Language to Action: Towards Interactive Task Learning with Physical Agents

Joyce Y. Chai[1], **Qiaozi Gao**[1], **Lanbo She**[2], **Shaohua Yang**[1], **Sari Saba-Sadiya**[1], **Guangyue Xu**[1]
[1] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824
[2] Microsoft Cloud & AI, Redmond, WA 98052
{jchai, gaoqiaoz, yangshao, sadiyasa, xuguang3}@cse.msu.edu, shelb26@gmail.com

Abstract

Language communication plays an important role in human learning and knowledge acquisition. With the emergence of a new generation of cognitive robots, empowering these robots to learn directly from human partners becomes increasingly important. This paper gives a brief introduction to interactive task learning where humans can teach physical agents new tasks through natural language communication and action demonstration. It discusses research challenges and opportunities in language and communication grounding that are critical in this process. It further highlights the importance of commonsense knowledge, particularly the very basic physical causality knowledge, in grounding language to perception and action.

1 Introduction

As AI starts to enter our everyday life, it's important for end users who are not technical experts to be able to teach artificial agents new knowledge and skills. Imagine in the future, you can purchase or rent a robot assistant. This robot comes with pre-programed knowledge and pre-trained skills. However the robot does not know anything about your household. There is no large amount of data available about your specific needs. You also cannot wait for the robot to explore your house by itself (and certainly don't want to risk the mess or destruction possibly brought to your kitchen). So what is ideal is for you to teach the robot the new environment and tasks as if you were teaching a human assistant.

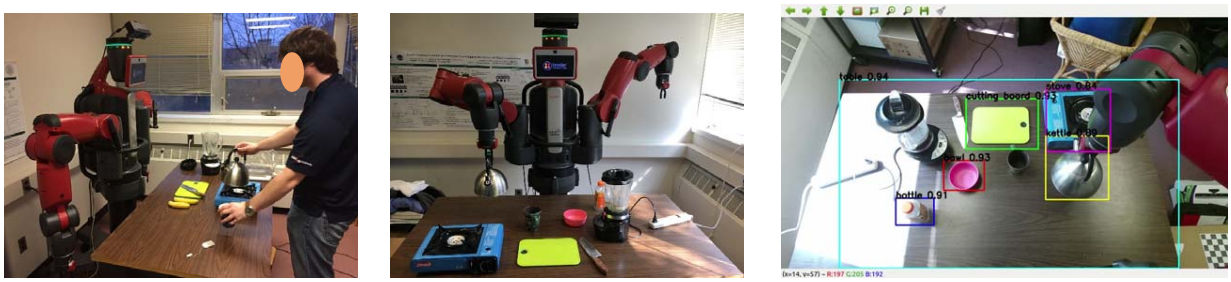
To address this issue, a new research area on Interactive Task Learning (ITL) is emerging [Gluck and Laird, 2018]. ITL is broadly defined as "any process by which an agent (A) improves its performance (P) on some task (T) through experience (E), when E consists of a series of sensing, effecting, and communication interactions between A, its world, and crucially other agents in the world." [Mitchell *et al.*, 2018]. In this paper, we discuss a specific form of ITL - communicative task learning, where humans can teach embodied agents (e.g., robots) in a shared physical world through language communication and action demonstration.

Communication provides a natural way for humans to acquire generic knowledge. Through a single exchange of information, teachers can selectively manifest the information to be acquired by learners. Such knowledge transfer can take the form of linguistic communication and manual demonstration, and is previously termed as *natural pedagogy* [Csibra and Gergely, 2009]. This is a kind of social learning which can accelerate learning by avoiding trials-and-errors and statistical generalization based on observations [Thomaz *et al.*, 2018]. Studies in developmental psychology have shown evidence of receptiveness and adaptation for natural pedagogy in young infants [Csibra and Gergely, 2006]. Through childhood to adulthood, humans have developed the ability to learn and teach through natural pedagogy, which appears universal across cultures and can be traced back to our ancestors [Tehrani and Riede, 2008]. As communication plays an important role in human learning, one question becomes important: how to enable natural pedagogy between humans and artificial agents and empower the agents to acquire new knowledge through communication with humans?

Recent years have seen an increasing amount of work on teaching robots new tasks through demonstration and instruction [Rybski *et al.*, 2007; Mohseni-Kabir *et al.*, 2018]. For example, learning from demonstration (LfD) [Thomaz and Cakmak, 2009; Argall *et al.*, 2009] learns a mapping from world states to robots' manipulations based on the human demonstration of desired robot behaviors. Recent work has also explored the use of natural language and dialogue to teach robots new actions [Mohan and Laird., 2014; Scheutz *et al.*, 2017]. We have also applied natural language communication and action demonstration to teach robots new tasks [She *et al.*, 2014; She and Chai, 2016; Liu *et al.*, 2016; She and Chai, 2017]. This paper gives a brief introduction to this research effort and discusses research challenges and opportunities.

2 Language Grounding in Learning through Communication

Language can be used in various ways to teach robots new tasks. For example, a human can teach a robot how to make tea by "telling" and "showing" and the robot learns through observation (shown in Figure 1a), or through its own actions by following human instruction and/or demonstration (Figure 1b). During learning, the robot observes how the world has been changed by the actions either performed by the hu-



(a) Robot learning from human language instruction and action demonstration.

(b) Robot learning through its own actions by following human instruction and demonstration

(c) Robot's perception of the physical world during learning.

Figure 1: An example setup of teaching a Baxter robot how to make tea.

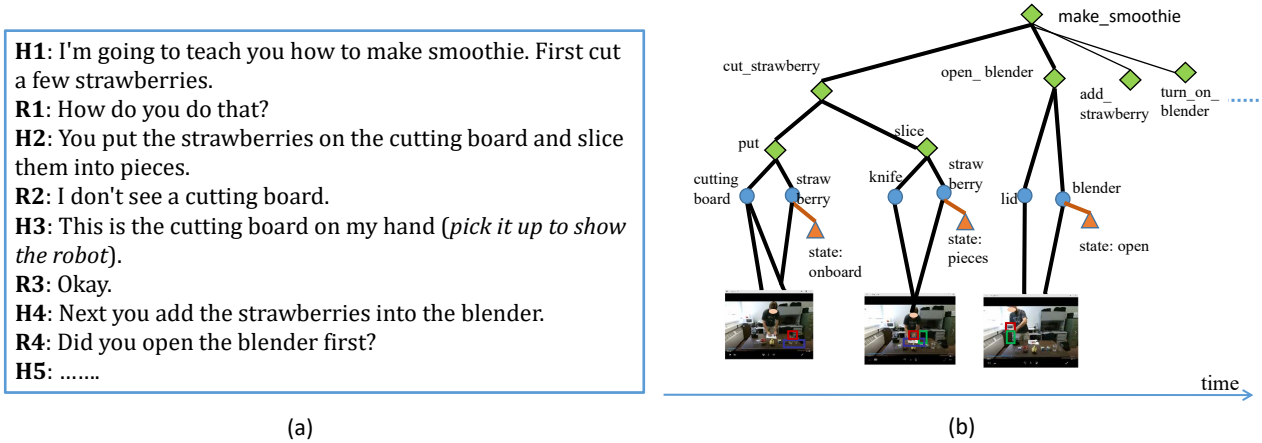


Figure 2: An example dialogue teaching the robot how to make smoothie (a) and the acquired grounded task structure (b).

man or by itself (Figure 1c, produced by YOLO [Redmon and Farhadi, 2017]). The robot can also communicate with the human back-and-forth to acquire tasks and task-related knowledge. Figure 2(a) shows an example dialogue where the human teaches the robot how to make smoothie. At the end of communication, the robot generates a grounded task structure that represents its understanding of this task as shown in Figure 2(b). Tasks are compositional in nature which can be captured by grammars [Liu *et al.*, 2016], Hierarchical Task Networks [Hogg *et al.*, 2009], or internal programming language [Wang *et al.*, 2017], etc. An overall task can be broken down into subtasks with possible constraints (e.g., temporal). A subtask can be decomposed into atomic actions which is *grounded* to the physical world where the agent can perceive or act. If the robot has the underlying manipulation ability (e.g., is able to perform the primitive action *cut*), the grounded task structure will allow the robot to plan and perform the learned task.

Enabling such communicative task learning faces many challenges. As shown in Figure 3, humans and robots are co-present in a shared environment. They both perceive from the environment and can potentially act to change the environment. However, they have significantly mismatched capabilities in perception, action, and reasoning. Their knowledge about the world is also vastly misaligned. All of these lead to disparities in their respective representations of the shared world and the task. The lack of common ground makes language communication between them difficult. Humans and agents will need to make extra collaborative effort to strive for a joint representation of the task structure. For example, they will need to keep track of each other's knowledge, beliefs, and intention (i.e., the Theory of Mind [Goldman, 2012]) as well as each other's abilities and limitations (i.e., user models) when interpreting or planning for communication. During this process, the robot acquires task-related knowledge and task structures to enrich its own knowledge base and also continuously updates its own representation of the shared world given the new knowledge. Thus, communicative task learning is more than just a process of acquiring grounded task structures. It is also intertwined with language learning (i.e., learning the grounded meanings of new words or language constitutes) and interactive knowledge acquisition (i.e., acquiring task-related knowledge and commonsense knowledge).

At the center of this process is the issue of *grounding*, a highly ambiguous term used in various context. In language communication with physical agents, two types of grounding are essential:

- **Semantic grounding** refers to the process where semantics of language is grounded to the agent's internal representations of perception from the world and actions to the world.
- **Communicative grounding** is the process for commu-

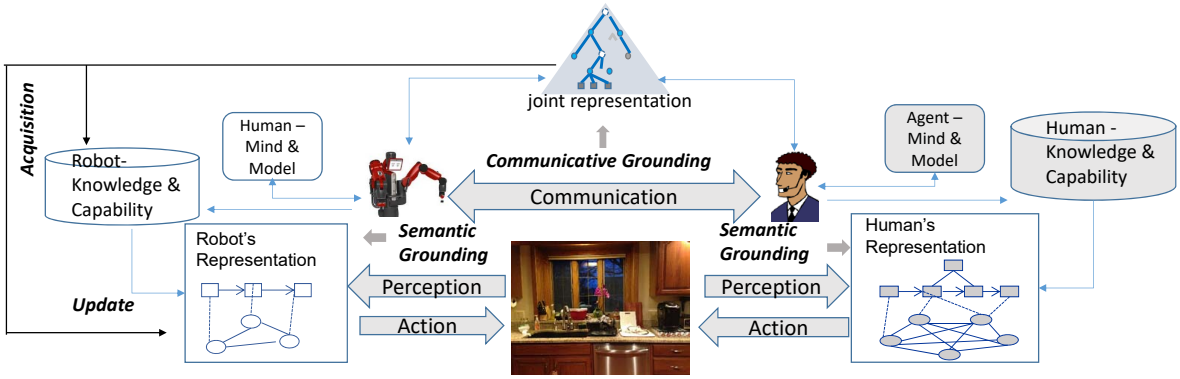


Figure 3: Semantic grounding and communicative grounding for learning a joint representation.

nication partners to reach a *common ground* - mutually agreed knowledge, beliefs, and assumptions. Communicative grounding is vital to keep partners on the same page to achieve joint communication goals.

2.1 Semantic Grounding

Semantic grounding relates to the classical concept of symbol grounding [Harnad, 1990] in Cognitive Science which proposes that meanings of symbols should be connected to the sensorimotor experience from the physical world. This notion has a particular significance in language communication with robots. In order for the robot to understand human language and act upon it, the meanings of language such as words, phrases, and utterances need to be grounded to the robot's sensors which perceive the environment and to the actuators which act to the environment.

Grounding to Perception

Grounding language to perception involves connecting meanings of words to machine perception [Roy, 2005; Matuszek *et al.*, 2012; Kennington and Schlangen, 2015; Thomason *et al.*, 2016] and grounding language expressions to visual objects [Liu *et al.*, 2012; Williams and Scheutz, 2018], to physical landmarks [Tellex *et al.*, 2011] and to actions or activities [Chen and Mooney, 2011; Artzi and Zettlemoyer, 2013]. In the context of communicative learning, grounding verbs and their arguments to the perceived world is crucial.

In Linguistics, verb semantics are often captured by semantic roles that specify arguments participating in an action such as `agent` (i.e., the one who performs the action), `patient` (the object the action is directed upon), `instrument` (the instrument used in the action) and so on [Baker *et al.*, 1998; Palmer *et al.*, 2005]. As shown in Figure 2(a), the verb *cut* takes the patient "strawberries" (`H1`); the verb *put* takes the patient "strawberries" and the destination "the cutting board" (`H2`). The robot will need to first identify different roles from linguistic utterances and then ground them to the perceived environment. Some of these roles such as `patient` are explicitly specified in language (i.e., *explicit roles*), but other roles, for example, the `instrument` (i.e., knife) associated with *cut* and *slice* (`H2`) is not explicitly stated (i.e., *implicit roles*). Our previous work [Yang *et al.*, 2016] has shown that, for a set of commonly used verbs, the role `instrument` is almost never explicitly specified in language, however it can be inferred from perception. For some verbs such as *take*, the `source` (where the things are taken from) is less likely specified and the `destination` (where the things taken to) is almost never explicitly stated. Nevertheless, these implicit roles are important components of an action. Therefore, the ability to ground not only explicit roles, but also implicit roles is vital in order for the robot to fully understand the composition of an action and possibly perform it.

Grounding to Action

Grounding verb arguments to the environment is not sufficient for the agent to perform corresponding actions. What controls actions of a robot typically consists of a discrete planner which captures a space of possible actions and their associated states, and a continuous planner that computes the trajectory for the movement. A robotic arm such as a `SCHUNK` industrial arm only has specifications for three primitive actions such as *open-gripper*, *close-gripper*, and *move-to*. Any higher level actions (e.g., specified by an action verb) will need to be translated to a sequence of primitive actions for the agent to perform [Kress-Gazit *et al.*, 2008]. Recent work has applied deep learning models to directly map language instructions and raw visual observations to actions [Misra and Langford, 2017] or action representations [Arumugam *et al.*, 2017]. These approaches require a large amount of training data which may not be available for the task at hand. In addition, to strive for a common ground in ITL, it is important for the agent to be able to explain its decision and receive relevant human feedback to update its model (particularly when an action fails). Thus, approaches that can connect verbs with the planning system and the ability to explain the robot's internal representations and decision making become important.

2.2 Communicative Grounding

In human-human communication, what enables us to understand each other depends on *common ground* and *shared intentionality* [Clark, 1996; Tomasello, 2008]. It is well established that communication is a cooperative process where both parties cooperate with each other to achieve common communication goals. These findings from human communication not only provide basis but also have new implications in human-robot communication.

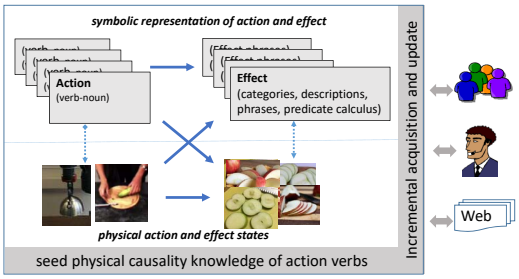


Figure 4: Physical causality knowledge of action verbs.

As shown in Figure 3, mismatched representations significantly jeopardize the common ground between humans and agents, making language communication difficult. When the common ground is missing, the intrinsic cooperative motivation will enable partners to collaborate and strive to establish a common ground. This is the process of communicative grounding. This cooperative principle brings challenges and opportunities in human-robot communication. For example, to mediate differences in the representation of the shared world, the speaker often produces language in an episodic and incremental manner to make sure the listener is following [Liu *et al.*, 2012]. The listener provides immediate feedback which may prompt the speaker to change language production in the middle of the planning. Therefore, from the robot perspective, algorithms for language interpretation will need to account for collaborative behaviors from the human, and algorithms for language generation will need to produce collaborative behaviors from the robot [Chai *et al.*, 2016]. Different mechanisms can be employed by humans and robots in communicative grounding such as using implicit or explicit confirmation [Thomaz *et al.*, 2018]. As shown in our previous work [Chai *et al.*, 2014], enabling transparency from the agent about its internal representations can reveal misunderstanding and significantly improve common ground. Moreover, the nature of embodiment in situated communication (i.e., non-verbal modalities such as gaze and deictic gestures) provides additional channels for communicative grounding [Chai *et al.*, 2018].

One of the key challenges to communicative grounding is presupposition. In human-human communication, much of background knowledge about the world is pre-assumed. The partners believe they share the same kind of background knowledge so do not need to explicitly state it in their communication. However artificial agents don't have the same kind of commonsense knowledge. To improve communicative grounding, one important solution is to equip the agent with an ability to acquire commonsense knowledge. Next we introduce some of our on-going efforts on acquiring commonsense knowledge, particularly, very basic causality knowledge associated with action verbs that are critical to task learning and execution.

3 Physical Causality of Action Verbs

Linguistics studies have shown that concrete action verbs can be divided into two categories *manner verbs* that "specify as part of their meaning a manner of carrying out an action" (e.g., laugh, run, swim), and *result verbs* that "specify the coming about of a result state" (e.g., empty, chop, open, enter) [Hovav, 2010]. While manner verbs play an important role in communicative task learning, our current work has focused on result verbs, particularly modeling causality - effects on the world given corresponding actions.

Different from other types of world knowledge (e.g., knowledge about places, people, events of the world), basic causality knowledge and the physics of the world is often presupposed and rarely explicitly stated in interpersonal or written communications. Applying NLP techniques to automatically populate knowledge base such as DBPedia, Freebase, and YAGO is not likely to result in basic causality knowledge of concrete actions. Existing resources such as VerbNet, FrameNet, and Propbank provide important information about the composition of a verb and its arguments, but they do not provide details on how the corresponding action may change the physical world. Recent work has investigated learning physics of the world from videos [Fire and Zhu, 2016] and simulations [Wu *et al.*, 2017]. However, except for a few works that look into physical properties of verbs [Forbes and Choi, 2017; Zellers and Choi, 2017], how verbs and their corresponding actions affect the state of the physical world is largely under-explored.

Representation. As shown in Figure 4, physical causality knowledge is represented by a mapping between *action* and *effect*. Symbolically, an *action* is specified as a *verb-noun* pair where a verb is a concrete result verb and a noun is a concrete noun which serves as a direct patient of the verb. An *effect* can be represented in various ways, for example, it can be as simple as categories to indicate the dimension of state change caused by the action to the direct object. It can also be captured by language descriptions (e.g., "the cucumber is chopped into pieces"), phrases (e.g., "cucumber + into pieces"), or predicate calculus that details the aspects of the changed world. As discussed in Section 3.1 and Section 3.2, different effect representations can be used in different tasks. The physical world captures physical actions (e.g., observed or manipulated by the agent) and perceived effect states (e.g., through vision and haptics). Symbolic actions and effects are grounded to physical actions and effects to facilitate language communication between humans and agents. As actions cause effects, such causality modeling will provide the agents basic knowledge about the physical world. Based on this knowledge, given an action, the agent can anticipate potential effects to the world; and given an effect state, the agent can reason about potential actions that may have led to that state. The acquired causality knowledge can be integrated with more formal models [Pearl, 2009] in the future to endow the agents more advanced abilities to do causal reasoning.

Acquisition. Causality knowledge can be acquired through three main channels. First, collective intelligence based on crowd-sourcing is used to create an initial seed knowledge base. After agents are deployed, it's likely they will encounter new verbs or new actions in a novel context for which there is no existing effect knowledge. Thus it is important to establish a process where the robot can incrementally and continuously acquire physical causality knowledge from collective intelli-

gence, web data (e.g., for images), and human partners the robot is working with. Next sections describe a few examples of our recent work that address acquisition of physical causality knowledge and how the acquired knowledge is applied to grounding language to perception and action.

3.1 Action-Effect Knowledge in Perception

Physical Causality in Grounding Verb Arguments. We first collected a dataset of commonly used verb-noun pairs (in a kitchen setting) and their effect descriptions using crowdsourcing [Gao *et al.*, 2016]. As shown from the data, result verbs often specify some movement along a scale [Hovav, 2010], which have similar behaviors of scalar adjective (e.g., big, small, long, short, etc.). Motivated by the typology for adjectives [Dixon and Aikhenvald, 2006], we have identified eighteen dimensions of physical change (i.e., physical causality categories) such as *size, shape, color, texture, visibility, solidity, temperature*, and *attachment*, etc. The changes along these dimensions can be potentially perceived from the environment, e.g., through visual or haptic sensors. Given a verb, the agent can anticipate the dimension of physical changes that can occur to the world. Given a noun, the agent can predict the affordance of the denoted object [Gibson, 1979; Chao *et al.*, 2015].

One important motivation of modeling physical causality is to provide top-down guidance for the agent to actively perceive the environment. When humans hear "pick up/take/put *something*", we anticipate the location of that *something* will change; when hearing "slice *something*", we anticipate that *something* will be changed into smaller pieces. Such anticipation is driven by the knowledge of the outcome associated with these result verbs during human language acquisition. If artificial agents have similar kinds of causality knowledge and can anticipate what may have happened or what is likely to happen to the physical environment, they can better perceive the environment and plan for lower level actions.

To validate this hypothesis, we incorporated causality knowledge into models to ground verb arguments to the environment [Gao *et al.*, 2016]. For each of the causality categories related to visual perception, we defined a set of visual detectors that aim to detect the kind of changes in the environment. For example, for the *attachment* category, the visual detector looks for "one object track breaks into multiple tracks". We then incorporate causality modeling in two different approaches. In the knowledge-based approach, the verb phrase (together with its causality category) from a human instruction will trigger corresponding visual detectors to ground arguments of verbs to objects that are most compatible with the detectors. In the learning-based approach, visual detectors are implemented as intermediate features and the association between the detectors and argument grounding is learned based on the training data. Our experimental results have shown that both approaches significantly improve argument grounding performance.

Action-Effect Reasoning. Given a verb-noun pair in language instruction, the anticipation of potential world change will enable the agent to better perceive the environment. Similarly, given a physical state of the world, the ability to infer what actions could cause that state of the world is equally important. For example, when teaching a robot, a human teacher may not explicitly describe every needed step. In `H4` (Figure 2), while the language instruction is "add the strawberries into the blender", but the human demonstrates by first opening the lid of the blender then putting the strawberries into the blender. When observing such a demonstration, the robot should ideally be able to infer that *add-strawberry* follows the step *open-blender* (i.e., `R4`) although this step is not explicitly instructed by the human. Partly inspired by this observation, our recent work has introduced a new task on naive physical action-effect prediction: given an effect state depicted by an image, predict actions (in the form of verb-noun pairs) that can potentially cause such effect [Gao *et al.*, 2018]. One problem of learning action-effect prediction models is the lack of training data. It is very expensive to have a large amount of image data (effect) which is annotated with corresponding causes (i.e., actions). To address this problem, we have applied a bootstrapping approach that harnesses web data through distant supervision for model training.

Although the performance on action-effect prediction is yet to be improved, our empirical results have shown that the web data can be used to complement a small number of seed examples (e.g., three images for a verb-noun pair in our experiments). This opens up possibilities for agents to learn physical action-effect relations for tasks at hand through communication with humans with a few examples. Given recent advances in distributional semantics, word embedding, and deep learning, our recent work has shown that there is a great potential the causality knowledge for known verb-noun pairs can be extended to new verb-noun pairs that may be encountered during task learning.

3.2 Action-Effect Knowledge in Planning

Incremental Acquisition of Grounded Verb Semantics. As discussed in Section 2.1, when following a natural language command, a robot needs to translate the action specified by a verb phrase to the lower-level primitive actions. Such translation often involves planning (e.g., classical STRIP or PDDL based planners or probabilistic planning based on Markov Decision Process). The core to these planners is action schemas or transition models which specify how primitive actions can cause the change of the world from one state to another. Therefore, another direction of our work is learning *grounded verb semantics* which explicitly models verb semantics (particularly for result verbs) as the desired goal states [She *et al.*, 2014; She and Chai, 2016]. Then given a verb and their arguments, planning algorithms can be applied to search for a sequence of primitive actions.

As social learning plays a pivotal role in child language acquisition [Tomasello, 2003], we developed an interactive learning framework where the agent can incrementally acquire grounded verb semantics by following the human teacher's instructions step by step. For example, a human can teach the agent the meaning of "fill the cup with water" by breaking it down into a sequence of steps, e.g., "pick up the cup, move to the sink, put down the cup, turn on the faucet, etc." As the robot performs each step, the teacher monitors its outcome. If the robot does not know how to perform a

particular step, additional instructions will be given, e.g., until all the way down to the primitive actions. At the end of teaching, after experiencing a sequence of changes in the environment, the robot will connect the state changes (i.e., a conjunction of predicates) with the new verb. In this case, the grounded semantic meaning for `fill(cup,water)` is $\lambda x \lambda y \lambda o_1 \lambda o_2\ isa(x, cup) \wedge isa(y, water) \wedge isa(o_1, sink) \wedge isa(o_2, table) \wedge has(x, y) \wedge in(x, o_1) \wedge \neg on(x, o_2)$. The hypothesis learned from this instance includes *cup in the sink* (i.e., $in(x, o_1)$) and *cup not on the table* ($\neg on(x, o_2)$). It is very specific and may not be relevant to a new situation. Our recent work thus extends a single hypothesis to form a hypothesis space to represent grounded verb semantics. The hypothesis space captures the specific-to-general hierarchy of all possible hypotheses applicable from this teaching instance [She and Chai, 2016]. During learning/teaching, a new hypothesis space can be acquired for each new verb (or new use of an existing verb) encountered. Existing hypothesis spaces are also combined, pruned, and updated given new experience. During execution, if there exists a hypothesis space for the verb in a language command, the robot will select the best hypothesis which is most relevant to the current situation (this selection can be learned through experience as well) and performs the action. If there is no knowledge of that verb or if the action is not correctly performed, a new teaching process as described above is initiated. This learning and execution form a loop which potentially allows the agent to continuously learn and acquire grounded verb semantics from human partners in the field. Using the data made available by [Misra *et al.*, 2015], our empirical results have shown the hypothesis space representation of grounded semantics significantly outperforms the single hypothesis representation.

Collaboration in Interactive Learning. As mentioned in Section 2.2, communication is a cooperative process. When humans and agents have mismatched capabilities and representations, both parties will make extra collaborative effort to achieve communication goals. For example, in our earlier work on teaching robots new verbs in a simplified blocks world [She *et al.*, 2014], we had human teachers perform two types of teaching: (1) teach one step at a time and make sure every step is correctly followed before moving to the next step; and (2) provide a complete instruction with multiple steps. Our empirical studies have shown that, as expected, the time taken for teaching is significantly higher in the *one-step-at-a-time* setting. However, when the agent applied the learned verb semantics in novel situations, the representations acquired from the *one-step-at-a-time* setting led to higher performance in action planning. These results have demonstrated that teaching style affects learning outcome, no exception for robots. Then the question is, how to make human teachers aware of the agent's abilities and limitations so that teachers can be more cooperative, for example, by providing the right kind of scaffolding in the teaching process. This makes explainable AI particularly important. The agent needs to communicate to the human teacher to explain its prediction, decision making, and action so that the teacher can provide the right kind of feedback (e.g., correction, additional instructions, etc.).

Agents should also be proactive in learning, especially when there are many levels of uncertainties. For example, the perceived world is full of uncertainties and is error-prone. How to make the agent to learn a reliable model of grounded verb semantics given the noise from the environment becomes an important question [She and Chai, 2017]. Motivated by previous work on interactive robot learning of new skills [Cakmak and Thomaz, 2012], we identified a set of questions for the agent to inquire about the state of the environment. We used an existing dataset [Misra *et al.*, 2015] to simulate different levels of noise of the environment and simulate interaction with a human teacher through question answering. Reinforcement Learning (RL) was applied to learn when to ask what question in order to maximize the long-term reward. Our results have shown that the policy learned from RL leads to not only more efficient interaction but also better models for the grounded verb semantics. Although this is encouraging, how to apply the learned policy to real world interaction across different tasks remains a challenging and important research question.

4 Conclusion

Language communication provides an efficient and natural means for artificial agents to acquire new tasks and task-related knowledge directly from humans. This paper gives a brief introduction to the key challenges in language and communication grounding to enable communicative task learning. What's presented here is only the tip of the iceberg. There are many research challenges ranging from commonsense reasoning, knowledge acquisition and sharing, to explainable AI, and human-agent collaboration. And many more problems are yet to be discovered. Given recent advances in language, vision, robotics, cognitive modeling, machine learning, and many other related disciplines, it has never been a better time to explore this exciting, highly multidisciplinary, and less studied territory.

Acknowledgments

The authors would like to thank Malcolm Doering, Sarah Fillwock, James Finch, and Kenneth Stewart for their contributions to data collection, annotation, and experiments. This work was supported by IIS-1208390 and IIS-1617682 from the National Science Foundation and the DARPA XAI program under a subcontract from UCLA (N66001-17-2-4029).

References

[Argall *et al.*, 2009] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[Artzi and Zettlemoyer, 2013] Y. Artzi and L. Zettlemoyer. Weakly supervised learning of semantic parsers for mapping instructions to actions. *TACL*, 1:49–62, 2013.

[Arumugam *et al.*, 2017] D. Arumugam, S. Karamcheti, N. Gopalan, L. Wong, and S. Tellex. Accurately and efficiently interpreting human-robot instructions of varying granularities. *Robotics: Science and Systems*, 2017.

[Baker *et al.*, 1998] C. Baker, C. Fillmore, and J. Lowe. The Berkeley FrameNet project. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 86–90, San Francisco, California, 1998.

[Cakmak and Thomaz, 2012] M. Cakmak and A. L. Thomaz. Designing robot learners that ask good questions. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 17–24, 2012.

[Chai *et al.*, 2014] J. Y. Chai, L. She, R. Fang, S. Ottarson, C. Littley, C. Liu, and K. Hanson. Collaborative effort towards common ground in situated human robot dialogue. In *The 9th ACM/IEEE International Conference on Human-Robot Interaction*, Bielefeld, Germany, 2014.

[Chai *et al.*, 2016] J. Y. Chai, R. Fang, C. Liu, and L. She. Collaborative language grounding towards situated human robot dialogue. *AI Magazine*, 37(4):32–45, 2016.

[Chai *et al.*, 2018] J. Y. Chai, M. Cakmak, and C. Sidner. Teaching robots new tasks through natural interaction. In K. A. Gluck and J. E. Laird, editors, *Interactive Task Learning: Agents, Robots, and Humans Acquiring New Tasks through Natural Interactions*, chapter 9. MIT Press, 2018. (In Press).

[Chao *et al.*, 2015] Y. Chao, Z. Wang, R. Mihalcea, and J. Deng. Mining semantic affordances of visual object categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4259–4267, 2015.

[Chen and Mooney, 2011] D. L. Chen and R. J. Mooney. Learning to interpret natural language navigation instructions from observations. In *AAAI*, 2011.

[Clark, 1996] H. H. Clark. *Using language*. Cambridge University Press, Cambridge, UK, 1996.

[Csibra and Gergely, 2006] G. Csibra and G. Gergely. Social learning and social cognition: the case for pedagogy. In Y. Munakata and M. H. Johnson, editors, *Processes of Changes in Brain and Cognitive Development*, pages 1249–274. Oxford University Press, 2006.

[Csibra and Gergely, 2009] G. Csibra and G. Gergely. Natural pedagogy. *Trends Cogn Sci.*, 13(4):148–153, 2009.

[Dixon and Aikhenvald, 2006] R. M. W. Dixon and A.Y. Aikhenvald. *Adjective Classes: A Cross-linguistic Typology*. Explorations in Language and Space. Oxford Press, 2006.

[Fire and Zhu, 2016] A. Fire and S. Zhu. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):23, 2016.

[Forbes and Choi, 2017] M. Forbes and Y. Choi. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 266–276, 2017.

[Gao *et al.*, 2016] Q. Gao, M. Doering, S. Yang, and J. Y. Chai. Physical causality of action verbs in grounded language understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1814–1824, 2016.

[Gao *et al.*, 2018] Q. Gao, S. Yang, J. Y. Chai, and L. Vanderwende. What action causes this? towards naive physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[Gibson, 1979] J. J. Gibson. *The Ecological Approach to Visual Perception*. 1979.

[Gluck and Laird, 2018] K. A. Gluck and J. E. Laird, editors. *Interactive Task Learning: Agents, Robots, and Humans Acquiring New Tasks through Natural Interactions*. MIT Press, 2018. (In Press).

[Goldman, 2012] A. Goldman. Theory of mind. In *The Oxford Handbook of Philosophy of Cognitive Science*. 2012.

[Harnad, 1990] S. Harnad. The symbol grounding problem. In *Physica D 42*, pages 335–346, 1990.

[Hogg *et al.*, 2009] C. Hogg, U. Kuter, and H. Muñoz-Avila. Learning hierarchical task networks for nondeterministic planning domains. In *IJCAI*, pages 1708–1714, 2009.

[Hovav, 2010] B. Hovav, M.and Levin. Reflections on Manner / Result Complementarity. *Lexical Semantics, Syntax, and Event Structure*, pages 21–38, 2010.

[Kennington and Schlangen, 2015] C. Kennington and D. Schlangen. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of Association for Computational Linguistics (ACL)*, 2015.

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

[Liu *et al.*, 2012] C. Liu, R. Fang, and J. Y. Chai. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 140–149, Seoul, South Korea, 2012.

[Liu *et al.*, 2016] C. Liu, S. Yang, S. Saba-Sadiya, N. Shukla, Y. He, S. Zhu, and J. Y. Chai. Jointly learning grounded task structures from language instruction and visual demonstration. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1492, 2016.

[Matuszek *et al.*, 2012] C. Matuszek, N. Fitzgerald, L. Zettlemoyer, L. Bo, and D. Fox. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1671–1678, 2012.

[Misra and Langford, 2017] D. K. Misra and Y. Langford, J.and Artzi. Mapping instructions and visual observations to actions with reinforcement learning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2017.

[Misra *et al.*, 2015] D. K. Misra, K. Tao, P. Liang, and A. Saxena. Environment-driven lexicon induction for high-level instructions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 992–1002, 2015.

[Mitchell *et al.*, 2018] T. Mitchell, S. Garrod, J. Laird, S. Levinson, and K. Koedinger. Framing the problem of interactive task learning. In K. A. Gluck and J. E. Laird, editors, *Interactive Task Learning: Agents, Robots, and Humans Acquiring New Tasks through Natural Interactions*, chapter 2. MIT Press, 2018. (In Press).

[Mohan and Laird., 2014] S. Mohan and J. E. Laird. Learning goal-oriented hierarchical tasks from situated interactive instruction. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014.

[Mohseni-Kabir *et al.*, 2018] A. Mohseni-Kabir, C. Li, V. Wu, D. Miller, B. Hylak, S. Chernova, D. Berenson, C. Sidner, and C. Rich. Simultaneous learning of hierarchy and primitives (slhap) for complex robot tasks. *Autonomous Robotics*, 2018.

[Palmer *et al.*, 2005] M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.

[Pearl, 2009] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

[Redmon and Farhadi, 2017] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[Roy, 2005] D. Roy. Grounding words in perception and action: computational insights. *TRENDS in Cognitive Sciences*, 9(8):389–396, 2005.

[Rybski *et al.*, 2007] P. E. Rybski, K. Yoon, J. Stolarz, and M. M. Veloso. Interactive robot task training through dialog and demonstration. In *The 2nd ACM/IEEE International Conference onHuman-Robot Interaction (HRI)*, pages 49–56, 2007.

[Scheutz *et al.*, 2017] M. Scheutz, E. Krause, B. Oosterveld, T. Frasca, and R. Platt. Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)*, pages 1378–1386, 2017.

[She and Chai, 2016] L. She and J. Y. Chai. Incremental acquisition of verb hypothesis space towards physical world interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 108–117, Berlin, Germany, 2016.

[She and Chai, 2017] L. She and J. Y. Chai. Interactive learning of grounded verb semantics towards human-robot communication. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1634–1644, 2017.

[She *et al.*, 2014] L. She, S. Yang, Y. Cheng, Y. Jia, J. Y. Chai, and N Xi. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Proceedings of the SIGDIAL 2014 Conference*, 2014.

[Tehrani and Riede, 2008] J. J. Tehrani and F. Riede. Towards an archeology of pedagogy: Learning, teaching and the generation of material culture traditions. *World Archaeol*, 40:316–331, 2008.

[Tellex *et al.*, 2011] T. Tellex, S.and 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 *AAAI*, 2011.

[Thomason *et al.*, 2016] J. Thomason, J. Sinapov, M. Svetlik, P. Stone, and R. J. Mooney. Learning multi-modal grounded linguistic semantics by playing "i spy". In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16)*, pages 3477–3483, New York City, 2016.

[Thomaz and Cakmak, 2009] A. L. Thomaz and M. Cakmak. Learning about objects with human teachers. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, HRI '09, pages 15–22, New York, NY, USA, 2009. ACM.

[Thomaz *et al.*, 2018] A. L. Thomaz, E. Lieven, M. Cakmak, J. Y. Chai, S. Garrod, W. Gray, S. Levinson, A. Paiva, and N. Russwinkel. Interaction for task instruction and learning. In K. A. Gluck and J. E. Laird, editors, *Interactive Task Learning: Agents, Robots, and Humans Acquiring New Tasks through Natural Interactions*. MIT Press, 2018. (In Press).

[Tomasello, 2003] M. Tomasello. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, 2003.

[Tomasello, 2008] M. Tomasello. *The Origins of Human Communication*. MIT Press, 2008.

[Wang *et al.*, 2017] S. I. Wang, S. Ginn, P. Liang, and C. D. Manning. Naturalizing a programming language via interactive learning. In *Association for Computational Linguistics (ACL)*, 2017.

[Williams and Scheutz, 2018] T. Williams and M. Scheutz. Reference in robotics: A givenness hierarchy theoretic approach. In Jeanette Gundel and Barbara Abbott, editors, *The Oxford Handbook of Reference*. 2018.

[Wu *et al.*, 2017] J. Wu, E. Lu, P. Kohli, W. Freeman, and J. Tenenbaum. Learning to see physics via visual deanimation. In *Proceedings of the 31th Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

[Yang *et al.*, 2016] S. Yang, Q. Gao, C. Liu, C. Xiong, S. Zhu, and J. Y. Chai. Grounded semantic role labeling. In *Proceedings of NAACL*, pages 149–159, 2016.

[Zellers and Choi, 2017] R. Zellers and Y. Choi. Zero-shot activity recognition with verb attribute induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.