9

Teaching Robots New Tasks through Natural Interaction

Joyce Y. Chai, Maya Cakmak, and Candace Sidner

Abstract

This chapter focuses on the main challenges and research opportunities in enabling *natural interaction* to support interactive task learning. Interaction is an exchange of communicative actions between a teacher and a learner. Natural interaction is viewed as an interaction between a human and an agent that leverages ways in which humans naturally communicate and does not require the human to have any prior expertise. The goal of communication is to achieve common ground and allow the learner to acquire new task knowledge. This chapter outlines the different types of knowledge that can be transferred between agents and discusses the perception, action, and coordination capabilities that enable teaching-learning interactions.

Introduction

Extending the framework introduced by Mitchell et al. (this volume), our focus in this chapter is on *natural interactions* between a human and an agent to enable interactive task learning. To reflect most prior work on this topics, we focus on interactive task learning scenarios where the teacher is a human and the learner is a physically embodied agent (e.g., robot) as opposed to a software agent.

Imagine an elderly couple, Katie and John Smith, who purchased a robot "Mia" as their personal assistant. Mia comes equipped with general knowledge of household chores and perceptual capabilities to recognize common household objects, such as those sold in grocery stores and hardware stores. Mia also has basic manipulation skills, such as grasping common objects or opening different types of containers. Despite these preexisting capabilities, Mia is unable to perform many tasks at Katie and John's house right out of the box. Not only does it need to be taught the unique tasks that the Smiths desire, it also must acquire new knowledge and capabilities that will enable those tasks. The process of learning these tasks as well as task-relevant knowledge and capabilities happens through various forms of interaction with people, as in the following scenarios:

1. On the day of delivery, David, an employee from the company that manufactured Mia, arrives at the Smiths' with the new robot. David has an associate degree in robotic technology and has completed training on how to teach robots. The process starts with teaching Mia a map of the Smiths' house. David manually drives Mia to different rooms to construct the map and also verbally provides information about each room and different points and regions in the room, such as where the main entrance is and locations of appliances, trash bins, tools, and supplies. Next, David programs a set of basic skills tailored for the Smiths' house, such as how to open or close their cabinets, drawers, and appliances as well as how to operate various tools and appliances. He teaches Mia these skills by moving the robot's arm to demonstrate them. Then, under various scenarios, David tests the learned skills to ensure they are robust.
2. Once Mia is settled in the new house, the Smiths continue to teach Mia new knowledge and tasks. For example, they show where they put their groceries or kitchen tools by pointing where they are and verbally describing their locations with natural language: "The waffle maker goes in the bottom cabinet next to the stove." Katie teaches Mia how to make their favorite dish from a family recipe. Using natural language and deictic gestures, she shows Mia different ingredients and demonstrates how and in what order

to mix the ingredients. Mia sometimes has difficulty understanding Katie's instruction. For example, when Katie asks Mia to "*grind the onion*," Mia does not understand what "grind" means and subsequently asks Katie for further instructions. Katie then provides detailed step-by-step instructions to show Mia how to perform the action "grind": "*cut the onion in half, put them into the blender, and press the top button.*" By following Katie's instruction and observing the change of the onion, Mia learns the meaning of the verb "*grind*" with respect to how the corresponding action changes the physical world. Mia can now transfer this understanding and perform related actions, such as "*grind the carrot,*" assuming Mia understands what a carrot is. Through this type of interaction, Mia continuously optimizes its task performance based on feedback from Katie, such as "That looks slightly overcooked. Try reducing the baking time next time around."

3. For outdoor chores (e.g., a simple car maintenance task) John instructs Mia similarly to how he taught his son: John demonstrates to Mia how to (a) open the hood of the car, (b) check the engine oil, (c) check the radiator coolant and fill if needed, (d) check the windshield wiper fluid and fill if needed, and (e) replace the air filter if it is dirty. John and Mia both use language and deictic gestures to establish shared attention during the teaching-learning process. Once John explains and demonstrates "how to fill radiator coolant," Mia can apply the learned skill to "fill windshield wiper coolant." To teach the task, John uses conditional statements (e.g., "if the oil is below this line, then add coolant") and purposive descriptions (e.g., "you hold it *because the funnel is too big*," "put it *so that the screw comes through the narrow part*," or "place it right where the middle center opens into the screw *so that the screw goes through the middle hole where it's open*."). Mia extracts causal-effect relations and converts them into schemas to support action planning and execution. The process also involves learning background knowledge mentioned in conditional statements, such as a too large funnel, the air filter being dirty, the time needed to hold an object in place, or the colors of objects though demonstrations or examples.
4. To understand Mia's capabilities and limitations, the Smiths can ask Mia different questions about its knowledge and its representation of the shared environment and tasks. These questions not only include "what" questions, but also "why" and "how" questions to assess Mia's reasoning and decision-making capabilities. Mia also proactively communicates with the Smiths about its internal representations of the world and the tasks, as well as the underlying reasoning that might take place to reach certain conclusions or decisions. Mia can even teach the Smiths' grandson how to cook their favorite dish and how to do car maintenance.

These scenarios illustrate different types of natural interaction that humans can use to teach robots new tasks or task-relevant knowledge and capabilities: by performing the task themselves, by verbally or kinesthetically guiding the robot, or through situated language instructions and gestures. This natural interaction between humans and agents instantiates the general framework of interactive task learning, as shown in Figure 9.1. The human teacher has some *target task knowledge* in mind and intends to transfer this knowledge to the robot through various forms of interaction. Let S represent the set of states of the physical world relevant to the task and S_c represent the set of states of communication, such as the verbal utterances or focus of attention of the teacher at each step of the interaction. The robot learner perceives a *task-related world state* $s \in S$ through its sensors and constructs a *communicative state* $s_c \in S_c$ based on its perception of the teacher's communicative actions. Let A represent the set of task-related actions (e.g., pick up an object) and A_c the set of communicative actions (e.g., asking for confirmation for its interpretation of a world state) available to the robot through its effectors. At each step of the interaction, the robot needs to decide what *task-related actions* $a \in A$ and/or *communicative actions* $a_c \in A_c$ it should take, given its current state and learning goals. The sequence of states and actions that a robot goes through during interactive tasks

learning constitutes its *interaction experience*. The robot needs to then extract *learning experience* from its interaction experience to obtain examples, specifications, and feedback from which it can obtain new *task knowledge*.

Enabling interactive task learning on robots through natural interactions requires a wide range of capabilities for perception, action, reasoning, learning, decision making, and communication. Here, we discuss the challenges and open questions associated with these capabilities. Specifically, we explore:

1. Forms of human teaching and the different kinds of knowledge that can be taught or learned through interaction.
2. Capabilities to perceive and infer task-related state and communicative state through sensors, including visual scene understanding, language understanding, and grounding language to visual perception (e.g., the environment, perception of human gestures, and perception of human actions)
3. Capabilities to act in the environment through effectors, including acting to manipulate the environment and communicating to the human during interaction.
4. Capabilities to manage and coordinate interaction and establish common ground.

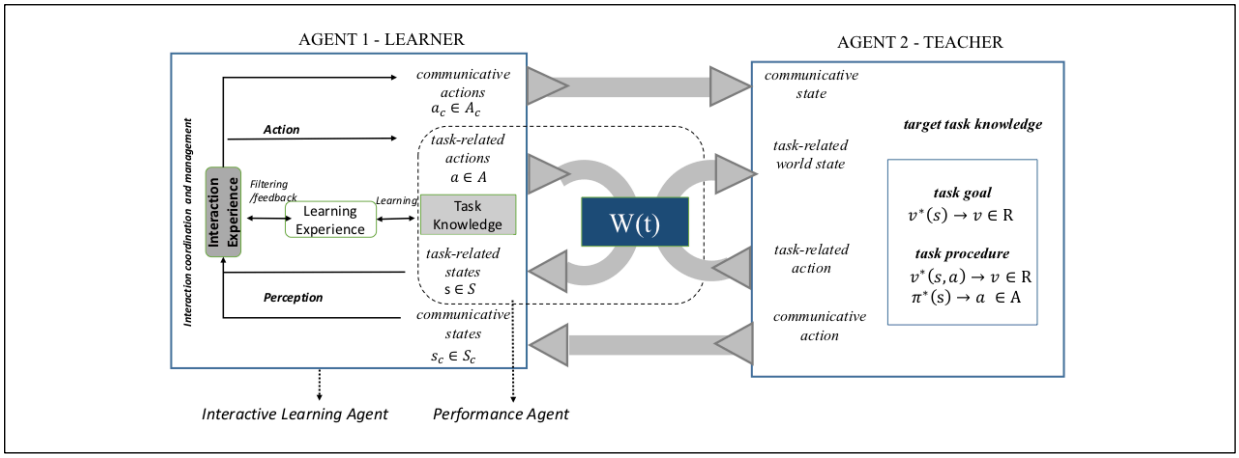

Figure 9.1 Extended two agent + world model separating task-related actions and communicative state and actions. Task-related states (S) and actions (A) are the minimal set of states and actions that an agent needs to perform the target task successfully. Communicative states (S_c) and actions (A_c) are what an agent needs to communicate to extract useful data and provide feedback for learning.

Types of Task Knowledge and Forms of Interaction

Humans can learn new tasks from other humans through various means: watching each other perform the task, doing the task themselves accompanied by instructions and guidance, or conversing and imagining the task without performing any actions (e.g., acquiring a new recipe). Similarly, as illustrated in our example scenario, robots can learn from humans in analogous ways. During interactive task learning, task knowledge from human teachers (i.e., targeted task knowledge) is transferred to the robot learner.

As shown in Figure 9.2, in interactive task learning the robot needs to extract learning experience from interaction experience through interaction. The learning experience can involve examples of goal states, examples of action sequences that lead to a goal, or evaluations of action sequences generated by the robot. These different learning experiences can be expressed in terms of the physical world state (s_i), task-related actions (a_i), and values assigned to them (v_i). The goal of task learning is to extract different types of task knowledge such as task goal (e.g., $v(s) \rightarrow v$) and task procedure (e.g., a policy to perform the task $\pi(s) \rightarrow a$) from these experiences. Different learning algorithms require specific types of experience data (e.g., direct policy learning requires sequences of state-action pairs). The role of the communicative

actions is to extract this data from the unstructured stream of data that the agent experiences. For instance, communicative actions by the teacher might indicate the start and end of a demonstration to help the learning process, even though the communicative states and actions are excluded from the learning data. As we discuss next, the way in which task knowledge is transferred and the role of communicative actions in that process largely depends on the type of task knowledge.

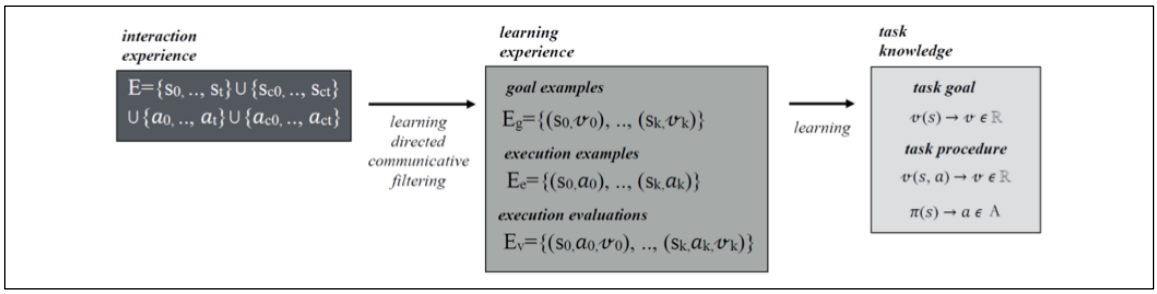

Figure 9.2 Interactive task learning is the process of converting the learning agent's experience into task knowledge. Different types of knowledge are learned from different types of example data.

Task Knowledge Types

The main goal of task learning is to acquire *task knowledge*, which defines what a task is and provides sufficient information to permit the robot to perform the task on its own. There are different types of task related knowledge and capabilities (described later in this section) that can be acquired during interaction. As discussed by Laird et al. (this volume), task related knowledge often includes goals, actions, and constraints which define the problem space as well as procedural/policy knowledge which is needed to perform the task. In this chapter, we particularly focus on two types of task knowledge and their representations: *task procedures* and *task outcomes.*

Task procedure information captures what the agent needs to do to complete the task, as shown in Figure 9.2. Most existing agent frameworks represent procedural information as a policy, which is a function that maps the perceived state to an action ($\pi(s) \rightarrow a$). Such functions can be represented with many different types of classifiers or regressors and can be learned from examples. Process information can also be captured in more explicit forms such as plans, programs, Finite State Machines, or Hierarchical Task Networks. Although these different representations do not necesarily provide a full mapping of states to actions, they still capture procedural knowledge by specifying a sequence, a partial ordering, a schedule, or a hierarchical organization of actions in the context of a task. For example, Pardowitz et al. (2007) introduced task precedence graphs (TPGs) that capture ordering constraints between actions involved in a task. Similarly, Ekvall and Kragic (2008) represent tasks with a set of ordering constraints between pairs of actions. Alexandrova et al. (2015) use a flow diagram to represent tasks with actions that have pre- and post-conditions that can cause branching in the program. Huang and Cakmak (2017) use the general-purpose visual programming language, Blockly, to represent various tasks that branching and looping.

Task outcome information relates to the goals or desired outcomes of a task, independent of the process followed to achieve them. This is different from the actual outcomes when performing a task (which can be expected or uninteded). Task goals are often captured by the reward or value functions associated with states and actions, assuming the agent is maximizing reward or value. In practice, task goals might be easier to express in terms of world states in which the task is considered complete; for instance, a conjunction of state variables that need to be true or other arbitrary functions that evaluate a given state in terms of whether the goal is achieved. A value can then be associated with each state based on how close they are to a goal state. The task "tidy up the living room," for example, could be specified with the list of items

in the room and their desired locations, without any information on how to get them there. Such a representation was used by Chao et al. (2011) to represent simple object re-configuration tasks. The ability to carry out tasks based solely on specified goals often requires the robot to have planning capabilities.

Some task representations involve combinations of process and outcome information. For example, a recipe for a particular dish specifies not only a sequence of actions but also mentions what to expect at the end of the process or when a task is considered complete.

Forms of Interaction in Transferring Task Knowledge

There are many forms of interaction that enable transfer of task knowledge. Our focus here is on two key types of information transferred in those interactions: *demonstrations* of the task or direct *specifications* of task constraints or properties.

In learning task processes from **task demonstrations**, multiple demonstrations provide alternative ways of achieving the same task (Argall et al. 2009). Different task representations capture this information in different ways. For example, a partially ordered plan captures alternative orderings of low-level actions. Hence different demonstrations of the task might involve a different ordering of actions. Similarly, a program with conditionals and loops captures alternative ways of performing a task, depending on a perceivable "condition" or different numbers of repetitions contingent on user-specified or environmental parameters. Different demonstrations of such tasks will involve alternative traces of the program.Task outcomes can also be taught by demonstration. Multiple examples provide variations of the states in which the task is considered successfully completed. One of the key computational challenges is to identify parts of the state that are relevant/irrelevant for the task. It is therefore important for demonstrations provided by the teacher to involve such variations.

Tasks can be demonstrated through different forms of interaction, for example, by the teacher performing the task, or provided directly to the robot, with guidance from the human teacher

- In tasks performed by humans, one of the most intuitive ways to demonstrate a task is for a human to perform it herself. For the robot to learn from this type of demonstration, the robot must be able to perceive the human's actions and/or the effects of human action on the environment. Perception of human actions can be facilitated through external sensors or wearable sensors on the human. Once a robot perceives human actions, they need to be mapped to corresponding robot actions. This is referred to as the retargeting problem. In some cases, perception of actual actions is not necessary, as long as the robot can detect the state changes that result from the task demonstration and learn the task based on that information (Baisero et al 2015; Mollard et al 2015).
- For tasks performed by a robot with guidance from human, the human teacher demonstrates a task to the robot by guiding it through the task. This mode of teaching bypasses the retargeting problem but requires the teacher to have a good understanding of the robot's action capabilities. The guidance to the robot can be provided in various ways, from kinesthetic movements to verbal instructions:
- *Kinesthetic guidance* involves physically holding the robot and moving its manipulators to perform the task (e.g., Akgun et al 2012; Phillips et al 2016).
- *Natural language guidance* involves instructing the robot on what to do to perform the task. Mohan et al. (2012) and She et al. (2014), for instance, use step-by-step language instructions to teach new tasks to a robot.
- *Multimodal language guidance* uses multimodal instructions (speech and gestures) to guide a robot through the task.
- *Gestures* often serve to reference parts of the environment.
- *Joystick-based guidance* involves driving the robot and triggering prespecified actions with the help of a special device to perform the task.

- *GUI-based guidance* employs a graphical interface to control the robot and trigger prespecified actions to perform the task.

In **task specifications**, alternative ways of achieving a task are directly specified by the teacher in a format compatible with the robot's task representation. For example, for a partially ordered plan representation, the teacher might verbally state:

> First *bring all of the ingredients and tools* to the kitchen counter (in any order).
> Second, pour *all the dry ingredients* into the mixing bowl (in any order).

While teaching by demonstration inevitably involves a particular ordering of the actions and hence requires multiple demonstrations to capture order invariances, direct specification provides an efficient way to provide the same information. Similarly, if the representation is a program, the user can directly specify loops or conditionals by literally writing a program or verbally specifying those with instructions like:

> Insert a toothpick into the center of the cake. If it comes out clean, take out the cake; otherwise continue to bake. Alternatively, for each cup on the muffin pan, pour until ¾ full.

Similarly, direct specification of task goals involves the teacher directly indicating parts of the world state that are relevant or irrelevant to the robot's task, rather than trying to exemplify variations of positive and negative goal states. For example, the teacher may verbally describe the desired goal state when teaching a robot to set up a table:

> The red bowl should be on top of the green plate and the napkin should be to the right of the plate.

Task specifications can be provided through natural language or graphical user interfaces (GUIs):

- *Natural language specifications* involve the use of language to directly specify certain properties or constraints about the task representation. For example, Cantrell et al. (2012) use natural language to specify precondition and effects of action schemas for task planning.
- *GUIs* can be used to specify properties or constraints about a task being taught to a robot.

Humans often combine these two means of communicating task knowledge (demonstrations and specifications). For example, a teacher might demonstrate the physical act of adding different ingredients to the mix in a particular order as part of teaching a recipe, while verbally specifying partial ordering constraints by saying: "Add all dry ingredients in any order." Similarly, a person might set up the table themselves to show an example of how they want the table to be set up, but then specify invariance constraints by saying "The salt and pepper can be anywhere in the center area of the table." In any case, regardless of the specific form of interaction, during learning, symbolic representations of human inputs (e.g., GUI, natural language) need to be tightly grounded to the robot's internal representations of perception and action.

Task-Relevant Background Knowledge and Capabilities

When we speak about a robot learning new tasks, we often assume that the robot has the necessary background knowledge and capabilities. The ability to perform new tasks, however, might equally be due to the acquisition of other knowledge or capabilities, not solely due to newly acquired task knowledge. Hence, the ability to acquire these different kinds of background knowledge and capabilities through interactions is also highly relevant for task learning. For example, the capabilities of a robot that already knows the task of sorting objects, based on different properties, can be expanded by the acquisition of new perceptual capabilities (e.g. the ability to detect new object properties) or new action capabilities (e.g. the ability to manipulate new types of objects). We identify the following four types of knowledge and capabilities relevant for task learning:

1. *Perception capabilities* refer to the ability to perceive the task-relevant environment and interpret human language, including:
   - state and actions of humans,
   - state, properties, and affordances of objects,
   - scene composition (surfaces, objects, humans, and their relationships),
   - changes of the state that occurred to the environment, and
   - state of communication such as communicative intent and focus of attention.
2. *Action capabilities* refer to lower-level policies that control a robot's actuators to carry out tasks and/or communicate with humans. These include capabilities that allow robots to:
   - navigate the environment,
   - manipulate objects in the environment, and
   - communicate with humans in the environment.
3. *Linguistic knowledge* concerns the meanings of words and phrases. For physical robots (which need to sense from and act upon the physical world, as opposed to the symbolic world), this knowledge cannot be purely symbolic as in a dictionary or thesaurus. Word semantics need to be grounded to the robot's sensorimotor skills.
4. *World knowledge* captures any other task-relevant knowledge about the world and how the world works, including.
   - *Facts* about the world and the robot's task environment: "My owner's name is Katie Smith" or "I was built in 2017."
   - *Common-sense knowledge* allows a robot to reason about how to achieve task goals (Tenorth and Beetz 2009; Al-Moadhen et al. 2013): to "boil the water," the water must first be placed in the boiling pot.
   - *Action knowledge* captures the existing knowledge about subtasks and subgoals previously acquired or learned. Formal action models capture preconditions and effects of actions (Fox and Long 2003). Preconditions specify world states in which the action is applicable; effects specify the expected changes to the world state.
   - *Domain knowledge* corresponds to information specific to a particular task environment or user that a robot needs to perform its task. For example, a robot that performs object deliveries to hotel rooms needs to have a map specific to the hotel within which it is deployed, with room numbers annotated on the map.

Some of these knowledge and capabilities can be programmed into a robot, they can also be acquired through interactions with humans although the means of acquisition is less clear than that for task knowledge.

Forms of Interaction for Learning Task-Relevant Knowledge or Capabilities

The types of interactions that support acquiring task-relevant knowledge and capabilities are similar to those involved in learning the task itself. As shown in Table 1, the forms of interaction often depend on the kind of knowledge or capabilities to be learned. For example, to help train the robot's visual perception capabilites, the teacher may use language descriptions and also show target objects from different angels. To acquire the navigation map, tele-operation (e.g., through joystick guidance) can be employed as well as language descriptions. Acquisition of low-level action knowledge (e.g., lower-level policies to generate trajectories) may benefit from kinesthetic demonstration whereas higher-level task knowledge (e.g., partial orderings) may best benefit from language instructions. Linguistic knowledge certainly involves the use of language, which is often combined with deictic gestures or action demonstrations because the semantics of words need to be grounded to visual perception and the change of state in the physical world.

Table 1 Example forms of interaction for different types of knowledge

Knowledge	Example forms of interaction
Perception capabilities	• Natural language and deictic gestures to teach labels of objects and indicate their relations • Natural language to specify object affordances
Action capabilities	• Kinesthetic demonstration to teach low-level control policies to generate arm trajectories or navigation strategies
Linguistic knowledge	• Natural language combined with deictic gestures to teach nouns and adjectives • Natural language combined with action demonstration to teach action verbs
World knowledge	• Natural language to specify order constraints among sub-actions • Natural language to specify causality (i.e., precondition and effect) of an action • Demonstrations performed by the human to show how basic actions/verbs change the state of the world • Joystick guidance to build a map of the robot's environment for navigation

Open Questions in Enabling Effective Task Learning Interaction

Teaching presupposed task-relevant knowledge

While previous work has investigated the acquisition of many types of task-related knowledge and capabilities, the acquisition of common sense world knowledge in task learning has largely gone unexplored. In human-to-human interactions, knowledge about the world and the domain is often presupposed. The speaker and the listener believe they share the same kind of world knowledge, so it does not need to be explicitly stated. However, in human-robot interactions, huge discrepancies in world knowledge can exist between humans and robots. Often, the robot does not have sufficient background knowledge to learn a new task. Thus human teachers need to be able to assess what kind of background knowledge the robot has and to teach the robot background knowledge pertinent to the task at hand. What sort of background knowledge must a robot possess? For example, the result states of basic action verbs are not usually specified, and humans naturally take them for granted. Existing verb semantic models (such as Verbnet, FrameNet) and preexisting knowledge bases (e.g., Google's Knowledge Graph, Freebase, among others) offer sources of information, but not at the level of detail required for the robot to understand the very basic principles about the conditions for their actions (e.g., "put A on B" requires A generally smaller and lighter than B) and how their actions may change the world (e.g., slicing a cucumber may lead to the change of the shape, size, and pieces of the cucumber).

Thus, it is important to understand *what a human must teach a robot about the domain of a task.* Some background knowledge (e.g., time as duration, units of time, and time relations) may

be best taught once for many domains, but much human knowledge is domain specific. Learning domain-specific knowledge will lead to a whole new set of research questions:

- How does the human know what knowledge the robot (e.g., sub-actions) has so that it can be used to teach new tasks?
- During task learning, what signals indicate the lack of background knowledge and clarify when human teaching is required?
- How can existing resources be leveraged to acquire the correct level of background knowledge during teaching?
- What level of granularity should background knowledge be taught by a human?
- How should background knowledge be represented and used for effective reasoning and inference?

Combining different forms of interaction for task learning/teaching

Most previous work on task learning has focused on a single form of interaction for teaching. Except for a limited few (Rybski et al. 2007; Niekum et al. 2015; Kirk et al., 2016; Mohseni-Kabir et al. 2018), techniques that combine language, dialogue, and action demonstration to teach complex tasks are in critical need. As discussed above, different forms benefit different types of knowledge. In addition, as the situation changes (e.g., the lighting situation changes from being good to poor), the form of interaction may need to adapt (e.g., switch from visual demonstration to language instruction). Thus we need to know *how to seamlessly combine and adapt different forms of teaching to enable the most effective teaching*. Is combining and adapting a problem for human teachers or a problem for robot learners? The answer is both.

Teaching Humans How to Teach Robots

After working with a robot, an experienced human teacher—in our scenario involving Mia, this would be David, the employee from robotic manufacturer—should be able to discern which form of interaction is necessary to teach a specific kind of knowledge to meet specific circumstances. Experienced human teachers should know when to provide a particular kind of feedback (such as reward or punishment) so that the robot can learn from such feedback and adjust its behaviors to maximize future rewards. Experienced human teachers may also apply scaffolding, intentionally vary the situation, and design different experiences for the robot to learn the task and aspects associated with the task.

Thus, similar to the setting in human skill learning, human teachers' behaviors and experience have a massive influence on the success of robot task learning. *How, then, should we train a new generation of human partners/teachers, so that robots can be effectively taught through their collaborations*?

Enabling Robots to Engage Proactively in Learning

We cannot expect that every human partner will be capable of identifying and employing the most effective means to teach the appropriate kind of knowledge. Thus a robot needs to be able to share the burden of selecting effective strategies. A crucial issue, not yet studied, is: *How can a robot be made to be aware of its own learning situation—one in which it is capable of communicating to the human its limitations and proactively requesting the right kind of teaching from the human?*

Capabilities to Perceive the Environment and Human Inputs

The ability to perceive the environment and human inputs as well as to infer current task-related states and communicative states is fundamental to interactive task learning. A robot must be able to recognize task-relevant objects in the environment, the change of the environment cause by an action, task demonstration from humans, and verbal and non-verbal human

communicative behaviors. It must also be able to infer human intent, interpret instructed actions and their involved objects, and derive task structures by grounding language to perception.

Visual Perception

Performing or learning tasks inevitably requires an understanding of objects and environments integral to the tasks. This includes objects, their properties, fluents (i.e., attributes which can potentially change), and relations, as well as an understanding of external actions and how they may have changed the perceived state of the physical world. As humans can perform actions to teach robots and apply nonverbal modalities (such as deictic gestures, iconic gestures, and gaze directions) to facilitate communication, the robot should also have the capability to recognize the state and actions of its human partners.

Acquiring perceptual capability has been the main research goal for the computer vision community. Most of the learning algorithms for perception are trained offline and rely on large training data for object recognition, activity recognition, etc. Recent years have seen significant progress on recognition of common objects from static scenes (e.g., images) (Grauman and Leibe 2011). However, in a dynamic scene such as is encountered in task learning, object tracking and human action recognition still face many challenges (for recent reviews, see Aggarwal and Ryoo 2011; Sargano et al. 2017). In addition, during task learning, it is likely that neither relevant computer vision models nor sufficient data are available. Thus, it is critical for the robot to continuously acquire new models for object recognition through interaction with its human teacher. The teacher can use language to provide the name, the object type, and related properties to a perceived object in the environment. However, the ability to efficiently learn a generalized model (e.g., for object recognition) that can be applied in new situations still faces many challenges. Some key research questions include:

- *How can a robot learn reliable models based on a small number of examples with limited human supervision during interaction?*
- *How can it transfer and adapt models learned from previous experience to a new situation (e.g., transfer learning), perhaps with limited human intervention?*

Language Understanding

Language serves as a main mode of interaction in interactive task learning. From a human's linguistic utterance, the robot needs first to understand the underlying intent of the teacher (e.g., whether it is to teach the robot a new step or to correct the robot's current understanding of a learned step/action). When a referring expression is involved, the agent needs to understand what entities, from the interaction discourse or the shared environment, are being referenced. When the utterance describes some task steps, the agent needs to understand what actions are specified and what participants are involved (e.g., agent, patient, instrument, source, destination, etc.). The robot also needs to be able to extract any information from the utterance that specifies preconditions, effects, and constraints (e.g., temporal orders) associated with actions and tasks. To help achieve the above-mentioned abilities, recent advances in natural language processing—particularly in syntactic parsing, semantic processing, and discourse processing—can be applied (Jurafsky and Martin 2008). In the event that the robot cannot successfully understand human utterances, dialogue can clarify human intent and disambiguate different interpretations of linguistic expressions.

In situated interaction, language communication is often accompanied by other nonverbal modalities, such as gesture. Deictic gestures (e.g., pointing to objects in the environment) and iconic gestures (e.g., waving hello or indicating an action or a particular type of object) are vital to an understanding of the teacher's intent. Pointing gestures are essential to task instruction because the array of objects in a task (which may be difficult to describe verbally) lead to the need to point at them rather than rely solely on language descriptions. Matuszek et al. (2014), for example, combine language and gesture to interpret directives in human-robot interaction.

Speech communication is perhaps one of the most natural means of interaction in task learning. Speech recognition has made significant progress over the last decade. More recently, advances in deep neural networks have made it possible for machines to achieve recognition performance on par with human performance. At the time of writing this article (June 2017), Google reported a 4.9% word error rate in recognition while human performance is estimated to be around 4% word error rate (Saon et al. 2016). Although encouraging, these results were often obtained based on offline benchmark data. Thus, it is not clear whether the same performance can be attained in a real-time, interactive, and unconstrained environment. *How can recent advances in speech recognition be successfully applied to real-time interactive systems for task learning*?

Unlike traditional natural language processing, to enable communication with physical robots, linguistic knowledge must go beyond pure symbolic representations, as in a dictionary or thesaurus. The meanings of words need to be grounded to the robot's internal representations that are connected with sensors and effectors. Concrete nouns, for instance, need to be grounded to the types of objects or object attributes perceived from the environment (e.g., color words grounded to color histograms). Adjectives are often grounded to the perceived attributes (e.g., the size of the bounding boxes, the weights of an object) and fluents (e.g., door open or closed, box open or closed). Verbs need to be grounded to the underlying action representations, which can be accessed by the robot's control system to plan and execute the corresponding actions. On one hand, existing knowledge of grounded word semantics will be applied to ground language to perception and action (discussed in the next section). On the other, as new words are often encountered during interaction, they should be acquired continuously through situated interaction (Mohan et al. 2012). When a situation changes (e.g., a change in the environment), the learned word representation may not fit the new situation (e.g., a lighting change in the environment may affect grounded word models for color words). Thus, it is important word models need to be adaptable to new situations (Liu and Chai 2015; Thomason et al. 2015).

Grounding Language to Perception

The capability to ground human language to the perceived physical environment is particularly important for task learning. Suppose a human teaches the robot how to boil water by demonstrating to the robot how to achieve this task through step-by-step instructions: *pick up the pot, fill the pot with water, boil the water,….* To learn how to perform this task, the robot must first understand what perceived objects are involved in each step of instruction by grounding the arguments of action verbs, such as the noun phrase *the pot*, to the perceived objects in the environment.

This task of grounding language to perception of the environment has received an increasing amount of attention (Mooney 2008; Tellex et al. 2011; Krishnamurthy and Kollar 2013; Yu and Siskind 2013; Matuszek et al. 2014; Tellex et al. 2014; Yang et al. 2016). Most previous approaches first process language and vision separately, and then integrate the partial results together. In a dynamic scene with ongoing activities, computer vision algorithms still have difficulty reliably recognizing and tracking objects and actions; this leads to a bottleneck in grounding language to vision. Recent deep learning approaches directly fuse raw features from language and vision and have achieved state-of-the-art empirical results on applications such as caption generation from images/videos and visual question answering. These approaches, however, require a large amount of training data. *To integrate language and vision in the context of interactive task learning, what would be the optimal architecture?*

Another line of recent work has explored causality modeling for action verbs (Gao et al. 2016). Here the idea is that knowledge of how concrete action verbs (e.g., cut, slice, pick up, etc.) might alter the world can drive visual detection. For example, from the directive "slice the cucumber," knowledge about expected changes to the cucumber will provide high-level guidance to look specifically for grounded objects with relevant features (or the change of

features) in the visual scene. Recent work has also explored common-sense physical knowledge about objects that are implied by action verbs (Forbes and Choi 2017). For example, "he threw the ball" implies that "he" is bigger, heavier, and faster than "the ball." This kind of implicit knowledge can potentially provide additional cues to ground language to perception.

Capabilities to Act and Communicate

Enabling a robot to learn new tasks requires action capabilities to carry out task-related actions as well as actions that facilitate communication. These capabilities span a wide range, from navigation and manipulation to communication.

Task-Related Actions and Grounding Language to Action Representation

A robot's action capabilities can be based on manually designed and tuned controllers, as well as policies learned from human demonstrations or through reinforcement learning. In some robotic applications, it is essential for the robot already to possess all of the action capabilities needed to complete a task. For example, previous work in the robotics community aimed to translate natural language instructions to robotic operations (Kress-Gazit et al. 2007; Spangenberg and Henrich 2015), but they were not designed for learning new actions or tasks. In other cases, tasks and actions can be learned simultaneously. For example, Mohan and Laird (2014) developed a system where a robot can learn a hierarchical representation of a new task based on linguistic interaction with the human. Similarly, Liu et al. (2016) applied grammar induction to learn a hierarchical and/or graph representation for a new task from human's language instructions and visual demonstrations.

To support action learning from language instructions, recent work has begun to explore the connection between semantics of concrete action verbs and action planning (She et al. 2014; Misra et al. 2016) and explicitly represented grounded verb semantics as desired goal states of the physical world as a result of the corresponding actions. Such representations are learned based on example actions demonstrated by the human. For example, a human may teach the robot how to "boil water" by issuing step-by-step language instructions which the robot knows how to perform: "move to the kettle, grasp the kettle, move to the stove, …" By following these steps, the robot will experience the change of the physical world. By capturing the differences between the goal state and the initial state, the robot is able to acquire the semantics of the verb frame "boil (water)". Once acquired, these grounded representations will allow the robot to interpret verbs/commands issued by humans in new situations and apply planning to execute actions. One limitation of previous work is that the algorithms were mainly developed based on simulations (e.g., simulated Baxter robots). Except for a few (e.g., She and Chai 2017), uncertainties from the environment were largely ignored. However, the world is full of uncertainties at various levels: from motion planning to perception and language grounding. To extend task learning from language instructions to the physical world, it is paramount to address *how to integrate uncertainties at multiple levels together, so that new actions associated with concrete action verbs can be learned*.

Verbal and Nonverbal Communicative Action

Separate from its task-related actions, a robot will need to perform communicative actions to facilitate its learning/teaching interactions. In situated interaction, both verbal and nonverbal modalities are available for the robot to communicate to its human partner. This communication includes such capabilities as:

- Generating speech and deictic gestures to confirm understanding of instructions or refer to objects in the environment (Fang et al. 2015),

- Generating gaze direction, communicative head gestures (e.g., nodding and shaking head), or facial expressions (confused or confident face) to respond to human input at different points in the interaction (Holroyd et al. 2011), or
- Displaying visualizations of learned concepts to enable humans to inspect them.

In particular, the embodiment of a physical robot can take advantage of nonverbal modalities (e.g., gaze and gesture) for efficient communication. The robotics community has learned from psychologists that gazing at others and at objects in the environment are quintessential human behaviors. Gaze that is used to convey information to a collaborator is referred to as *social gaze*. Gaze at a collaborator functions to gather attention from the other, to indicate social presence, and to indicate attention to the individual (e.g., turn taking via gaze aversion). Gaze at objects serves to indicate what one is paying attention to, is about to point at, what one intends to do next, or to indicate that what another has focused on should now be the object of mutual gaze. Collaborators use gaze information to assess how well their partners comprehend their collaborations as well as to assess the collaborators' level of continued engagement (Rich et al. 2010). Every one of these abilities is valuable in task learning, as they enable the assessment of how the learning is progressing, whether the learner is looking in the right direction, and what the teacher intends for the learner to do. Gestures also have similar effects in coordinating interaction, establishing shared attention, and providing feedback. Proxemics, which models the stance of individuals to others and how they approach one another, can be significant in tasks because where the learner stands in performing a task may be crucial. *How to effectively generate verbal and nonverbal communicative behaviors to facilitate task learning remains an important focus for research research.*

Capabilities to Manage and Coordinate Interaction

Managing interactions between humans and robots is critical to support task learning/teaching. At any point in the interaction, robots need to decide what to do next based on interaction history, current situation, and learning goals. These decisions can be made by following simple decision rules that are manually crafted or interaction policies that are learned from experience.

Interaction Management and Active Learning

Decades of work on dialogue modeling are relevant for interactive task learning. Different approaches have been developed, for example, driven by intention and collaboration (e.g., Grosz and Sidner 1986; Rich and Sidner 1998), based on information states (Larsson and Traum 2000) or interaction policies learned from reinforcement learning (Kaelbling et al., 1996; Young et al. 2013). Despite recent progress, dialogue modeling remains a significant challenge. Dialogue models need to be able to accommodate interruption, turn taking, and other dialogue behaviors, which neither the intention-based nor information state approach have successfully addressed, but are essential in task instruction.

Specifically to learn new tasks, active learning has been shown to be an important component that contributes to effective interaction management. Most work on task learning assumes a learner that passively receives information from the teacher. However, humans are often suboptimal in their teaching when the learner is passive. One line of work explores active task learning whereby the learner actively requests specific information that it evaluates as most useful. Active questioning enables much more efficient learning. For example, Chao et al. (2010) and Cakmak et al. (2010) demonstrated that an active learner which requests labels (positive/negative) for specific instances of a task goal outperforms a passive learner taught by examples selected by naïve human teachers. In particular, Cakmak and Thomaz (2012) identified three types of questions that can be used by a human/robot student as part of active task learning: (a) demonstration queries asking for a full or partial demonstration of the task, (b) label queries asking whether an execution is correct, and (c) feature queries asking about the relevance or invariance of specific aspects of the task. Recent work by She and Chai (2017)

extended this question/answer style of interaction and applied reinforcement learning to acquire an interaction policy that allows the robot to handle noisy environment and learn new verbs and corresponding actions.

To improve interactive task learning, we need to know how to engage in a full range of interaction that can incorporate active learning with other communicative goals (e.g., clarification and disambiguation) to acquire more reliable models of skills.

Extra Collaborative Effort and Transparency

In human-human task learning, background knowledge is largely presupposed. The speaker and the listener believe they share the same kind of background knowledge, so it does not need to be explicitly stated. In addition, human partners often share similar perceptual capabilities. There is basic common ground where human teacher/learner can ground to without much effort.

In human-robot task learning, however, there are huge discrepancies in background knowledge between humans and robots. Often the robot does not have sufficient background knowledge to learn a new task. Furthermore, although co-present in a shared environment, humans and robots have mismatched capabilities in reasoning, perception, and action. Their representations of the shared environment and joint tasks can be significantly misaligned. A significant challenge in interaction and communication with cognitive robots involves the lack of common ground and discrepancies in the human's mental model of what a robot knows and is capable of doing. Previous work (Chai et al. 2016) has shown that to bridge the gap and strive for a common ground of shared representations, humans and robots need to make extra effort to establish common ground. This extra collaborative effort in interaction not only has implications in algorithms for language grounding, but also affects interaction management.

Transparency plays an important role in achieving common ground and promoting accurate mental models during interaction. For example, Thomaz and Breazeal (2006) show that natural transparency mechanisms like gaze can steer the human's behavior while demonstrating a task. Pejsa et al. (2014) used facial expressions to provide transparency about dialog uncertainties. Alexandrova et al. (2015) employed interactive visualizations of learned actions to enable teachers to verify tasks that are learned from a single demonstration and correct any mistakes they detect. Guha (2016) used pointing to communicate the robot's understanding of a referenced object, and Whitney et al. (2016) used heat map visualizations and facial expressions to communicate uncertainty about its inference. Recent work by Hayes and Shah (2017) allows a robot to automatically generate verbal description of its learned policy (i.e., which actions it takes in which contexts).

To enable common ground for effective task learning, there are many research questions to pursue:

- How can an agent make its internal representations (e.g., causal-effect relations) transparent to the human?
- How can an agent explain its autonomy or decision so that the human can better understand the agent's capabilities and limitations?
- What are the mechanisms to manage interaction so that it can encourage human's collaborative behaviors and simultaneously create more collaborative behaviors from the robot?

Conclusions

To fully support interactively teaching robots new tasks through various means, many challenges and open questions remain as discussed above. While the scenarios in the introduction section focused on in-home settings, teaching robots new tasks is applicable in many situations, especially ones with highly structured environments. Already robots are being trained by people in ad-hoc ways to work in manufacturing assembly lines (cf. the Baxter robots

of Re-Think robotics). Robots working in warehouses are largely programmed by hand, but it is not difficult to envision the need for them to be taught tasks by human co-workers. The same applies to robots in the service industry (e.g. hotel helpers).

One key challenge in task learning that has not been discussed above is evaluation. Evaluation has long been a critical and difficult issue in interactive systems because many confounding factors are involved. In the context of interactive task learning, many new questions arise:

- How do we know the task is learned?
- What additional metrics should be used to evaluate the success of task acquisition beyond traditional metrics for evaluating interaction (e.g., efficiency and task completion)?
- What are reasonable baselines and upper bound (e.g., human-human interaction)?
- How do researchers conduct longitudinal studies and evaluation?
- What kinds of products are available that may make longitudinal evaluation (e.g., putting robots in people's house) possible?

While this paper is mainly about task learning where humans serve as teachers and robots serve as learners, it is not difficult to imagine that a well-trained and capable robot can also teach humans new tasks. In the intelligent tutoring world, computer programs have been teaching humans in various ways for more than three decades. Virtual agents teach humans all sorts of tasks, from turbine engine operation (Rickel and Johnson 2000) to negotiation (Gratch et al. 2015) to cross cultural communication (Johnson and Zaker 2012). The idea that robots might teach humans has received relatively little attention, perhaps in part due to the lack of capabilities. Robots are not yet teachers, but for many tasks, from doing experiments, to manipulation of heavy equipment, the form factor of a robot will be useful in ways that computer programs and virtual agents are not. As robots become more capable, the teacher/learner role reverse is foreseeable in the future, which will bring new research challenges and opportunities.

References

Aggarwal, J. K., and M. S. Ryoo. 2011. Human Activity Analysis: A Review. ACM Computing Surveys. https://dl.acm.org/citation.cfm?doid=1922649.1922653. (accessed Oct. 4, 2017). [09]

Akgun, B., Cakmak, M., Yoo, J. W., & Thomaz, A. L. (2012, March). Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective. In Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (pp. 391-398). ACM.

Al-Moadhen, A., R. Qiu, M. Packianather, Z. Ji, and R. Setchi. 2013. Integrating Robot Task Planner with Common-Sense Knowledge Base to Improve the Efficiency of Planning. *Procedia Comput. Sci.* **22**:211–220. [09]

Alexandrova, S., Z. Tatlock, and M. Cakmak. 2015. Roboflow: A Flow-Based Visual Programming Language for Mobile Manipulation Tasks. In: Robotics and Automation (ICRA), Proceedings of the 2015 IEEE International Conference, Institute of Electrical and Electronics Engineers. http://ieeexplore.ieee.org/document/7139973/. (accessed Oct. 4, 2017). [09]

Argall, B. D., S. Chernova, M. Veloso, and B. Browning. 2009. A Survey of Robot Learning from Demonstration. *Rob. Auton. Syst.* **57**:469–483. [09]

Baisero, A., Mollard, Y., Lopes, M., Toussaint, M., & Lütkebohle, I. (2015, September). Temporal segmentation of pair-wise interaction phases in sequential manipulation demonstrations. In Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on (pp. 478-484). IEEE.

Cakmak, M., C. Chao, and A. L. Thomaz. 2010. Designing Interactions for Robot Active Learners. *IEEE Trans. Auton. Ment. Dev.* **2.2**:108–118. [09]

Cakmak, M., and A. L. Thomaz. 2012. Designing Robot Learners That Ask Good Questions. In: Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction. https://dl.acm.org/citation.cfm?id=2157693. (accessed Oct. 4, 2017). [09]

Cantrell, R., K. Talamadupula, P. Schermerhorn, et al. 2012. Tell Me When and Why to Do It! Run-Time Planner Model Updates via Natural Language Instruction. In: Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction. https://dl.acm.org/citation.cfm?id=2157840&CFID=991822793&CFTOKEN=69761956. (accessed Oct. 4, 2017). [09]

Chai, J. Y., R. Fang, C. Liu, and L. She. 2016. Collaborative Language Grounding Towards Situated Human-Robot Dialogue. *AI Magazine* **37**:32–45. [09]

Chao, C., M. Cakmak, and A. L. Thomaz. 2010. Transparent Active Learning for Robots. In: Human-Robot Interaction (HRI), Proceedings of the 5th ACM/IEEE International Conference. https://dl.acm.org/citation.cfm?id=1734562&CFID=991822793&CFTOKEN=69761956. (accessed Oct. 4, 2017). [09]

———. 2011. Towards Grounding Concepts for Transfer in Goal Learning from Demonstration. In: Development and Learning (ICDL), Proceedings of the 2011 IEEE International Conference. http://ieeexplore.ieee.org/document/6037321/. (accessed Oct. 4, 2017). [09]

Ekvall, S., and D. Kragic. 2008. Robot Learning from Demonstration: A Task-Level Planning Approach. *Int.J . Adv. Robot. Syst.* **5**:223–234. [09]

Fang, R., M. Doering, and J. Y. Chai. 2015. Embodied Collaborative Referring Expression Generation in Situated Human-Robot Dialogue. In: Proceedings of the 10th ACM/IEEE Conference on Human-Robot Interaction (HRI). https://dl.acm.org/citation.cfm?id=2696467. (accessed Oct. 4, 2017). [09]

Forbes, M., and Y. Choi. 2017. Verb Physics: Relative Physical Knowledge of Actions and Objects. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). http://www.aclweb.org/anthology/P17-1025. (accessed Oct. 4, 2017). [09]

Fox, M., and D. Long. 2003. Pddl 2.1: An Extension to Pddl for Expressing Temporal Planning Domains. *J. Artif. Intell. Res.* **20**:61–124. [09]

Gao, Q., M. Doering, S. Yang, and J. Y. Chai. 2016. Physical Causality of Action Verbs in Grounded Language Understanding. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). http://www.aclweb.org/anthology/P16-1171. (accessed Oct. 4, 2017). [09]

Gratch, J., D. DeVault, G. Lucas, and S. Marsella. 2015. Negotiation as a Challenge Problem for Virtual Humans. In: 15th International Conference on Intelligent Virtual Agents, ed. W. P. Brinkman et al. http://ict.usc.edu/pubs/Negotiation%20as%20a%20Challenge%20Problem%20for%20Virtual%20Humans.pdf. (accessed Oct. 4, 2017). [09]

Grauman, K., and B. Leibe. 2011. Visual Object Recognition. In: Synthesis Lectures on Artificial Intelligence and Machine Learning, ed. R. J. Brachman and T. G. Dietterich, pp. 1–181, vol. 5. Williston, VT: Morgan and Claypool Publishers. [09]

Grosz, B., and C. L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Comput. Linguist.* **12**:175–204. [09]

Guha, A. I. 2016. Towards Meaningful Human-Robot Collaboration on Object Placement. Undergraduate thesis, Department of Computer Science, Brown University. [09]

Hayes, B., and J. Shah. 2017. Improving Robot Controller Interpretability and Transparency through Autonomous Policy Explanation. In: ACM International Conference on Human-Robot Interaction. https://dl.acm.org/citation.cfm?id=3020233&CFID=991822793&CFTOKEN=69761956. (accessed Oct. 4, 2017). [09]

Holroyd, A., C. Rich, C. L. Sidner, and B. Ponsler. 2011. Generating Connection Events for Human-Robot Collaboration. In: Proceedings of Ro-Man 2011, 20th IEEE International Symposium on Robot and Human Interactive Communication. http://ieeexplore.ieee.org/document/6005245/. (accessed Oct. 4, 2017). [09]

Huang, J., and M. Cakmak. 2017. Code3: A System for End-to-End Programming of Mobile Manipulator Robots for Novices and Experts. In: Proceedings of the Twelfth Annual ACM/IEEE International Conference on Human-Robot Interaction. https://dl.acm.org/citation.cfm?id=3020215. (accessed Oct. 4, 2017). [09]

Johnson, W. L., and S. B. Zaker. 2012. The Power of Social Simulation for Chinese Language Teaching. Alelo Corporation Memo. https://www.alelo.com/wp-content/uploads/2014/06/TCLT7_Presentation_Johnson_Zakar_May2012.pdf. (accessed Oct. 4, 2017). [09]

Jurafsky, D. and Martin, J. 2008. Speech and Langauge processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition, Prentice Hall.

Kirk, J., Mininger, A., Laird, J. 2016. Learning task goals interactively with visual demonstrations. *Biologically Inspired Cognitive Architectures. New York, New York*

Kaelbling, L. P., M. L. Littman, and A. W. Moore. 1996. Reinforcement Learning: A Survey. *J. Artif. Intell. Res.* **4**:237–285. [09]

Kress-Gazit, H., G. E. Fainekos, and G. J. Pappas. 2007. From Structured English to Robot Motion. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). http://ieeexplore.ieee.org/abstract/document/4398998/. (accessed Oct. 4, 2017). [09]

Krishnamurthy, J., and T. Kollar. 2013. Jointly Learning to Parse and Perceive: Connecting Natural Language to the Physical World. *Trans. Assoc. Comput. Linguist.* **1**:193–206. [09]

Larsson, S., and D. R. Traum. 2000. Information State and Dialogue Management in the Trindi Dialogue Move Engine Toolkit. *Nat. Lang. Eng.* **6**:323–340. [09]

Liu, C., and J. Y. Chai. 2015. Learning to Mediate Perceptual Differences in Situated Human-Robot Dialogue. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI Press. http://www.cse.msu.edu/~jchai/Papers/AAAI2015.pdf. (accessed Oct. 4, 2017). [09]

Liu, C., S. Yang, S. Saba-Sadiya, et al. 2016. Jointly Learning Grounded Task Structures from Language Instruction and Visual Demonstration. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, TX. https://aclweb.org/anthology/D/D16/D16-1155.pdf. (accessed Oct. 4, 2017). [09]

Matuszek, C., L. Bo, L. Zettlemoyer, and D. Fox. 2014. Learning from Unscripted Deictic Gesture and Language for Human-Robot Interactions. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI Press. https://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8327. (accessed Oct. 4, 2017). [09]

Misra, D. K., J. Sung, K. Lee, and A. Saxena. 2016. Tell Me Dave: Context Sensitive Grounding of Natural Language to Manipulation Instructions. *Int. J. Rob. Res.* **35**:281–300. [09]

Mohan, S., and J. E. Laird. 2014. Learning Goal-Oriented Hierarchical Tasks from Situated Interactive Instruction. In: Proceedings of the 28th AAAI Conference on Artificial

Intelligence. https://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8630. (accessed Oct. 4, 2017). [09]

Mohan, S., A. Mininger, J. Kirk, and J. E. Laird. 2012. Acquiring Grounded Representations of Words with Situated Interactive Instruction. *Adv. Cog. Syst.* **2**:113–130. [09]

Mohseni-Kabir, A., C. Li, V. Wu, D. Miller, B. Hylak, S. Chernova, D. Berenson, C. Sidner, and C. Rich. 2018. Simultaneous Learning of Hierarchy and Primitives (SLHAP) for Complex Robot Tasks. *Autonomous Robotics*. [09]

Mollard, Y., Munzer, T., Baisero, A., Toussaint, M., & Lopes, M. (2015, September). Robot programming from demonstration, feedback and transfer. In Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on (pp. 1825-1831). IEEE.

Mooney, R. 2008. Learning to Connect Language and Perception. In Proceedings of 23rd AAAI Conference on Artificial Intelligence.

Niekum, S., S. Osentoski, G. D. Konidaris, et al. 2015. Learning Grounded Finite-State Representations from Unstructured Demonstrations. *Int. J. Rob. Res.* **34**:131–115. [09]

Pardowitz, M., S. Knoop, R. Dillmann, and R. D. Zollner. 2007. Incremental Learning of Tasks from User Demonstrations, Past Experiences, and Vocal Comments. *IEEE Trans. Syst. Man. Cybern. B Cybern.* **37**:322–332. [09]

Pejsa, T., D. Bohus, M. F. Cohen, et al. 2014. Natural Communication About Uncertainties in Situated Interaction. In: ACM Intl. Conf. on Multimodal Interaction (ICMI). https://dl.acm.org/citation.cfm?id=2663249. (accessed Oct. 4, 2017). [09]

Phillips, M., Hwang, V., Chitta, S., & Likhachev, M. (2016). Learning to plan for constrained manipulation from demonstrations. Autonomous Robots, 40(1), 109-124.

Rich, C., B. Ponsler, A. Holroyd, and C. L. Sidner. 2010. Recognizing Engagement in Human-Robot Interaction. In: Proceedings of the ACM Conference on Human-Robot Interaction. https://dl.acm.org/citation.cfm?id=1734580. (accessed Oct. 5, 2017). [09]

Rich, C., and C. L. Sidner. 1998. Collagen: A Collaborative Manager for Software Interface Agents. *User Model. User-adapt. Interact.* **8**:315–350. [09]

Rickel, J., and W. L. Johnson. 2000. Task-Oriented Collaboration with Embodied Agents in Virtual Worlds. In: Embodied Conversational Agents, ed. J. Cassell et al., pp. 95–122. Cambridge, MA: MIT Press. [09]

Rybski, P. E., K. Yoon, J. Stolarz, and M. M. Veloso. 2007. Interactive Robot Task Training through Dialog and Demonstration. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction. https://dl.acm.org/citation.cfm?id=1228724&CFID=816340353&CFTOKEN=71046559. (accessed Oct. 5, 2017). [09]

Saon, G., T. Sercu, S. Rennie, and H.-K. J. Kuo. 2016. The Ibm 2016 English Conversational Telephone Speech Recognition System. In: Proceedings of International Speech Communication Association Annual Conference. https://arxiv.org/pdf/1604.08242.pdf. (accessed Oct. 5, 2017). [09]

Sargano, A. B., P. Angelov, and Z. Habib. 2017. A Comprehensive Review on Handcrafted and Learning-Based Action Representation Approaches for Human Activity Recognition. *Applied Sci.* **7**:doi:10.3390/app7010110. [09]

She, L., and J. Y. Chai. 2017. Interactive Learning of Grounded Verb Semantics Towards Human-Robot Communication. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). http://www.cse.msu.edu/~jchai/Papers/ACL2017.pdf. (accessed Oct. 5, 2017). [09]

She, L., S. Yang, Y. Cheng, et al. 2014. Back to the Blocks World: Learning New Actions through Situated Human-Robot Dialogue. In: Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue. http://www.sigdial.org/workshops/sigdial2014/proceedings/pdf/W14-4313.pdf. (accessed Oct. 5, 2017). [09]

Spangenberg, M., and D. Henrich. 2015. Grounding of Actions Based on Verbalized Physical Effects and Manipulation Primitives. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). http://ieeexplore.ieee.org/document/7353470/. (accessed Oct. 5, 2017). [09]

Tellex, S., T. Kollar, S. Dickerson, et al. 2011. Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. In: Proceedings of the National Conference on Artificial Intelligence (AAAI 2011). https://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3623. (accessed Oct. 5, 2017). [09]

Tellex, S., P. Thaker, J. Joseph, and N. Roy. 2014. Learning Perceptually Grounded Word Meanings from Unaligned Parallel Data. *Mach. Learn.* **94**:151–167. [09]

Tenorth, M., and M. Beetz. 2009. Knowrob—Knowledge Processing for Autonomous Personal Robots. In: Proceedings of the IEEE/RSJ International Conference (IROS 2009) Intelligent Robots and Systems. http://ieeexplore.ieee.org/document/5354602/. (accessed Oct. 5, 2017). [09]

Thomason, J., S. Zhang, R. Mooney, and P. Stone. 2015. Learning to Interpret Natural Language Commands through Human-Robot Dialog. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI). http://www.ijcai.org/Proceedings/15/Papers/273.pdf. (accessed Oct. 5, 2017). [09]

Thomaz, A. L., and C. Breazeal. 2006. Transparency and Socially Guided Machine Learning. In: Proceddings of the IEEE International Conference on Development and Learning (ICDL). https://pdfs.semanticscholar.org/e482/85256ed8ff285b69b22634cd59c65c2e1cd2.pdf. (accessed Oct. 5, 2017). [09]

Whitney, D., M. Eldon, J. Oberlin, and S. Tellex. 2016. Interpreting Multimodal Referring Expressions in Real Time. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). http://h2r.cs.brown.edu/wp-content/uploads/2016/08/whitney16.pdf. (accessed Oct. 5, 2017). [09]

Yang, S., Q. Gao, C. Liu, et al. 2016. Grounded Semantic Role Labeling. In: Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). http://aclweb.org/anthology/N16-1019. (accessed Oct. 5, 2017). [09]

Young, S., M. Gasic, B. Thomson, and J. Williams. 2013. Pomdp-Based Statistical Spoken Dialogue Systems: A Review. *Proc. IEEE* **101**:1160–1179. [09]

Yu, H., and J. M. Siskind. 2013. Grounded Language Learning from Video Described with Sentences. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. https://aclweb.org/anthology/P/P13/P13-1006.pdf. (accessed Oct. 5, 2017). [09]