Abstract: This note outlines a case for robot learning researchers to introduce natural language into their research agenda. The importance of language in robot learning goes beyond the potential of natural language interfaces. Language touches on core aspects of the robot learning problem, including modularity, sample complexity, simulation, and the gap between short- and long-term behaviors.

Keywords: natural language, robot learning, representation learning

1 Introduction and Motivation

The interest in robot learning methods that integrate natural language is slowly increasing. However, the place of natural language within robot learning research is still often questioned. The objective of this note is to answer this question by advocating for a tight research integration between robot learning and situated natural language processing (NLP), describe some of the current work in the area, and illustrate how natural language exposes robot learning to new research avenues.

The most established use case for natural language in robotics is as an interface for user-robot interaction. Natural language provides an interface that is both expressive and accessible to non-expert users. While the need for such an interface is not new, it is of increased urgency given the objective of robot learning to expand robotic function beyond carefully controlled environments to complex real-world situations. Key to achieving this goal is allowing non-expert users to control capable robotic agents. However, as the span of robot behaviors expands, programmatic and visual interfaces become increasingly difficult to operate without a high level of expertise. This makes the building of natural language interfaces critical and integral to the robot learning problem.

Natural language has been studied in robotics research mostly within modular approaches [e.g., 1, 2, 3, 4, 5, 6, 7]. The modular approach reflects traditional division of labor within and between fields, and allows for the adoption of abstractions that potentially simplify the study of complex systems. However, current understanding of representation learning effectiveness points away from a strict modular design. Natural language places unique requirements on every part of a robot learning system, including the overall architecture. These requirements affect how modularity should be viewed within a robot learning system, what reasoning is required of the system and how it is performed, and how information about the world is represented and extracted. This calls for tighter algorithmic integration between NLP and robot learning, similar to ongoing trends in perception [8].

2 Decoupling NLP and Robotics Research with Modular Approaches

For the purpose of this note, the modular approach (also known as the pipeline approach) to the robotics problem is characterized by decomposing complex robotic systems into separate modules, each solving a different sub-problem. For example, a potential decomposition will include modules for perception, mapping, planning, and control. The modules are then built, or trained separately and integrated together. This process requires designing interfaces between the different modules, often using symbolic representations. The modular approach eases the design of complex systems and enables the study of specific problems in isolation. Other advantages include module re-usability across systems, and interpretability through the representations used in the inter-module interface.

More informally, this note aims to help the perplexed roboticist that wonders (a) why are language people coming to my conference; (b) why should I care, pay attention, and actually start working on these problems; and (c) how should I get about doing that.

Submitted to the 5th Conference on Robot Learning (CoRL 2021). Do not distribute.
Natural language interaction with robotic agents has been studied extensively within the modular approach, both directly with physical robotic agents [e.g., 1, 2, 3, 4, 5, 6, 7] or through proxy tasks with varying degrees of real-world fidelity [e.g., 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. As far as the language problem, adopting a modular approach often entails developing a discrete symbolic representation that forms an interface between the language processing module and the other modules in the system. Beyond the fit for modular system design, this is aligned well with using symbolic representations to capture language meaning, an approach that originates from the study of formal semantics in linguistics [22, 23, 24, 25] and, until recently, has been popular in NLP [e.g., 26, 27, 28, 29, 30, 31]. Such representations organize concepts into ontologies, for example of object classes and attributes, action types and effects, and spatial and containment relations.

This allowed researchers in robotics to abstract away the language problem, and focus on integrating tools from the NLP community into robotic systems [e.g., 2, 32, 6]. This is not dissimilar to how perception is often addressed within modular systems [e.g., 33, 34, 36, 37, 38]. In practice, this expresses an underlying expectation that NLP research will produce a way to recover a formal meaning representation that will give robotic systems the actionable information needed.

3 The Opportunity in Joint Language and Robot Learning Research

The modular approach in robotics requires significant engineering and integration efforts, which robot learning largely trades off with learning challenges. A potentially more limiting property of the modular approach is reliance on manually-designed representations. Despite significant efforts [39, 40], such representations remain difficult to design at the scale needed for real-life situations in a way that is usable and generalizable. For example, simply consider designing a symbolic formal language to represent all object types observed in a regular home environment, their states, and their affordances. This is not necessarily an issue for heavily-engineered restricted environments, such as in manufacturing, but poses challenges for deployment in more dynamic environments, such as homes. The answer of robot learning to this challenge is representation learning.

This transition is mirrored in the NLP community, where hand-crafted linguistic representations [41, 42] are being replaced with implicit learned representations, either acquired from raw text [43, 44] or from labeled data [45, 46]. In addition to the general representation design challenge, the symbolic representations commonly used in NLP are often not suitable for robotics systems. This is because they are often designed for and tightly coupled with commonly studied domains in NLP, such as newswire [41], and tasks, such as semantic labeling [47, 48] and textual entailment [49]. In practice, this left the need for a formal robotics-oriented robust language representation largely unanswered.

The shift to robot learning and the convergence of techniques across the communities (i.e., via the use of neural networks) creates a unique opportunity to develop language technologies within robot learning, including models and learning methods primarily designed for robotic applications.

As with the modular approach, the problem of natural language interaction is often studied with representation learning using proxy tasks focused on constrained environments [50, 51], realistic perception [52, 53, 54], household tasks [55, 56], and high-fidelity robot simulation [57, 58, 59]. Deployment on physical robots remains less frequent. Blukis et al. [60] presented an approach to map raw observations and natural language instructions to continuous control of a quadcopter drone. The key to deploying on a physical drone was a decomposition of the model, which allowed for concurrent reinforcement learning in simulation and physical environments without autonomous real-world flight during learning. Blukis et al. [61] extended this approach to few-shot learning, where the robot reasons about previously unseen objects during testing. Anderson et al. [62] demonstrated transfer from a panorama-based virtual environment to instructing a physical ground robot, largely highlighting the challenging gap between the two. Banerjee et al. [63] studied maze robot navigation, focusing on supervisor-follower dialogue. While all these tasks are far from solved, the tight integration of robot learning and NLP brought about results beyond the reach of previous methods.

Beyond the user interface problem, the integration of robot learning and NLP enables new research avenues. Human language provides a conceptual framework (i.e., a schema) to relay information about the world, including many of the concepts (i.e., objects, actions, states) that robotic agents need to reason about. While informal, for example in contrast to logical ontologies, its expressivity and coverage are unparalleled. For robot learning, it can provide a form of inductive bias that is easy to obtain from non-expert users. As such, it forms an organizational schema to bias representation
learning, with the potential of enabling more efficient learning, maybe even during robot deploy-
ment when human users linguistically discuss robot operation. Variants of such language-informed
learning have been studied in reinforcement learning domains [64, 65, 66, 67], including with robot
simulations [68]. Largely though, the area remains understudied, especially with physical robots.

4 Natural Language Expands Robot Learning

The integration of robot learning and NLP expands the envelope of research in the field, propelling
the community beyond learning specific skills or gym-like tasks [69, 70, 71, 72]. Not just due to the
added complexity of yet another type of input, but because of the flexibility of natural language in
referring to robot behavior, and the many layers of abstraction it encompasses.

Modularity vs. Integration in System Design  Robot learning is largely motivated by the weak-
nesses of modular approaches in scaling to complex, real-world tasks (Section 3). However, mod-
ular approaches provide several advantages, including the decomposition of complex problems to
simpler ones, a higher potential for interpretability, and re-usability of modules (Section 2). The
flexibility of representation learning is key to reasoning about language meaning, and as such instru-
mental in enabling contemporary robot learning systems that use natural language. However, the
complexities of such systems, their tasks, and their accompanying learning challenges require us to
reconsider the cost of completely giving up on modularity. This raises the need to re-visit modularity
in the context of robot learning. Can we design modular abstractions that retain the learning benefits
without the drawbacks of conventional modular systems?

As an example of how this question is currently studied, consider the model decomposition ap-
proach of Blukis et al. [57, 58, 60, 61], which separates the control network from the language and
perception inputs via an internal plan-like visual representation. Language and perception are tightly
integrated with plan generation, but the approach stops short of considering these inputs in making
specific control decisions. The plan is simply a distribution over observed locations in the world,
specifying for each location the probability of visiting it during execution. This representation relies
on modeling basic concepts like trajectories, but avoids the intractable challenges of symbolic on-
tology design. It provides sample-complexity and interpretation benefits, and is sufficient to model
many of the tasks currently studied. However, such strict decomposition depends on data that allows
training the two model parts, and the intermediate representation capturing all language meaning
necessary for the task. As tasks become increasingly complex, both requirements will be harder
to satisfy. The intermediate representation may be harder to design and train in a way that scales
to more complex tasks. For example, consider the instructions pull the lever slower and slower or
push the door strongly. In both, the control output is tightly dependent on the language instruction,
and any intermediate representation will have to capture the required velocity and thrust dynam-
ics. This entails either a more expressive intermediate representation, or directly reasoning about
the language when making control decisions. Both avenues pose open research problems that are
language-focused instantiations of some of the most pressing questions in robot learning.

Simulation, Synthetic Language, and Data Fidelity  Obtaining data for training is a core chal-
lenge in robot learning. Currently, maybe the most common way to address it is via simulation [73].
Because building simulations of sufficient fidelity for complex tasks remains an open problem, the
eventual utility of this approach remains a topic of frequent discussion in the community. Regard-
less, much of the current progress in the field relies on simulations for training.

The inclusion of natural language in robot learning further brings the simulation problem into focus.
Simulation allows observing a very high number of states during training, with very little cost.
However, when including natural language in the world state, for example when an instruction is
given to a robot, a physics simulation is no longer sufficient to model the dynamics of the complete
world state. This exposes new shortcomings, and open problems in the use of simulation.

A potential solution is to expand the use of simulation to language by generating synthetic language
during training, for example with templates [74] or a generative grammar [75]. In contrast to simu-
ation in robotics, transferring models that are trained on synthetic language to natural language has
provided mixed results. A slightly more effective approach is to use a corpus of language in the

2A common pitfall that may obscure this is using synthetic data for testing, and generalizing the results to
success on natural language. In general, such evaluation provides little indication for performance on human
language. Regardless of the training regime, it is critical that the test data comprises of natural language.
target domain to train a generative model of in-domain language. This approach has shown performance improvements when the generated data is used for data augmentation [76], but applying it to completely new environments and tasks, as would be needed for simulation, remains an open problem. The simplest, and most common approach is to limit the language data seen during training to a static training corpus only, thereby observing only human-created natural language. Effectively, this constrains the simulated state space.

A less understood data problem is the effect of using non-embodied tasks for collecting natural language data for robotic tasks. Even relatively restricted NLP domains often require $>10k$ examples for training. For example, the robotic drone task of Blukis et al. [60] uses 23k instructions for training, a modestly sized corpus by NLP standards. However, collecting this amount of data using a physical robot in a lab is costly and slow. The common solution is crowdsourcing, often using virtual environments [51]. While more scalable and cost effective, the proxy simulation tasks cannot replicate the shared embodiment of users and robots interacting in the same space. This may be a reasonable way to collect training data, but it is unclear if use of such data for evaluation provides a valid way to estimate system performance when human users instruct the robot sharing the same space. Although this practice is common, this validity problem is largely an open question.

Finally, data issues become even more complex when the task entails back-and-forth user-robot interaction, such as in dialogue settings [63]. Such systems are often trained, and even tested on wizard-of-oz (WOZ) data, collected via interaction between a user and an expert operator. However, when system performance influences the user’s language and behavior, static WOZ data is a poor proxy. This challenge is not specific to robotics, and appears in many interactive NLP problems. These issues impact training and testing, but it is the latter that requires more caution. Generally, this calls for in-person evaluation with human users, which is difficult to conduct at the necessary scale. At a minimum though, evaluation protocols can explicitly discuss such limitations. These are general evaluation challenges in robot learning, which the inclusion of language brings into focus.

**Short- vs. Long-term Reasoning** Natural language can seamlessly specify both high-level goals (e.g., *reheat leftover soup*) and low-level constraints (e.g., *carry the soup slowly and carefully*). While robot learning methods show significant success on learning sensorimotor skills [77, 78, 8, 79, 80], reasoning about natural language brings to front the problem of bridging the gap between high-level goals and low-level actions, potentially through long-term planning and hierarchical reasoning. Despite progress, these challenges are not adequately captured in current learning benchmarks. For example, the ALFRED [56] benchmark includes both high- and low-level instructions in a virtual 3D environment, and has enabled studying long-term task execution, hierarchical reasoning, and representation learning. However, it abstracts over manipulation, control, and real-world deployment challenges. On the other hand, navigation benchmarks [e.g., 60] provide a fuller view of the robotic system, but are focused on short-term behavior with limited planning.

There remains a need for benchmarks that target the full scope of natural language as an interface for and a reasoning framework within physical robot systems. Here too, the integration of natural language as part of the research problem crystallizes the robot learning problem more completely, encompassing both short- and long-term skill learning.

## 5 Conclusion

This brief note only touches on the potential of the study of natural language within the robot language community. This is far from an exhaustive discussion, or a complete enumeration of relevant work. We do not discuss how natural language enables avenues for collaboration with users and exposes shifting user behavior (e.g., via adaptation). While these topics are studied in the HRI community, they are also core to robot learning systems, especially when learning from interaction with users. We also do not review the paradigm similarities between NLP systems that reason about previously unseen language (e.g., for instruction following) and meta learning systems, a learning approach receiving significant attention in robot learning. Finally, although we briefly mention few-shot learning [61], we do not elaborate on how the inclusion of language impacts this problem. All these require a longer discussion, beyond the scope of this note.

Currently, language-using research in robotics is often solely affiliated with HRI, both by roboticists and funding agencies. We hope this note illustrates the importance and potential of expanding the fundamental algorithmic study of natural language within the robot learning community.
References


