problem of learning a model that associates (i.e., “grounds”) each word in a free-form utterance to its corresponding referent in the robot’s model of its state and action space (Harnad, 1990; Howard et al., 2014a,b; Tellex et al., 2011b). Most existing methods assume that the robot’s model of the environment (the “world model”) is known a priori, typically in the form of a map that expresses the semantic and metric properties of objects and locations necessary to interpret the command.

Instead, we have proposed and evaluated a probabilistic framework that enables robots to exploit multimodal observations, including linguistic, visual, and haptic measurements, to infer latent properties of its environment necessary for human-robot collaboration in partially observed settings. Earlier work in this area includes that of Duvallet et al. (2013), 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 (Anderson et al., 2018; Mei et al., 2016a). More closely related to our framework are methods that leverage metric and semantic information implicit or explicit in the command to learn a distribution over world models that facilitates natural language understanding in a priori unknown environments (Duvallet et al., 2014; Hemachandra et al., 2015; Oh et al., 2016; Walter et al., 2014b). We address a different element of “partial observability” by inferring the state of

object attributes as opposed to hypothesized locations of objects or landmarks that exist beyond the robot’s field of view or its 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.

Meanwhile, recent methods similarly exploit multimodal observations to learn object attributes. Of this body of work, some approaches incorporate human gestures as an input modality to learn object and relation classifiers, as well as attributes such as color (Kollar et al., 2013a; Matuszek et al., 2014; Whitney et al., 2016). Others incorporate audio and haptic measurements as sensing modalities to learn attributes that are not visually observable (Chu et al., 2015), such as whether a container is full or not based on the sounds produced while picking up and shaking (Sinapov and Stoytchev, 2009). Related, some methods directly learn behavior- or sensorimotor-grounded classifications (Hogman et al., 2013), such as the work of Sinapov et al. (2014) that uses vision, proprioception, and audio to learn semantic labels for objects while the robot interacts with them.

Relevant to the goals of this work are methods that consider the problem of understanding instructions that are ambiguous in the context of the robot’s model of its state and action space. Among these methods are those that employ inverse groundings (Gong and Zhang, 2018; Tellex et al., 2014) as a means of asking targeted questions that are believed to be most informative in an estimation-theoretic sense (Tellex et al., 2012). Related, a number of techniques have been proposed to learn a priori unknown grounding models by exploring models that relate novel linguistic predicates to the robot’s world model or directly to its percepts (She and Chai, 2017; Thomason et al., 2018, 2016; Tucker et al., 2017). Our work differs in that we assume that the concepts are known, but that the instantiation of these concepts in the robot’s environment are unknown.

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 (Kollar et al., 2013b; Matuszek et al., 2012a; Paul et al., 2017; Perera and Allen, 2013; Thomason et al., 2016) assume 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.

In the event that the robot identifies discrepancies between the declared knowledge and its observation of the environment, our framework conveys this disagreement to the user via generated language. Our approach is related to recent work on inverse symbol grounding (Tellex et al., 2014), which is typically considered in the context of engaging the user in dialog to resolve ambiguities in the

task (Deits et al., 2013; Hemachandra and Walter, 2015; Raman et al., 2013; Tellex et al., 2012). With this approach, we invert our learned language understanding model to identify the set of phrases that are most likely to correspond to the particular properties of the environment of interest. Unlike Tellex et al. (2014), who used generalized grounding graphs (Tellex et al., 2011b), we use the distributed correspondence graph language model (Howard et al., 2014a), which affords more efficient inference. We also identify phrases by explicitly optimizing over their likelihood rather than maximizing over a scoring function.

Highly relevant is work on referring expression generation, which is concerned with producing a textual description that allows a human to correctly identify a target object or other entity that is known only to the generator. In the computer vision and natural language processing communities, the task typically involves conveying information about objects or locations within an image (Kazemzadeh et al., 2014; Luo and Shakhnarovich, 2017; Mao et al., 2016; Yu et al., 2016). Contemporary approaches to this problem employ neural network architectures for language generation, and thus require access to large datasets for training, which are typically not available for robotics or other embodied domains. In robotic applications, referring expression problems often involve reasoning over spatially extended 3D environments (e.g., at the room, floor, or building level). Consequently, generation algorithms (Fang et al., 2015; Kelleher and Kruijff, 2006; Zender et al., 2009) must provide enough information for the listener, whose context will often be limited.

Related, other researchers have endowed robots with language generation capabilities as a means of conveying task information to their human partners (Andrist et al., 2013; Dzindolet et al., 2003; Wang et al., 2016). Among these are methods that consider the problem of producing free-form instructions that allow humans to perform a task, such as navigation (Curry et al., 2015; Goeddel and Olson, 2012; Oswald et al., 2014). Traditionally, solutions to this problem have relied upon hand-crafted rules that are designed to mimic the way in which humans generate instructions (e.g., via a set of composition rules and language templates). Much like language understanding, recent work employs statistical and learned models (Cuayáhuitl et al., 2010; Daniele et al., 2017b; Oswald et al., 2014) that can be trained from natural language corpora, and are thus able to produce utterances that are easier for people to follow.

Significant effort in the natural language processing community has focused on the problem of generation. This includes work on selective generation, which considers the problem of producing a natural language utterance that effectively expresses the content of a rich database. Selective generation has traditionally been formulated as two separate problems: content selection (Barzilay and Lapata, 2005; Barzilay and Lee, 2004), which reasons over *what* to talk about, and surface realization (Liang

et al., 2009; Walker et al., 2001), which decides *how* to convey the selected content via natural language. Relevant to our inverse semantics approach, Wong and Mooney (2007) effectively inverted a semantic parser to generate natural language text from formal meaning representations using synchronous context-free grammars.

Recent work performs selective generation via a single framework (Angeli et al., 2010; Chen and Mooney, 2008; Kim and Mooney, 2010; Konstas and Lapata, 2012; Mei et al., 2016b), rather than treating it as two separate sub-problems. Angeli et al. (2010) formulate content selection and surface realization as local decision problems via log-linear models, and employ templates for generation. Mei et al. (2016b) proposed a recurrent neural network encoder-aligner-decoder model that jointly learns to perform content selection and surface realization from database-text pairs, thereby treating the selective generation as an end-to-end learning problem.

9. Discussion and conclusion

We have introduced a probabilistic model for inferring the latent semantic properties of the world to correctly follow high-level human instructions in partially observable environments. We have 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 required for task execution. Further, we have leveraged background commonsense knowledge corpora to learn an informed prior when initializing the model for efficient subsequent inference.

We have also presented an approach for generating linguistic feedback to the human in the case where discrepancies are observed between the robot's and the human's semantic knowledge about the world. Finally, we have addressed the issue of reducing latency in both instruction interpretation and feedback generation that stems from the computation complexity of associating language with semantic entities in the world. We have introduced a proactive grounding approach that predicts future utterances and selectively computes candidate interpretations from incremental observations of the world. We have demonstrated the approach on fixed and mobile manipulators executing high-level tasks by "filling in" semantic knowledge about world entities from both declarative knowledge sources as well as physical interactions.

The experiments in this work contribute towards bridging the gap between higher-order inputs such as language from the human and low-level representation such as interaction forces for the robot via grounded learning of semantic concepts by fusing acquired semantic knowledge. The experimental evaluations on multiple platforms and the field deployment test contribute toward validating the reproducibility and robustness in the presence of uncertain environment conditions. Further, the ability to

provide online linguistic feedback for resolving differences in the robot's and the human's mental models contributes to addressing the transparency and op-tempo communication requirements of real-world human-robot teaming scenarios.

There are several avenues for future work that emerge from the current investigation. Our current approach for deciding and taking information gathering actions is myopic because we only utilize a one-step look ahead. The decision is also based on the entropy of the underlying distribution but does not explicitly compute the information gain associated with actions. There is scope to integrate multistep planning to gain information about uncertain semantic properties. Further, we considered semantic attributes associated with an object to be independent while fusing knowledge from multiple sources. Often, physical properties are correlated. For example, heavy objects are often difficult to slide. Hence, future work will explore Bayesian priors that preserve correlations. There is scope to leveraging similar work in *correlated topics modeling* (Blei and Lafferty, 2006).⁷ Similarly, there is scope to use a correlated measurement model that accounts for correlated observations. For example, observing items such as cups and tables are highly predictive of the presence of humans in a building.

The current model assumes that the space of semantic concepts is fixed a priori, thereby making the overall system less robust for situations in which the plan execution requires knowledge of a novel semantic property that was not seen during training. This limitation can be addressed in two ways. First, we can incorporate non-parametric Bayesian models that expand with data complexity (Blei and Jordan, 2006). Second, we may explore ways to detect the presence of a new concept and acquire new recognition models online with limited interaction (Tucker et al., 2017), thereby allowing our model to grow its space of semantic concepts in an online fashion. Our experiments so far have focused on the robot interacting with the world to improve its understanding. There is further scope to acquire semantic knowledge by observing the behavior of other agents, either during an intentional demonstration or via happenstance while executing a collaborative task. As an example, if the robot observes a person struggle to lift a box, it can incorporate that observation as evidence about the box's heaviness.

The present formulation incorporated binary predicate symbols to represent symbolic states. The model can be extended in case of ternary or multi-ary properties as well by incorporating a multi-dimensional conjugate distribution. For example, we can extend the Beta-Bernoulli prior to a Dirichlet-multinomial prior to incorporate multi-ary properties.

This work has explored the use of natural language to inform the latent properties of objects in the robot's world model that were corroborated or corrected by the haptic modality. However, unlike touch, language utterances are often ambiguous and may only implicitly communicate information. For example, the a language instruction may

be ambiguous in terms of which objects are referenced. Consider the utterance, “the barrel on the left is empty” when there are two barrels on the left side of the robot. Such ambiguity can be addressed by engaging in dialog with the operator. The natural language generation system presented in this work can be extended and used to generate disambiguation queries to resolve the ambiguity. Now, we turn our attention to the problem of implicit knowledge that we did not consider in this work. Consider the scenario where the operator informs the robot that “all the oil in the barrel was removed today.” Common sense reasoning informs us that the barrel is now empty. However, the presented system would not use such knowledge as it cannot reason about implicit knowledge. The problem can be addressed by incorporating (learning) common sense knowledge and performing a form of logical inference or logical state estimation to determine the implicit states from the explicitly stated knowledge. Exploration in this direction remains part of future extensions.

Further, the current model assumes that the linguistic, haptic, and knowledge-based priors are equally weighted. In practical contexts, one modality may be more informative than others. Learning per-modality sensor models and context-specific weightings remains part of future work.

Finally, we seek to expand the scope of language feedback to also include explanations (Parkash and Parikh, 2012; Selvaraju et al., 2017). We envision that the robot should be able to communicate not only that a piece of factual knowledge is incorrect, but describe how it arrived at such a conclusion, for example, by interacting with it. We intend to explore richer multimodal communication as part of future research.


Acknowledgement


We thank Michael Noseworthy for valuable feedback on this manuscript.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the Robotics Consortium of the US Army Research Laboratory under the Collaborative Technology Alliance Program (RCTA; ARO grant W911NF-15-1-0402), the Toyota Research Institute (TRI; award number LP-C000765-SR), and Lockheed Martin Co.

ORCID iDs

Jacob Arkin  <https://orcid.org/0000-0002-1074-9248>

Daehyung Park  <https://orcid.org/0000-0002-1287-9433>

Supplemental material

Supplemental material for this article is available online.

Notes

1. Our technical exposition uses two time scales. The subscript t denotes the time scale at which the language utterances,

visual observations, and interaction measurements are used to update the robot’s world model. We assume a finer discretization of this update time instant t into n time steps t_0, \dots, t_n in which the robot executes a low-level motion plan and receives measurements that are collectively used to update knowledge about the world.

2. Note that the beta distribution models the distribution over the true likelihood of the Bernoulli distribution. Each sample from the beta distribution forms a histogram over truth value of a semantic property.
3. We implement the HMMs using the general hidden Markov model library (GHMM) (Schliep et al., 2004).
4. We used $\alpha_0 = (2, 2)$ to initialize a symmetric beta distribution acting as a diffuse uninformed prior over K_0 at model initialization at time t_0 .
5. Note that the same language understanding model (Paul et al., 2017) was used in Section 3 to infer declarative facts from language utterances. In this section, we use the model to infer grounded actions based on knowledge acquired from past observations and prior knowledge.
6. The higher performance of the Bilinear-diag similarity function corroborates findings by Yang et al. (2014) in link prediction tasks.
7. Correlated topic models (Blei and Lafferty, 2006) use a logistic normal prior instead of Dirichlet priors to model correlations between discrete word expression.

References

- Anderson P, Wu Q, Teney D, et al. (2018) Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3674–3683.
- Andrist S, Spannan E and Mutlu B (2013) Rhetorical robots: Making robots more effective speakers using linguistic cues of expertise. In: *Proceedings ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Tokyo, Japan, pp. 341–348.
- Angeli G, Liang P and Klein D (2010) A simple domain-independent probabilistic approach to generation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 502–512.
- Arkin J and Howard TM (2018) 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)*.
- Arkin J, Paul R, Park D, Roy S, Roy N and Howard T (2018) Real-time human–robot communication for manipulation tasks in partially observed environments. In: *Proceedings of the International Symposium on Experimental Robotics (ISER)*.
- Barber DJ, Howard TM and Walter MR (2016) A multimodal interface for real-time soldier-robot teaming. In: *SPIE Defense + Security*. 98370M.
- Barzilay R and Lapata M (2005) Collective content selection for concept-to-text generation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 331–338.
- Barzilay R and Lee L (2004) Catching the drift: Probabilistic content models, with applications to generation and summarization. *arXiv preprint arXiv:0405039*.
- Bhattacharjee T, Kapusta A, Rehag JM and Kemp CC (2013) Rapid categorization of object properties from incidental

- contact with a tactile sensing robot arm. In: *Proceedings IEEE-RAS International Conference on Humanoid Robots (Humanoids)*. IEEE, pp. 219–226.
- Bishop CM (2006) Probability distributions. In: Jordan M, Kleinberg J and Schölkopf B (eds.) *Pattern Recognition and Machine Learning*. Berlin: Springer-Verlag.
- Blei D and Lafferty J (2006) Correlated topic models. In: *Advances in Neural Information Processing Systems*, Vol. 18, p. 147.
- Blei DM and Jordan MI (2006) Variational inference for Dirichlet process mixtures. *Bayesian Analysis* 1(1): 121–143.
- Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.
- Bordes A, Usunier N, Garcia-Duran A, Weston J and Yakhnenko O (2013) Translating embeddings for modeling multi-relational data. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2787–2795.
- Chen DL and Mooney RJ (2008) Learning to sportscast: A test of grounded language acquisition. In: *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 128–135.
- Chitta S, Sturm J, Piccoli M and Burgard W (2011) Tactile sensing for mobile manipulation. *Transactions on Robotics* 27(3): 558–568.
- Chu V, McMahon I, Riano L, et al. (2015) Robotic learning of haptic adjectives through physical interaction. *Robotics and Autonomous Systems* 63: 279–292.
- Cuayáhuil H, Dethlefs N, Frommberger L, Richter KF and Bateman J (2010) Generating adaptive route instructions using hierarchical reinforcement learning. In: *Proceedings of the International Conference on Spatial Cognition*, pp. 319–334.
- Curry AC, Gkatzia D and Rieser V (2015) Generating and evaluating landmark-based navigation instructions in virtual environments. In: *Proceedings of the European Workshop on Natural Language Generation (ENLG)*, Brighton, UK, pp. 90–94.
- Daniele A, Howard TM and Walter MR (2017a) A multiview approach to learning articulated motion models. In: *International Symposium on Robotics Research*.
- Daniele AF, Bansal M and Walter MR (2017b) Navigational instruction generation as inverse reinforcement learning with neural machine translation. In: *Proceedings ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, Vienna, Austria.
- Deits R, Tellex S, Thaker P, Simeonov D, Kollar T and Roy N (2013) Clarifying commands with information-theoretic human–robot dialog. *Journal of Human–Robot Interaction* 2(2): 58–79.
- Duvallet F, Kollar T and Stentz A (2013) Imitation learning for natural language direction following through unknown environments. In: *Proceedings IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1047–1053.
- Duvallet F, Walter MR, Howard T, et al. (2014) Inferring maps and behaviors from natural language instructions. In: *Proceedings of the International Symposium on Experimental Robotics (ISER)*, Marrakech/Essaouira, Morocco.
- Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG and Beck HP (2003) The role of trust in automation reliance. *International Journal of Human–Computer Studies* 58(6): 697–718.
- Fang R, Doering M and Chai JY (2015) Embodied collaborative referring expression generation in situated human–robot interaction. In: *Proceedings ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, pp. 271–278.
- Forbes M and Choi Y (2017) Verb physics: Relative physical knowledge of actions and objects. In: *Proceedings of the Association for Computational Linguistics (ACL)*.
- Goeddel R and Olson E (2012) DART: A particle-based method for generating easy-to-follow directions. In: *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1213–1219.
- Gong Z and Zhang Y (2018) Temporal spatial inverse semantics for robots communicating with humans. In: *Proceedings IEEE International Conference on Robotics and Automation (ICRA)*.
- Grimmett H, Triebel R, Paul R and Posner I (2016) Introspective classification for robot perception. *The International Journal of Robotics Research* 35(7): 743–762.
- Harnad S (1990) The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42(1): 335–346.
- Hemachandra S, Duvallet F, Howard TM, Roy N, Stentz A and Walter MR (2015) Learning models for following natural language directions in unknown environments. In: *Proceedings IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, WA.
- Hemachandra S and Walter M (2015) Information-theoretic dialog to improve spatial-semantic representations. In: *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Hamburg, Germany.
- Hogman V, Bjorkman M and Kragic D (2013) Interactive object classification using sensorimotor contingencies. In: *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2799–2805.
- Howard TM, Chung I, Propp O, Walter MR and Roy N (2014a) Efficient natural language interfaces for assistive robots. In: *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Workshop on Rehabilitation and Assistive Robotics*.
- Howard TM, Tellex S and Roy N (2014b) A natural language planner interface for mobile manipulators. In: *Proceedings IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 6652–6659.
- Kazemzadeh S, Ordonez V, Matten M and Berg T (2014) Referringgame: Referring to objects in photographs of natural scenes. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798.
- Kelleher JD and Kruijff GJM (2006) Incremental generation of spatial referring expressions in situated dialog. In: *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1041–1048.
- Kim J and Mooney RJ (2010) Generative alignment and semantic parsing for learning from ambiguous supervision. In: *Proceedings of the International Conference on Computational Linguistics*, pp. 543–551.
- Kollar T, Krishnamurthy J and Strimel GP (2013a) Toward interactive grounded language acquisition. In: *Proceedings Robotics: Science and Systems (RSS)*, pp. 721–732.
- Kollar T, Perera V, Nardi D and Veloso M (2013b) Learning environmental knowledge from task-based human–robot dialog. In: *Proceedings IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4304–4309.
- Kollar T, Tellex S, Roy D and Roy N (2010) Toward understanding natural language directions. In: *Proceedings ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, pp. 259–266.
- Konstas I and Lapata M (2012) Unsupervised concept-to-text generation with hypergraphs. In: *Proceedings of the*

- Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 752–761.
- Liang P, Jordan MI and Klein D (2009) Learning semantic correspondences with less supervision. In: *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 91–99.
- Liang P, Jordan MI and Klein D (2013) Learning dependency-based compositional semantics. *Computational Linguistics* 39(2): 389–446.
- Luo R and Shakhnarovich G (2017) Comprehension-guided referring expressions. In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7102–7111.
- Mao J, Huang J, Toshev A, Camburu O, Yuille AL and Murphy K (2016) Generation and comprehension of unambiguous object descriptions. In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20.
- Massa F and Girshick R (2018) Mask R-CNN-benchmark: Fast, modular reference implementation of instance segmentation and object detection algorithms in PyTorch. Available at: <https://github.com/facebookresearch/maskrcnn-benchmark> (accessed 16 July 2019).
- Matuszek C, Bo L, Zettlemoyer L and Fox D (2014) Learning from unscripted deictic gesture and language for human–robot interactions. In: *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.
- Matuszek C, FitzGerald N, Zettlemoyer L, Bo L and Fox D (2012a) A joint model of language and perception for grounded attribute learning. In: *Proceedings of the International Conference on Machine Learning (ICML)*.
- Matuszek C, Fox D and Koscher K (2010) Following directions using statistical machine translation. In: *Proceedings ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, pp. 251–258.
- Matuszek C, Herbst E, Zettlemoyer L and Fox D (2012b) Learning to parse natural language to a robot execution system. Technical Report UW-CSE-12-01-01, University of Washington.
- Mei H, Bansal M and Walter MR (2016a) Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In: *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Phoenix, AZ.
- Mei H, Bansal M and Walter MR (2016b) What to talk about and how? Selective generation using LSTMS with coarse-to-fine alignment. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Misra DK, Sung J, Lee K and Saxena A (2016) Tell me Dave: Context-sensitive grounding of natural language to manipulation instructions. *The International Journal of Robotics Research* 35: 281–300.
- Oh J, Howard TM, Walter MR, et al. (2016) Integrated intelligence for human–robot teams. In: *Proceedings of the International Symposium on Experimental Robotics (ISER)*.
- Oswald S, Kretschmar H, Burgard W and Stachniss C (2014) Learning to give route directions from human demonstrations. In: *Proceedings IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3303–3308.
- Park D, Kim H and Kemp CC (2018) Multimodal anomaly detection for assistive robots. *Autonomous Robots* 43(3): 611–629.
- Parkash A and Parikh D (2012) Attributes for classifier feedback. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Berlin: Springer, pp. 354–368.
- Paul R, Arkin J, Aksaray D, Roy N and Howard TM (2018) Efficient grounding of abstract spatial concepts for natural language interaction with robot platforms. *The International Journal of Robotics Research* 37(10): 1269–1299.
- Paul R, Barbu A, Felshin S, Katz B and Roy N (2017) Temporal grounding graphs for language understanding with accrued visual-linguistic context. In: *Proceedings International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4506–4514.
- Paul R, Grimmett H, Triebel R and Posner I (2013) Introspective classification for mission-critical decision making. In: *RLDM 2013*, p. 63.
- Pennington J, Socher R and Manning C (2014) GloVe: Global vectors for word representation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.
- Perera IE and Allen JF (2013) Sall-e: Situated agent for language learning. In: *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, p. 1241–1247.
- Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2): 257–286.
- Raman V, Lignos C, Finucane C, Lee KCT, Marcus M and Kress-Gazit H (2013) Sorry Dave, I’m afraid I can’t do that: Explaining unachievable robot tasks using natural language. In: *Proceedings Robotics: Science and Systems (RSS)*, Berlin, Germany.
- Rashkin H, Sap M, Allaway E, Smith NA and Choi Y (2018) Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*.
- Russell SJ and Norvig P (2016) *Artificial Intelligence: A Modern Approach*. New York: Pearson Education Limited.
- Schliep A, Rungtarityotin W and Georgi B (2004) General hidden Markov model library. Available at <http://www.ghmm.org/> (accessed March 2020).
- Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D and Batra D (2017) GRAD-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 618–626.
- She L and Chai J (2017) Interactive learning of grounded verb semantics towards human–robot communication. In: *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1634–1644.
- Shridhar M and Hsu D (2018) Interactive visual grounding of referring expressions for human–robot interaction. *arXiv preprint arXiv:1806.03831*.
- Sinapov J, Schenck C, Staley K, Sukhoy V and Stoytchev A (2014) Grounding semantic categories in behavioral interactions: Experiments with 100 objects. *Robotics and Autonomous Systems* 62(5): 632–645.
- Sinapov J and Stoytchev A (2009) From acoustic object recognition to object categorization by a humanoid robot. In: *Proceedings of the RSS 2009 Workshop on Mobile Manipulation*.
- Socher R, Chen D, Manning CD and Ng A (2013) Reasoning with neural tensor networks for knowledge base completion. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 926–934.
- Tellex S, Knepper R, Li A, Rus D and Roy N (2014) Asking for help using inverse semantics. In: *Proceedings Robotics: Science and Systems (RSS)*.

- Tellex S, Kollar T, Dickerson S, et al. (2011a) Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine* 32(4): 64–76.
- Tellex S, Kollar T, Dickerson S, et al. (2011b) Understanding natural language commands for robotic navigation and mobile manipulation. In: *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.
- Tellex S, Thaker P, Deits R, Simeonov D, Kollar T and Roy N (2012) Toward information theoretic human–robot dialog. In: *Proceedings Robotics: Science and Systems (RSS)*, Sydney, Australia.
- Thomason J, Sinapov J, Mooney RJ and Stone P (2018) Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In: *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.
- Thomason J, Sinapov J, Svetlik M, Stone P and Mooney RJ (2016) Learning multi-modal grounded linguistic semantics by playing “I Spy”. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3477–3483.
- Thomason J, Zhang S, Mooney RJ and Stone P (2015) Learning to interpret natural language commands through human–robot dialog. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1923–1929.
- Thrun S, Burgard W and Fox D (2005) *Probabilistic Robotics*. Cambridge, MA: MIT Press.
- Triebel R, Grimmer H, Paul R and Posner I (2016) Driven learning for driving: How introspection improves semantic mapping. In: *The 16th International Symposium on Robotics Research*, pp. 449–465.
- Tucker M, Aksaray D, Paul R, Stein GJ and Roy N (2017) Learning unknown groundings for natural language interaction with mobile robots. In: *International Symposium on Robotics Research*.
- Vedantam R, Lin X, Batra T, Lawrence Zitnick C and Parikh D (2015) Learning common sense through visual abstraction. In: *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 2542–2550.
- Walker MA, Rambow O and Rogati M (2001) SPoT: A trainable sentence planner. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Pittsburgh, PA.
- Walter MR, Antone M, Chuangsuwanich E, et al. (2014a) A situationally aware voice-commandable robotic forklift working alongside people in unstructured outdoor environments. *Journal of Field Robotics* 32(4): 590–628.
- Walter MR, Hemachandra S, Homberg B, Tellex S and Teller S (2013) Learning semantic maps from natural language descriptions. In: *Proceedings Robotics: Science and Systems (RSS)*, Berlin, Germany.
- Walter MR, Hemachandra S, Homberg B, Tellex S and Teller S (2014b) A framework for learning semantic maps from grounded natural language descriptions. *The International Journal of Robotics Research* 33(9): 1167–1190.
- Wang N, Pynadath DV and Hill SG (2016) Trust calibration within a human–robot team: Comparing automatically generated explanations. In: *Proceedings ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, Christchurch, New Zealand.
- Wang Q, Wang B and Guo L (2015) Knowledge base completion using embeddings and rules. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.
- Whitney D, Eldon M, Oberlin J and Tellex S (2016) Interpreting multimodal referring expressions in real time. In: *Proceedings IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3331–3338.
- Wong YW and Mooney RJ (2007) Generation by inverting a semantic parser that uses statistical machine translation. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 172–179.
- Yang B, Yih Wt, He X, Gao J and Deng L (2014) Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.
- Yatskar M, Ordonez V and Farhadi A (2016) Stating the obvious: Extracting visual common sense knowledge. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 193–198.
- Yu L, Poirson P, Yang S, Berg AC and Berg TL (2016) Modeling context in referring expressions. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 69–85.
- Zender H, Kruijff GJM and Kruijff-Korabayova I (2009) Situated resolution and generation of spatial referring expressions for robotic assistants. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.
- Zettlemoyer LS, Pasula HM and Pack Kaelbling L (2008) Logical particle filtering. In: *Dagstuhl Seminar Proceedings*.
- Zhang M and Chen Y (2018) Link prediction based on graph neural networks. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5165–5175.

Appendix. Index to multimedia extensions

Archives of IJRR multimedia extensions published prior to 2014 can be found at <http://www.ijrr.org>, after 2014 all videos are available on the IJRR YouTube channel at <http://www.youtube.com/user/ijrrmultimedia>

Table of Multimedia Extensions

Extension	Media type	Description
1	Video	Demonstration of physical interaction with closed cases for inferring hidden states via a Clearpath Husky A200 with a Universal Robotics UR5 manipulator.
2	Video	Demonstration of physical interaction with barrels to estimate their pliability/pushability via a Husky with a UR5 manipulator.
3	Video	Demonstration of physical interaction with cups in a tabletop domain to estimate their internal state as empty or full on a Rethink Robotics Baxter Research Platform. The determination of latent states allows completion of a tabletop clearing task.