Grounding Robot Plans from Natural Language Instructions with Incomplete World Knowledge

Daniel Nyga[*1], **Subhro Roy**[*2], **Rohan Paul**[*2], **Daehyung Park**[2], **Mihai Pomarlan**[1], **Michael Beetz**[1], and **Nicholas Roy**[2]

[1]Institute for Artificial Intelligence, University of Bremen, Germany
[2]Computer Science and AI Laboratory, Massachusetts Institute of Technology, USA

Abstract: Our goal is to enable robots to interpret and execute high-level tasks conveyed using natural language instructions. For example, consider tasking a household robot to, "prepare my breakfast", "clear the boxes on the table" or "make me a fruit milkshake". Interpreting such underspecified instructions requires environmental context and background knowledge about how to accomplish complex tasks. Further, the robot's workspace knowledge may be incomplete: the environment may only be partially-observed or background knowledge may be missing causing a failure in plan synthesis. We introduce a probabilistic model that utilizes background knowledge to infer latent or missing plan constituents based on semantic co-associations learned from noisy textual corpora of task descriptions. The ability to infer missing plan constituents enables information-seeking actions such as visual exploration or dialogue with the human to acquire new knowledge to fill incomplete plans. Results indicate robust plan inference from under-specified instructions in partially-known worlds.

Keywords: Human-Robot Interaction. Grounding Language Instructions.

1 Introduction

We envision collaborative robots in homes, factories, and workplaces that can be instructed to perform high-level tasks such as clearing an area, preparing a meal or performing a complex assembly task. Natural language instructions from a human are often incomplete and require environmental context and background knowledge about how activities are performed to fully determine a plan of actions to execute. Further, the robot may be asked to perform novel tasks for which the predetermined knowledge may be insufficient.

Consider a scenario where a human instructs the robot to "prepare breakfast". The robot must recognize an under-specified command and ask the human, "what do you want for breakfast?" Receiving a response, "get me cereal and a fruit", the robot must proceed with the breakfast preparation task by bringing a cereal, a bowl and a fruit, and placing these items on a table. In doing this, it has to resolve the term 'fruit' to an appropriate object present in its environment, e.g., an apple. Additionally, the robot may be expected to know that milk is typically needed for cereal and hence determine where to retrieve it from. Accomplishing this simple task requires procedural knowledge, relational reasoning, and the ability to acquire new facts and concepts, see Figure 1.

Contemporary instruction following models [1, 2, 3, 4] relate input instructions with actions associated with perceived entities in the scene. These models lack the ability to incorporate abstract relational knowledge in the grounding process. Alternatively, symbolic reasoning systems [5, 6, 7, 8, 9] incorporate background knowledge while reasoning about appropriate plans for the agent. How-

*Authors contributed equally. We gratefully acknowledge funding support in part by the U.S. Army Research Laboratory under the Robotics Collaborative Technology Alliance (RCTA) Program, the Toyota Research Institute Award LP-C000765-SR, the Deutsche Forschungsgemeinschaft (DFG) collaborative research centre EASE (grant no. 1320) and the DFG research project PIPE (grant no. 322037152).

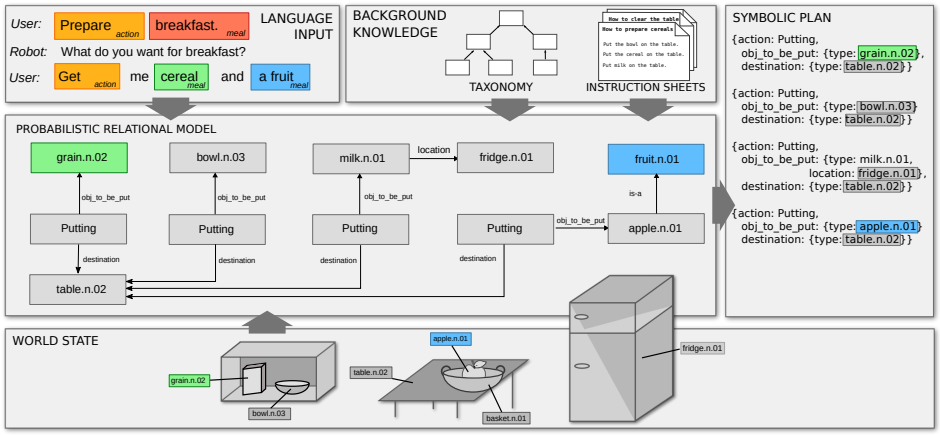


Figure 1: Illustration of semantic relational model and inference of an executable symbolic plan. A probabilistic relational model is learnt using knowledge from relational and taxonomic databases and instruction sheets for household tasks. The natural language input from the human and workspace knowledge provides only sparse evidence for the relational model. A symbolic plan comprising action sequences and their parameters must be inferred for plan execution (e.g., a sequence of four putting actions), and needs to be adapted to current environment: as the perceived world model does not contain an instance of milk, its most likely location is inferred (e.g., fridge) to inform exploratory motion.

ever, these approaches are typically non-probabilistic and do not directly integrate context from the current world state while reasoning about feasible plans.

The contribution of this work is a probabilistic model that enables a robot to infer symbolic plans from natural language commands in scenarios where the workspace is partially observed or the robot's background knowledge is insufficient. We observe that background conceptual knowledge encompassing semantic attributes, relationships, affordances etc. can be used to infer "missing" plan constituents (such as action subjects or goal locations) based on statistical correlations and taxonomic affinities. The model incorporates a probabilistic relational model over symbolic knowledge learned from noisy textual descriptions and taxonomic data bases. Reasoning over the entire space of knowledge is computationally infeasible. We adopt an online context-driven approach that hypothesizes missing knowledge components by integrating evidence from the language instruction and visual observations of the world state. The ability to infer latent plan components enables the robot to take information gathering actions such as visual exploration or or engaging in dialogue with the human to replenish its knowledge gap and enable plan synthesis. This work contributes towards language grounding models that are more robust towards incomplete world knowledge, enabling robots to make inferences or seek information about missing plan elements.

2 Preliminaries and Problem Formulation

We consider a robot operating in a workspace populated with a set of objects O and capable of executing a motion trajectory μ_t. Each object is associated with a pose, a geometric model and a semantic class label estimated using a perception system. The human operator communicates with a robot using natural language utterances Λ_t conveying instructions or factual knowledge relevant for the task. Following [10], we adopt a (first order) logical representation for workspace knowledge capturing the semantic properties and relationships that are true for the robot's environment. Formally, let C denote a set of constants associated with semantic entities such as a *box*, *cup*, *tray*, *breakfast* etc. that can potentially exist in the environment. Let L denote logical predicate symbols that express an attribute for an entity or relationships between entities, such as *On($\cdot, \cdot$)*. Predicates from L when propositionalized over constants from C form the set of "instantiated" predicates $\mathcal{X}_t$. An example of instantiated predicate is *On(block, table)*. Finally, a "grounded" predicate is an instantiated predicate whose arguments correspond to physical entities in the robot's workspace.

The robot's knowledge is encoded in terms of instantiated predicates, representing physical concepts and relations derived from perceived entities O. These include semantic types such as *block, table, robot, human* etc. or spatial relations such as *On(block, table)*, *Left(box, robot)* etc. The space of physical concepts defined over the set of detected objects constitute the perceptual world state Υ_t. A set of "factual" knowledge predicates denoted by K_t includes taxonomic subsump-

tion relations (*SubClass(apple, fruit)*), mereological relationships (*HasA(button, blender)*, object affordances (*Location(milk, fridge)*), unobserved object states (*IsUsed(carton)*) or abstract attributes (*IsMine(cup)*) communicated by a human. The space of factual concepts can expand over time through observation and interaction. The symbolic knowledge state K_t also incorporates "procedural" knowledge about complex activities such as cooking, clearing, assembly etc. Procedural knowledge can be encoded as a refinement relation $\mathcal{R}$ that associates a complex task (e.g., making a milkshake) with a sequence of constituent actions (pouring, mixing, cutting etc.).

We assume that the robot can affect the environment by executing actions such as moving an object from a source location to a destination, changing the state of a button or opening an articulated object. Let A denote the set of symbolic action predicates defined in terms of goals and constraints associated with the objects affected by the action. For example, *Move(cup, table, tray)*, *Push(button)*, *Open(fridge)*, *Pour(milk, cup)* etc., A temporally-ordered sequence of grounded symbolic actions $\{a_0, a_1, \ldots, a_n\}$ forms a "grounded" symbolic plan σ that accomplishes a high-level task. Let Σ denote the space of probable symbolic plans. The robot's goal is to infer a grounded plan given a language instruction, background knowledge and the observed world state, denoted as the likelihood $P(\sigma_t|\Lambda_t, \Upsilon_t, K_t)$. The estimated symbolic plan is provided as an input to a low-level motion planner to generate a metric trajectory to accomplish the goal. The estimation problem can be stated as:

$$\sigma_t^* = \arg\min_{\sigma \in \Sigma} P(\sigma_t|\Lambda_t, \Upsilon_t, K_t). \tag{1}$$

The estimated symbolic plan can be provided to a low-level motion planner to generate a platform-dependent trajectory. Note that the robot's workspace maybe only partially known and the language instruction may be underspecified. The factor $P(\sigma_t|\Lambda_t, \Upsilon_t, K_t)$ denotes the likelihood of a symbolic plan given sparse context from visual observations and language augmented with background knowledge about the workspace.

In this paper, we show how background knowledge can be used to infer the latent symbolic knowledge necessary for complete plan inference. Further, reasoning over possible plans with the expansive space of factual knowledge is computationally intractable. We show how evidence from language Λ_t and the observed world state Υ_t can be used to construct a reduced symbolic model for relevant background knowledge that is amenable to tractable inference. Next, we introduce a probabilistic representation for semantic knowledge and subsequently present a probabilistic model that leverages the representation for plan estimation under incomplete world knowledge.

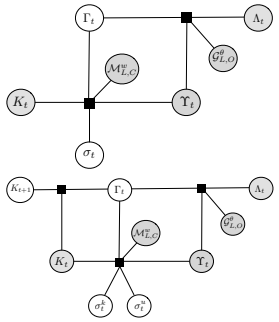
3 Knowledge Representation

The perceived and factual predicates along with the set of possible symbolic actions cumulatively form the robot's workspace knowledge $\mathcal{X}_t = \{\Upsilon_t \cup K_t \cup A_t\}$. Semantic concepts constituting $\mathcal{X}_t$ are often statistically correlated. For example, objects may be located in typical locations (milk in fridge), actions are typically performed on particular objects in certain ways (cutting a vegetable with a knife or empty boxes typically discarded in trash). Such co-associations allow humans to make inferences about plan attributes that may not be explicitly stated in the language instruction. We adopt a Markov logic network (MLN) representation [5] for modeling the probabilistic relationships between semantic concepts. Next, we briefly review the MLN formulation as applied to our domain.

Let $\mathcal{M}_{L,C}^w$ denote a MLN that models the joint likelihood of instantiated predicates derived from L defined over symbolic entities C. The MLN defines a Markov network with a node in the network for each instantiated predicate in L. The model assumes a set of formulae F_i that are indicative of probabilistic associations between binary logical predicates in the knowledge state. For example, the formula, $\{\forall x, y$ *IsEmpty*$(x) \wedge$ *Carton*$(x) \wedge$ *TrashCan*$(y) \implies$ *Destination*$(x, y)\}$ encodes the knowledge that used cartons are likely to be discarded in the trash can. The formulae are probabilistic, each associated with a weight w_i, indicative of how strong the association is: the higher the weight, the greater the difference in the log-probability between a knowledge state where the formula is true and one that does not [11]. Formulae are indicative of possible correlations between concepts and hence induce shared factors between instantiated predicates in the induced Markov network. The likelihood of a possible knowledge state can be expressed as:

$$P(\mathcal{X}_t = x|\mathcal{M}_{L,C}^w) = \frac{1}{Z}\prod_i \phi_i\left(x_{\{i\}}\right)^{n_i(x)} = \frac{1}{Z}\exp\left(\sum_i w_i n_i(x)\right), \tag{2}$$

Figure 2: Probabilistic model. The variable Υ_t denotes the metric world state obtained from visual observations. Let O denotes the detected objects. L denotes predicate symbols and the variable K_t represent factual knowledge. The robot estimates a symbolic plan σ_t from language instructions Λ_t from a human. **Up:** Grounding plans with incomplete background knowledge. The variable Γ_t denotes the (literal) symbolic interpretation of the instruction using a language grounding model $\mathcal{G}^{\theta}_{L,O}$. Here, L denotes predicate symbols and O denotes the detected objects and θ denote learned weights. A Markov logic network $\mathcal{M}^{w}_{L,C}$ define a joint distribution over the knowledge state that includes factual knowledge, grounding from language, candidate actions and perceptual context. Given partial observations of the environment, the relational model can infer "missing" symbols required for plan completion correlated with the observed workspace entities allowing information gathering actions. **Down:** Augmented model for incremental knowledge acquisition. The model estimates a partition over known and unknown plans, σ^{k}_{t} and σ^{u}_{t} based on whether the input utterance conveys a novel concept. A language query is generated if an unknown symbol is inferred. The grounding for the human's response updates the knowledge state as K_{t+1}.



where x is a binary vector $\{0,1\}^{|\mathcal{X}_t|}$ that denotes a truth assignment to all predicate instantiations of $\mathcal{X}_t$, $x_{\{i\}}$ denotes the state of the atoms appearing in F_i, $n_i(x)$ is the number of satisfying assignments of formula F_i in x, Z is normalization constant and $\phi_i\left(x_i\right) = e^{w_i}$.

4 Probabilistic Model

We introduce a probabilistic model for inferring symbolic plans from natural language instructions given the perceived world state and background knowledge expressed as the likelihood $P(\sigma_t|\Lambda_t, \Upsilon_t, K_t)$. Language utterances from a human can instruct the robot to perform high-level tasks such as "prepare milkshake but do not use the apples" or convey factual knowledge such as "the box on the left is empty". The process of assigning meaning or "grounding" a language instruction involves estimating a set of instantiated predicates Γ_t that compactly represents observations of the task specification and factual knowledge conveyed in the instruction. For example, the clause $\textit{Prepare}(\textit{milkshake}) \wedge \neg \textit{NeedsA}(\textit{milkshake}, \textit{apple})$ estimated from the milkshake preparation instruction. The inclusion of language grounding variables Γ_t allows the plan inference factor in Equation 1 to be expressed as:

$$\sigma^{*}_{t} = \arg\min_{\sigma \in \Sigma} \sum_{\Gamma_t} \overbrace{P(\sigma_t|\Gamma_t, \Upsilon_t, K_t; \mathcal{M}^{w}_{L,C})}^{\text{Knowledge reasoning}} \overbrace{P(\Gamma_t|\Lambda_t, \Upsilon_t; \mathcal{G}^{\theta}_{L,O})}^{\text{Language grounding}}. \tag{3}$$

The input language only provide *sparse* evidence for the intended plan to be executed by the robot and require joint reasoning with background knowledge. The factor $P(\Gamma_t|\Lambda_t, \Upsilon_t; \mathcal{G}^{\theta}_{L,O})$ infers a *literal* grounding Γ_t of a language instruction Λ_t given the observed world state Υ_t. The factor $P(\sigma_t|\Gamma_t, \Upsilon_t, K_t; \mathcal{M}^{w}_{L,C})$ determines a *complete* interpretation of the instruction in terms of a grounded symbolic plan σ_t conditioned on groundings from language Γ_t, background knowledge K_t and the observed workspace Υ_t. The language grounding model is realized using a structured log-linear model [12] parametrized as $\mathcal{G}^{\theta}_{L,O}$. The model relates phrases in the input instruction with semantic concepts derived from the observed environment. The knowledge reasoning factor is realized using probabilistic relational model $\mathcal{M}^{w}_{L,C}$ introduced previously in section 3. Figure 2a illustrates the graphical model.

4.1 Inferring Plans from Instructions in Partially-known Workspaces

We now detail the estimation of a symbolic plan σ given the input instruction and workspace knowledge, $P(\sigma_t|\Gamma_t, \Upsilon_t, K_t; \mathcal{M}^{w}_{L,C})$. The distribution over the full space of workspace knowledge is represented as a Markov logic network modeling the joint likelihood $P(\sigma_t, \Gamma_t, \Upsilon_t, K_t|M^{w}_{L,C})$ introduced previously in Section 3. The workspace knowledge is derived from two sources: the observed metric state Υ_t derived from perception and the set of grounding symbols Γ_t from language indicative of the intended actions to be performed by the robot. With the observed observed world state Υ_t and groundings from language Γ_t as "evidence", the model "queries" a likely grounded symbolic plan σ_t as the following conditional likelihood over the relational model:

$$P(\sigma_t|\Gamma_t, \Upsilon_t, K_t; \mathcal{M}^{w}_{L,C}) = \frac{\sum_{x \in \sigma_t \cap \Gamma_t \cap \Upsilon_t \cap K_t} \exp\left(\sum_i w_i \cdot n_i(x)\right)}{\sum_{x \in \Gamma_t \cap \Upsilon_t \cap K_t} \exp\left(\sum_i w_i \cdot n_i(x)\right)}. \tag{4}$$

The model uses statistical correlations between symbols of background ontological knowledge to infer plan components that may be implicit in an under-specified instruction or missing from the observed world model; in essence, estimating symbolic plans that are maximally correlated with visual and linguistic context.

Inferring missing plan components involves reasoning over the full space of background knowledge state K_t. In practice, the number of variables in the ground knowledge state can be very large, $\mathcal{O}(|L| \cdot |C|^N)$, where N denotes the maximum arity of a predicate in symbol space. We address this problem by using context from input language Λ_t and the observed world state Υ_t to postulate a smaller set of probable knowledge state variables $\widetilde{K_t}$. Formally, the model determines a smaller space of symbolic entities $\widetilde{C}$ and relations $\widetilde{L}$ forming a context-dependent relational model $\mathcal{M}^w_{\widetilde{L},\widetilde{C_t}}$ of size $\mathcal{O}(|\widetilde{L}| \cdot |\widetilde{C}|^N)$ where $(|\widetilde{C}| \ll |C|)$ and $(|\widetilde{L}| \ll |L|)$. The reduced knowledge state $\widetilde{K_t}$ allows plan inference to be factorized as:

$$P(\sigma_t|\Gamma_t, \Upsilon_t, K_t; \mathcal{M}^w_{L,C}) = \sum_{\widetilde{K_t}} \overbrace{P(\sigma_t|\Gamma_t, \Upsilon_t, \widetilde{K_t}; \mathcal{M}^w_{\widetilde{L},\widetilde{C}})}^{\text{Plan inference with reduced model}} \overbrace{P(\widetilde{K_t}|\Gamma_t, \Upsilon_t, K_t)}^{\text{Estimating reduced knowledge base}} \quad . \tag{5}$$

Here, the estimated knowledge state $\widetilde{K_t}$ inferred from linguistic context Γ_t and observations of the world state Υ_t forms a Markov blanket between the plan inference and most variables in the large knowledge state K_t. Probable plans are determined via lifted inference [13].

The inferred symbolic plan σ_t consists of a sequence of symbolic actions which may contain symbolic or "ungrounded" literals associated with objects that may not be detected as yet. For example, for the cereal preparation task, if the milk is located inside the refrigerator and hence not visible, the inferred plan is likely to contain an ungrounded symbol for the milk entity and a grounded symbol for refrigerator as a possible location to search for the milk entity. For grounded symbolic actions in the plan, a motion planner estimates a minimum-time, collision-free trajectory to accomplish the task. The inferred locations for ungrounded literals serve as an informed prior for guided exploration in the environment such as visual-scanning or opening the lid/cover to gather new observations.

4.2 Acquiring Missing Background Knowledge via Online Interaction

In realistic scenarios the robot's pre-trained set of concepts may not suffice. Hence, the robot needs an ability to estimate when its knowledge model lacks procedural knowledge for a task and then interact with the user to replenish its knowledge representation. We augment the graphical model to include a knowledge state variable K_{t+1} propagated to the next time step $t+1$.

Given the current estimated knowledge state, the model classifies a resulting plan as known σ^k_t if the high-level task possesses a known decomposition into plan sub-goals. Otherwise the plan is classified as unknown σ^u_t if the corresponding procedural knowledge is lacking in the background knowledge state. When a task specification is inferred as *unknown*, the the model generates a template-based language query Λ^q_t to the human operator seeking a factual description for stated concept. We assume a cooperative human operator who provides a task description in the form of a language utterance. The grounding of the language response serves as a new observation updating the current knowledge state with new relational knowledge resulting in the updated knowledge state K_{t+1}. Figure 2(b) illustrates the graphical model. The inclusion of the propagated knowledge state, and the known/unknown tasks augments the plan estimation factor as:

$$P(K_{t+1}, \sigma^k_t, \sigma^u_t|\Lambda_t, \Upsilon_t, K_t) = \sum_{\Gamma_t} \overbrace{P(K_{t+1}|K_t, \Gamma_t)}^{\text{Knowledge update}} \overbrace{P(\{\sigma^k_t, \sigma^u_t\}|\Gamma_t, K_t, \Upsilon_t)}^{\text{Unknown/Known Plan inference}} . \overbrace{P(\Gamma_t|K_t, \Lambda_t, \Upsilon_t)}^{\text{Instruction Grounding}} .$$

5 Evaluation

5.1 Quantitative Evaluation

For the MLN $\mathcal{M}^w_{L,C}$, we use predicates indicating whether words are associated with syntactic and semantic labels like part of speech tags, dependency relations, Wordnet [14] synsets, ConceptNet relations [15], and Framenet [16] actions and roles. We use instantiated predicates representing all

relations (synonymy, hyponymy, hypernymy) between the 117000 synsets in Wordnet. For training the weights, we use 30 instruction sheets from the cooking domain and an additional 5 task descriptions collected for the household domain indicative of typical locations of household objects. The grounding model $\mathcal{G}^{\theta}_{L,O}$ is the one presented in [12].

For evaluation, we use 19 high level instructions collected from WikiHow. Each instruction is paired with textual description of the steps that the robot has to perform, and the final symbolic robot plan that needs to be executed. The instructions range from meal preparation (e.g., "prepare a milkshake") to common household tasks (e.g., "clear the table"). Finally, each instruction is paired with multiple robot workspaces, each of which contains all required objects necessary to execute the plan and 4 extra distractor objects not required for the plan. Resolving each of the 19 instructions requires analyzing a large number of instructions, since each instruction can be decomposed into component sub goals with a corresponding set of instructions. Overall, the evaluation involves analysis of 114 unique natural language instructions.

We quantitatively evaluate the proposed model in the following scenarios:

1. **Workplace Context**: We evaluate the ability of the model to generate grounded executable robot execution plans using the context of the robot's workspace. We replace a randomly chosen object's type with its hyponym in the workspace. The system has to select the hyponym as a suitable replacement for the missing object. We compare against a baseline which chooses the most similar item as a replacement. Similarity is computed using pretrained word embeddings [17].
2. **Partially known Workspace**: If items are missing in the workplace, our model estimates ungrounded literals, along with a possible location to search for the missing entity (Section 4.1). We evaluate this capability – predicting plans with search locations for missing entities in this scenario. We replace an object in the workspace with a possible location where it can be stored. The goal is to generate a plan with possible location for the missing items. We compare against a baseline which chooses locations based on average similarity (based on word embeddings) to storage locations encountered in the training data.
3. **Acquiring Missing Knowledge**: We evaluate the ability of the system to infer missing knowledge in the database, and to query the human to online replenish the missing knowledge in the database (Section 4.2). We delete the corresponding textual descriptions for input instructions. Once the model detects that relevant knowledge is missing, we provide the descriptions as a response provided by a human. The goal of the model is to ground language instructions online and generate the correct executable plan.

Table 1 summarizes the results. The proposed model significantly outperforms the aforementioned similarity based baselines. State of the art linguistic similarity metrics fail to capture the relational reasoning required for our task. In comparison to traditional language grounding work, the symbol

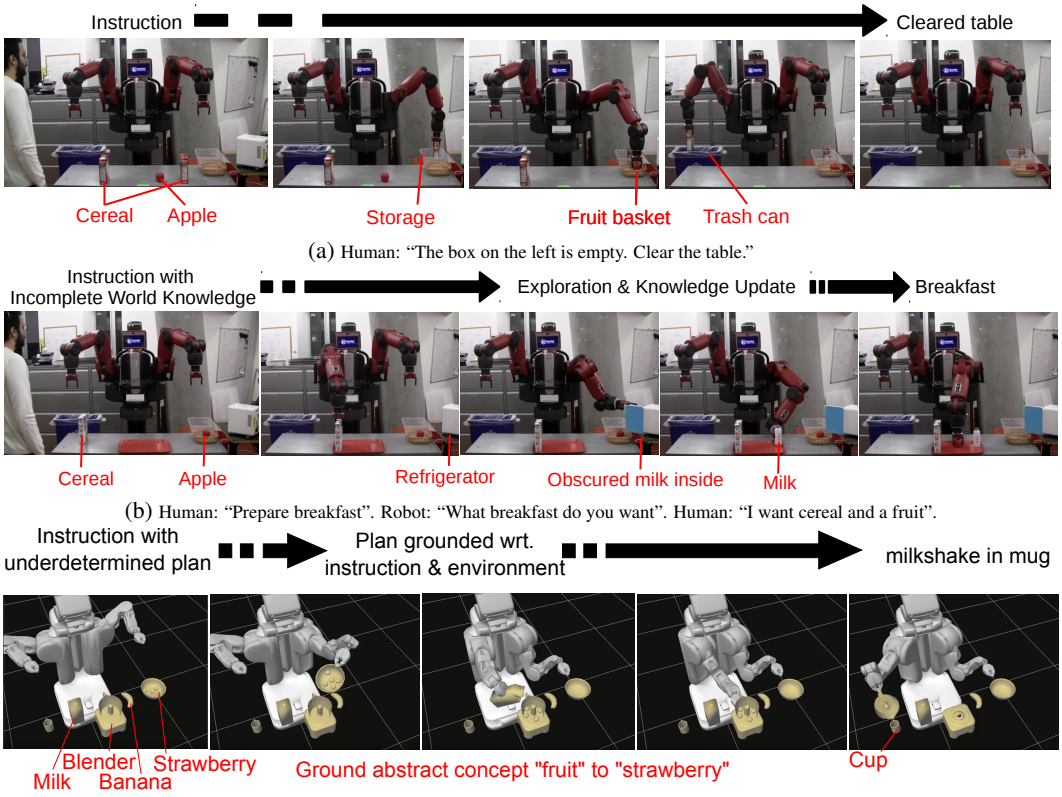
	Workspace Context	**Partially known Workspace**	**Acquiring Missing Knowledge**
Baseline	62	42	N/A
Proposed Model	96	70	48

Table 1: Accuracy (%) of generating executable robot plans under different scenarios. Each experiment was run on 50 examples.

space for our model is very large owing to the inference over concepts in Wordnet [14] and Framenet [16]. In particular, grounding a low level instruction involves inference over 42 actions, each action is associated with around 2 arguments, each argument is selected from a space of 117000 concepts. Our model leverages context provided by the world state to construct an approximate reduced set of relations (Section 4.1) to reason over during plan inference: concepts that are physically absent in the world do not need to be considered as potential alternatives in plan execution and can thus be excluded from the inference. This context-specific independences allowed search over a significantly reduced fraction of symbols – only 0.3% to 1.21% in our experiments.

5.2 Qualitative Results

The model is demonstrated on a Baxter Research Robot in a table top workspace and a PR2 robot in a simulated kitchen environment. During the experiments, the robot is commanded to perform



(a) Human: "The box on the left is empty. Clear the table."

(b) Human: "Prepare breakfast". Robot: "What breakfast do you want". Human: "I want cereal and a fruit".

(c) Human: "Make a strawberry milkshake".

Figure 3: Demonstration on a Baxter Research Robot and a simulated PR2 robot. (a) The clearing task requires relational reasoning based on object attributes (used or full) to determine the correct destinations (trash can or the storage tray respectively). (b) The sub-goals for the breakfast preparation task are *a priori* unknown to the robot. The robot infers the missing concept and queries the human to elicit a description. The model leverages relational knowledge and estimates that the milk ingredient is correlated with cereal and required for this preparation and further infers the fridge as a likely location to explore as a milk bottle cannot be seen in the workspace. (c) The recipe preparation task requires taxonomic generalization for plan completion.

clearing, assembly and recipe preparation tasks, Figure 3. A video for the demonstration is available at: `https://youtu.be/uWv-l7XMoB8`.

Table clearing task. In this scenario, two cracker boxes, a fruit, a trash can, a fruit basket and a storage box are placed on the table. The human subject conveys that, "the cracker box on the left is empty" and then tasks the robot to "clear the table". The model infers the grounding for this instruction as an abstract task composed of plan steps as learnt from prior knowledge in the instruction data. The second cracker box is assumed to be full. The relational model infers that the likely destination for the empty box is the trash can and determines that the full cracker box is to be stored away. Taxonomic reasoning enables the apple to be inferred as a fruit and the fruit basket is determined to be its most likely destination.

Breakfast preparation task. The robot's workspace consists of a can, a box, a fruit and a refrigerator with a milk carton inside. The location of the milk carton is unknown. The robot is asked to "prepare breakfast". The model infers this task specification as novel and generates a question, "what breakfast would you like to have". The human further presents a factual description as, "I would like cereal and a fruit". The model infers the apple as a plan constituent using background taxonomic knowledge. Using learned procedural knowledge, the model infers that cereal is highly correlated with milk and populates the plan with a symbolic abstraction for the milk concept and estimates its likely location to be the refrigerator. The robot performs exploratory actions by opening the fridge, and then detects the milk carton and moves the carton to the table completing the task. As the model maintains a distri-

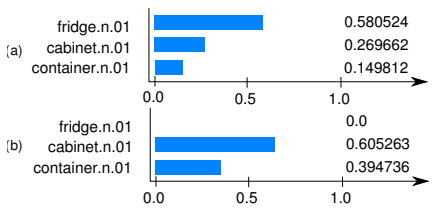


Figure 4: Distributions over likely locations of milk: (a) the original prior and (b) the posterior distribution updated with the fact that the milk is not in the fridge.

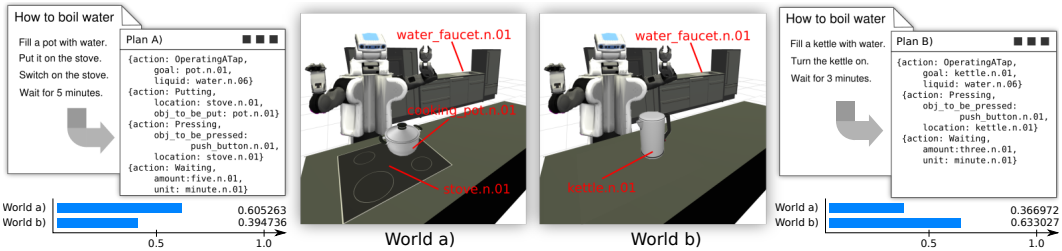


Figure 5: Inference over multiple plan hypotheses: A distribution over two equivalent robot plans for boiling water using different objects and devices, conditioned on different world models. World (a) is equipped with a cooking pot and a stove, World (b) contains a kettle. Worlds being more compatible with the respective plan (with respect to objects available), gain higher likelihoods.

bution over possible locations of objects, more informed posteriors can be computed with the world model having been updated. If, after exploration, for instance, the milk turns out to be not in the fridge, the world model can be updated and the distribution over locations is being recomputed with the new evidence. Figure 4 illustrates this reasoning case.

Recipe following task. A simulated PR2 robot in a kitchen environment is tasked to prepare a strawberry milkshake. The robot's background knowledge contains a natural language recipe for preparing a milkshake and serving a drink. The natural language recipe is specified only abstractly, referring to the flavor of the milkshake as 'fruit'. Here, the taxonomic knowledge is used to adapt the recipe to the instruction and the environment, such that the final plan contains an action to put the strawberries into the blender, but not the other fruits. The robot also infers that serving the milkshake requires pouring it into a glass. The workspace did not contain a glass but a cup, so it infers the cup as an appropriate alternative to adapt the plan.

A key feature of the proposed approach is the ability to compute probability distributions over multiple plan options (e.g. multiple ways of achieving the same goal) *conditioned* on a respective world model. This allows the robot to select the action sequence that is *most compatible* in the given environment. An example is shown in Figure 5: Two different recipes for a 'boiling water' task are given, one heating water in a pot on the stove, the other one using a water kettle. Under evidence of two different worlds, plans that are more compatible with a respective world gain higher probability.

6 Related Works

Contemporary language grounding models [1, 3, 4, 18] estimate correspondences between linguistic constituents and the semantic entities observed in the world model but do not directly incorporate background relational knowledge in inference. Past work in robot language grounding and semantic parsing [2, 19, 20, 21] map instructions to formal logical representations, but they place the onus on a downstream model [12, 22, 23, 24] to derive executable controllers from high level logical forms. In contrast, our method jointly infers both the high-level predicates a sequence of low-level sub-goals directly executable by the robot. In related work [25, 7, 8] propose cognitive architectures modeled as symbolic production systems. In this work, we take a probabilistic approach that enables probabilistic estimates for plan hypotheses. Approaches such as [5, 6, 8, 25] use symbolic data bases and ontologies but ignore the perceptual context. Approaches in [26, 6] construct context-sensitive relational knowledge bases from local observations but do not consider the plan inference task. Other efforts have focused on acquiring abstract concepts by interacting with the user [4, 27, 28, 29, 30].

7 Conclusions

We introduced a probabilistic model for interpreting complex activities communicated via natural language as a sequence of symbolic actions that an autonomous agent can execute. The model incorporates a learned relational knowledge representation that enables inference over plan elements not explicitly stated in language or missing from the perceptual view of the workspace. Relational inference is performed over a context-dependent model based on accrued context of observed entities and the history of linguistic interactions with the human. When the robot encounters a novel task the model allows the robot to interact with the human to acquire new knowledge. We demonstrate robot manipulators following high-level instructions in incompletely known workspaces.

References

[1] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. J. Teller, and N. Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *National Conference on Artificial Intelligence (AAAI)*, 2011.

[2] C. Matuszek, E. Herbst, L. Zettlemoyer, and D. Fox. Learning to parse natural language commands to a robot control system. In *Experimental Robotics*, pages 403–415. Springer, 2013.

[3] T. M. Howard, S. Tellex, and N. Roy. A natural language planner interface for mobile manipulators. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[4] C. Liu, S. Yang, S. Saba-Sadiya, N. Shukla, Y. He, S.-C. Zhu, and J. Chai. Jointly learning grounded task structures from language instruction and visual demonstration. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[5] D. Nyga and M. Beetz. Everything robots always wanted to know about housework (but were afraid to ask). In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[6] M. Waibel, M. Beetz, J. Civera, R. d'Andrea, J. Elfring, D. Galvez-Lopez, K. Häussermann, R. Janssen, J. Montiel, A. Perzylo, et al. Roboearth. *IEEE Robotics & Automation Magazine*, 18(2):69–82, 2011.

[7] J. R. Anderson. Act: A simple theory of complex cognition. *American Psychologist*, 51(4): 355, 1996.

[8] J. E. Laird, A. Newell, and P. S. Rosenbloom. Soar: An architecture for general intelligence. *Artificial intelligence*, 33(1):1–64, 1987.

[9] J. E. Laird, C. Lebiere, and P. S. Rosenbloom. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, 38(4), 2017.

[10] S. Russell and P. Norvig. Artificial intelligence: A modern approach. *Prentice-Hall, Egnlewood Cliffs*, 25:27, 1995.

[11] M. Richardson and P. Domingos. Markov logic networks. *Machine learning*, 62(1-2):107–136, 2006.

[12] R. Paul, A. Barbu, S. Felshin, B. Katz, and N. Roy. Temporal grounding graphs for language understanding with accrued visual-linguistic context. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[13] D. Jain, P. Maier, and G. Wylezich. Markov logic as a modelling language for weighted constraint satisfaction problems. In *Eighth International Workshop on Constraint Modelling and Reformulation, in conjunction with CP*, 2009.

[14] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

[15] C. Havasi, R. Speer, and J. Alonso. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, pages 27–29. Citeseer, 2007.

[16] C. F. Baker, C. J. Fillmore, and B. Cronin. The structure of the framenet database. *International Journal of Lexicography*, 16(3):281–296, 2003.

[17] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

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

[19] S. W.-t. Yih, M.-W. Chang, X. He, and J. Gao. Semantic parsing via staged query graph generation: Question answering with knowledge base. *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015.

[20] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

[21] D. Jain, L. Mosenlechner, and M. Beetz. Equipping robot control programs with first-order probabilistic reasoning capabilities. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[22] A. Boteanu, T. Howard, J. Arkin, and H. Kress-Gazit. A model for verifiable grounding and execution of complex natural language instructions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[23] H. Kress-Gazit, G. E. Fainekos, and G. J. Pappas. Translating structured english to robot controllers. *Advanced Robotics*, 22(12):1343–1359, 2008.

[24] C. Finucane, G. Jing, and H. Kress-Gazit. Ltlmop: Experimenting with language, temporal logic and robot control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010.

[25] K. Talamadupula, G. Briggs, T. Chakraborti, M. Scheutz, and S. Kambhampati. Coordination in human-robot teams using mental modeling and plan recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014.

[26] S. Chernova, V. Chu, A. Daruna, H. Garrison, M. Hahn, P. Khante, W. Liu, and A. Thomaz. Situated bayesian reasoning framework for robots operating in diverse everyday environments.

[27] J. Thomason, A. Padmakumar, J. Sinapov, J. Hart, P. Stone, and R. J. Mooney. Opportunistic active learning for grounding natural language descriptions. In *Conference on Robot Learning (CoRL)*, 2017.

[28] N. H. Kirk, D. Nyga, and M. Beetz. Controlled Natural Languages for Language Generation in Artificial Cognition. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[29] N. Gopalan and S. Tellex. Modeling and solving human-robot collaborative tasks using pomdps. In *RSS Workshop on Model Learning for Human-Robot Communication*, 2015.

[30] J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2009.