Recursive Spoken Instruction-Based One-Shot Object and Action Learning

Matthias Scheutz[1], **Evan Krause**[1], **Bradley Oosterveld**[1], **Tyler Frasca**[1], **and Robert Platt**[2]
[1] Human-Robot Interaction Laboratory, Tufts University, Medford, MA 02111, USA
[2] Northeastern University, Boston 02115, USA
matthias.scheutz@tufts.edu, evan.krause@tufts,bradley.oosterveld@tufts.edu,tyler.frasca@tufts.edu
rplatt@ccs.neu.edu

Abstract

Learning new knowledge from single instructions and being able to apply it immediately is highly desirable for artificial agents. We provide the first demonstration of spoken instruction-based one-shot object and action learning in a cognitive robotic architecture and briefly discuss the architectural modifications required to enable such fast learning, demonstrating the new capabilities on a fully autonomous robot.

1 Introduction

Quickly acquiring new knowledge during task performance, possibly in "one-shot" from a single instruction, being able to use it right away and also share it other agents would be an important feat for many agent applications. While most current artificial cognitive systems have the ability to acquire new knowledge, some even from natural language instructions (e.g., [Kirk and Laird, 2014; Mohan *et al.*, 2012]), they typically make various assumptions about perceptual and actuation capabilities (i.e., what primitive percepts and actions are available) as well as internal representations (e.g., what concepts and relations can be represented). As a result, they typically cannot accommodate truly novel objects or actions, or even novel object parts or known actions performed on those parts. Moreover, in learning new knowledge from instructions, an architecture also has to cope with unknown words in the instruction, in addition to the unknown concepts denoted by those words. This, in turn, requires the natural language subsystem (NLS) of the architecture to be able to cope with all aspects of unknown words: from their acoustic features, to their syntactic properties, to their semantic meaning, and possibly their pragmatic implictures. Hence, to be able to *truly learn from natural language instructions*, agent architectures need to allow for systematic representations of unknown entities such as words, percepts, actions, etc. that can be processed in almost every component of the architecture and subsequently refined based on the semantics of the natural language instructions (and possibly perceptual and other constraining contextual factors). This requires modifications to the component algorithms and deep interactions between the NLS and other components to allow for subsequent refinement and further specification of those initially incomplete representations.

We propose modifications to a cognitive robotic architecture that allow it to learn new objects and actions in *one shot*, i.e., from a single instruction, in such a way that (1) the acquired knowledge about objects and actions is integrated within the existing knowledge, (2) the knowledge can be used immediately by the learning agent for task performance, and (3) the knowledge can be shared immediately with other agents using the same architecture.

2 Instruction-Based One-Shot Learning

Instruction-based one-shot object and action learning can be defined as learning *conceptual definitions* for objects and actions as well as their aspects (e.g., object parts and action parameters) from natural language expressions that contain these definitions. For example, an object definition such as "A medical kit is a white box with red cross on it and a handle on top" defines a medical kit in terms of other shape, color, and object concepts referred to by color adjectives (e.g., "white" and "red"), shape nouns (e.g., "box" and "cross") and other object types (e.g., "handle") as well as relational expressions such as "on" and "on top" which relate the various object parts. A vision system that knows how to recognize and determine the various ingredients used in the definition can then recognize the new object by way of recognizing its constituent parts and their relationships (cp. to [Krause *et al.*, 2014]). Similarly, a definition such as "To follow means to stay within one meter of me", again assuming that all concepts after "means" are known, should allow an action execution component to construct an action script that captures the procedural meaning of the expression (cp. to [Cantrell *et al.*, 2011]).

Definition[Natural language object and action definition]. Let W_O be a set of natural language expressions denoting objects, object properties, or object parts in a set O, W_r be a set of relation expressions denoting relations in R among object parts O, and W_v be a set of natural language expressions denoting actions and action sequences as well as action modifications in a set of actions V. Then the natural language

expression $U(w, W_o, W_r, W_v)$ is a *definition* of a concept denoted by w if U contains w in a way that marks w as the *definiendum* (i.e., w is used to denote what is being defined such as saying in "A table is..." or "I will teach you how to pick up...") and the rest of the U, the *definiens*, involves any number of expressions from W_o, W_r, and W_v in a compositional fashion (e.g., "white box with red cross on it" or "stay within one meter") such that the composite meaning denoting a new object or action can be determined from the meaning of its parts.

One-shot object and action learning then amounts to (1) learning the linguistic aspects of the definiendum w, (2) determining the semantics of the definiens U, possibly recursing on unknown words, and (3) associating the constituent parts of the definiens U with different data representations in the agent architecture. Assuming that the architecture has all functional representations and processes for all W_o, W_r, W_v (e.g., object, object part and relation detectors in the vision system and action primitives and parameters in the action execution system), then invoking w in subsequent utterances will lead to retrieval and application of these data structures. In other words, by being able to understand natural language definitions of concepts cast in terms of either known concepts or other unknown concepts that are themselves defined in natural language, an agent can quickly acquire new knowledge by combining existing knowledge in a way prescribed by those definitions. To make this also practically possible, several changes and additions must be made to various architectural components to enable the agent to handle new words, generate novel data structures based on expressions that refer to other knowledge, and associate those data structures in ways that future invocations of the word will trigger the right kind of retrieval and application processes of the knowledge. For example, the speech recognizer must learn the acoustic signature of the new word on its first occurrence and be able to recognize it subsequently, the syntactic and semantic parsers have to be able to assign a grammatical type, syntactic structure and descriptive semantics to the word, and depending on whether it denotes an action or object, the new term has to be associated with knowledge in the vision and action components. Moreover, the agent has to detect when new knowledge is presented and understand from the utterance what type of knowledge it is.

3 Required Architectural Modifications

In this section, we very briefly describe some of the architectural changes required to allow for genuine one-shot learning and how we addressed them (for details, please refer to a longer version of this paper that appeared in 2017 [Scheutz *et al.*, 2017]).

3.1 Speech Recognition

The speech recognizer must be able to reliable detect unknown out-of-vocabulary words and generate new token entries for those words in its dictionary while also generating prototypes of the acoustic signals for subsequent recognition of the word (to avoid the addition of new tokens for the different instances of the same word).

We address this challenge by adding a special recognizer for words not recognized by standard speech recognizers based on the acoustic DP-ngram algorithm [Oosterveld *et al.*, 2017; Aimetti, 2009], which will generate a new token for each novel word (e.g., "UT1") that is added to the ASR dictionary and passed on down the language processing chain for each subsequently recognized occurrence of the word.

3.2 Parsing

The parser must be able to handle new words without parts-of-speech (POS) tags and attempt to infer their POS tags from the lexcial parsing context, and then generate descriptive semantic representation for the unknown words for subsequent refinement (i.e., through explicit definitions).

We address this challenge by determining whether the new word token is used as an action or as on object/object part based on its lexical context, then we generate Combinatory Categorial Grammar (CCG) types for the unknown word based on the assumption that the utterance was grammatically correct and generate generic lambda terms for noun phrases or generic actions based on the arity of the action expressions using extensions of the parser from [Dzifcak *et al.*, 2009].

3.3 Vision Processing

The vision system must be able to combine existing image processing algorithms based on natural language description to detect objects and parts determine potential grasp points on novel objects.

We address this challenge by using natural language instructions to determine how existing detectors and image processors in the vision system have to be combined to allow for the detection of novel objects (e.g., [Krause *et al.*, 2014]) and use a convolutional neural network to make grasp predictions based on projections of the portion of the point cloud contained between the fingers (e.g., [Gualtieri *et al.*, 2016]).

3.4 Action Processing

The action system needs to recursively generate action scripts from natural language instructions and determine appropriate action arguments, while interacting with the vision system to generate actions performed on the right object parts.

We address this challenge by incrementally assembling action scrips based on whether an utterance contains actions vs. control instructions (e.g., "while" or "if") and inferring action script signatures from the syntactic CCG representation (i.e., how many arguments an action requires, see also [Cantrell *et al.*, 2011]) and inferring their types based on the language context as well as the actual types used in the scripts.

4 Demonstration: Recursive One-Shot Learning

In this section, we briefly walk through an example (see Fig. 1) of recursively learning how to perform a complex action sequence, assume that the robot just learned what a plate is (a video demonstration is available at https://hrilab.tufts.edu/movies/recursiveoneshotlearning.mp4).

Human:	Pass the plate.
Robot:	Sorry, I do not know how to do that.
Human:	OK, I will teach you how to pass the plate.
Robot:	OK.
Human:	Pick up the plate.
Robot:	Sorry, I do not know how to do that.
Human:	OK, I will teach you how to pick up the plate.
Robot:	OK.
Human:	First, find the plate.
Robot:	OK.
Human:	Then grab the plate.
Robot:	OK.
Human:	Move the plate up.
Robot:	OK.
Human:	That is how you pick up the plate.
Robot:	OK.
Human:	Move the plate forward.
Robot:	OK.
Human:	Release the plate.
Robot:	OK.
Human:	That is how you pass a plate.
Robot:	OK.

Figure 1: The demo learning interaction.

The ASR recognizes the utterance except for "pass" and creates a new unique identifier "UT1" for it. The recognized text "UT1 the plate" is sent to the Parser which generates semantics of the form "UT1(self,ι(x).plate(x))" (where $\iota(x)$ determines the definite reference that is subsequently resolved to the plate in front of the robot) and passes it on to the Dialogue system. The utterance is interpreted as a literal command, acknowledged ("OK") and passed on to the KR and Inference component which generates a new goal "UT1(self,ι(x).plate(x))" that is sent to Action Manager. The Action Manager however does not have an action script for "UT1", hence action selection fails and control is returned to the Dialogue manger, where a response is generated to address the action failure by saying "I do not know how to do that". The subsequent "I will teach you how to pick up the plate" in an indication for the action component to start assembling a new action script and start monitoring the subsequent utterance for action instructions or control instructions. Note that from the utterance it is possible to infer the argument structure of the new action: "pass(actor,object)" (and also note that this is only a simplified version of passing an object, a more complex version includes the recipient as well). Since the first instruction contains another unknown action, learning proceeds recursively and the robot continues to learn the "pick up" action next. Currently the system needs an explicit instruction to look for the object ("find the plate"), otherwise it would not look for one when the script is being executed (although this can be inferred when attempts to pick up objects fail because the system does not know which object to pick up). The end of each instruction sequence is indicated by "This is how you...", at which point a new script is assembled, indexed, and available for immediate execution.

5 Discussion

The above walk-throughs show how new, initially meaningless token representations are generated as part of the learning process and become increasingly associated with different meaning representations. Note that instructions do not have to pertain to a particular set of sensors or actuators and that they do not depend on a particular robotic platform either. Nor do instructions have to be contained in one sentence, but can be spread over multiple sentences and dialogue interactions. Moreover, learning can be implicitly triggered using a novel word that the robot does not understand, making the robot prompt the user for a definition, in which case new acoustic, syntactic, and formal semantics representations will be generated that get associated with the content representations of the instructed knowledge after the relevant parts of the utterances were semantically analyzed (e.g., the visual representations of object part or the action script representation of the action sequence). Note that since all knowledge representations (i.e., the associations with the new token learned as part of the learning process) are *purely additive*, i.e., do not modify existing knowledge, it is possible to transmit the knowledge directly to other agents who do not yet have that knowledge for integration into their architectural components (e.g., see [Scheutz, 2014] for a discussion on how to do this).

It is also important to point out that the proposed architectural augmentations and the resultant one-shot learning scheme are not limited to the particular simple example demonstrated in the above walk-through. Rather, being implementations of the general one-shot learning definition in Section 2, they are very general themselves, only limited by the robot's knowledge of natural language as well as its perceptual and actuation capabilities. This raises then the question of how a system like the proposed system which implements a formally correct algorithm (i.e., using meaning expressions in logical definitions cast in natural language to associate the definiendum with the definiens) should then be evaluated. Clearly, empirical runs are important in the robotic case, since implementation details as well as real-time and real-world constraints matter. For this purpose, we provided an uncut video showing the algorithms at work in real-time on a fully autonomous robot. And the discussion of the NLS showed that the architecture can truly handle new words acoustically, syntactically, and semantically as well on the natural language side.

It is an interesting question to determine the extent to which knowledge acquired through one-shot learning is robust, and is another interesting aspect deserving of further investigation. The demonstrations discussed above has a nearly 100% success rate when repeatedly instructed after the initial learning instructions (i.e., if the robot is repeatedly taught how to pass an object). Similarly robust results are obtained using other definitions, but note that ultimately the robustness of application of a newly learned knowledge item depends on the robustness of its constituent parts (e.g., the detectors in the vision system that detect objects and their parts, the action and manipulation algorithms that plan motion parts and carry out action sequences, etc.). Critically, these are not evaluation criteria for one-shot learning, but rather evaluation criteria for

the learned content and should thus not be conflated with the latter. However, they might be useful in deciding whether knowledge learned quickly through one-shot instructions is sufficiently robust for a task or whether it will have to be altered or augmented to reach the required level of robustness.

There are also interesting open questions about knowledge transfer between robots ensuring that transferred knowledge leads to consistent knowledge bases (because it is still possible, that even though learned knowledge is additive, it could lead to inconsistencies in other systems that do not share exactly the same knowledge bases as the learner), but these will have to be left for another occasion. The important point here is that different from other learning schemes (e.g., neural networks) where new information can alter existing information, the learner itself will remain consistent (to the extent that consistent knowledge is instructed) and can extend its knowledge quickly from a series of instructions.

6 Related Work

While research involving teaching robots through spoken natural language instructions has achieved some successes for both navigation-based tasks [Lauria *et al.*, 2001] and more general tasks [Huffman and Laird, 1995], as well as through more highly structured dialogues which mimic programming [Meriçli *et al.*, 2013], instruction-based one-shot learning is still in its infancy. Current approaches to one-shot learning are very limited with respect to the allowable teaching inputs and usually can only learn simple behaviors, not complex action sequences (e.g., [Cantrell *et al.*, 2011]). And when more complex tasks can be learned through dialogues, additional assumptions are typically made (e.g., the words for the new concepts are already in the speech recognizer, the parser already knows what to do with the word, etc.). Moreover, several of the so-called "one-shot" learning approaches really require multiple trials (e.g., most approaches that focus on visual category, object, and concept learning, in particular, those based on Bayesian approaches, e.g., [Fei-Fei *et al.*, 2006; Lake *et al.*, 2012]).

7 Conclusion

We presented a general one-shot learning scheme together with modifications to various component representations and algorithms in a cognitive robotic architecture that allow for true one-shot learning of new objects and actions from spoken natural language instructions. Specifically, we demonstrate how the proposed mechanisms allowed different robots to recursively learn how to manipulate an object and immediately apply the acquired knowledge. Different from previous work for instruction-based learning, the proposed modifications allow a cognitive robotic architecture to truly acquire new knowledge at every level: from the unknown word and its linguistic properties, to the denoted object concepts and how to manipulate it, to how to perform whole sequences of instructed actions. Moreover, by way of how the newly acquired knowledge is represented and integrated with existing knowledge, it can be shared immediately with other agents running the same architecture.

8 Acknowledgements

This work has in part been funded by ONR grant #N00014-14-1-0149 and #N00014-14-1-0751 to the first author.

References

[Aimetti, 2009] Guillaume Aimetti. Modelling early language acquisition skills: Towards a general statistical learning mechanism. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 1–9. Association for Computational Linguistics, 2009.

[Cantrell *et al.*, 2011] Rehj Cantrell, Paul Schermerhorn, and Matthias Scheutz. Learning actions from human-robot dialogues. In *Proceedings of the 2011 IEEE Symposium on Robot and Human Interactive Communication*, July 2011.

[Dzifcak *et al.*, 2009] Juraj Dzifcak, Matthias Scheutz, Chitta Baral, and Paul Schermerhorn. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation (ICRA '09)*, Kobe, Japan, May 2009.

[Fei-Fei *et al.*, 2006] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4), 2006.

[Gualtieri *et al.*, 2016] Marcus Gualtieri, Andreas ten Pas, Kate Saenko, and Robert Platt. High precision grasp pose detection in dense clutter. *CoRR*, abs/1603.01564, 2016.

[Huffman and Laird, 1995] Scott B. Huffman and John E. Laird. Flexibly instructable agents. *arXiv preprint cs/9511101*, 1995.

[Kirk and Laird, 2014] J. Kirk and J.E. Laird. Interactive task learning for simple games. *Advances in Cognitive Systems*, (3):11–28, 2014.

[Krause *et al.*, 2014] Evan Krause, Michael Zillich, Thomas Williams, and Matthias Scheutz. Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. In *Proceedings of Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[Lake *et al.*, 2012] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Concept learning as motor program induction: A large-scale empirical study. In *Cognitive Science Conference*, 2012.

[Lauria *et al.*, 2001] Stanislao Lauria, Guido Bugmann, Theocharis Kyriacou, Johan Bos, and A Klein. Training personal robots using natural language instruction. *Intelligent Systems, IEEE*, 16(5):38–45, 2001.

[Meriçli *et al.*, 2013] Cetin Meriçli, Steven D Klee, Jack Paparian, and Manuela Veloso. An interactive approach for situated task teaching through verbal instructions. 2013.

[Mohan *et al.*, 2012] Shiwali Mohan, Aaron Mininger, James Kirk, and John E Laird. Learning grounded language through situated interactive instruction. In *AAAI Fall Symposium Series*, pages 30–37, 2012.

[Oosterveld *et al.*, 2017] Bradley Oosterveld, Richard Veale, and Matthias Scheutz. A parallelized dynamic programming approach to zero resource spoken term discovery. In *Proceedings of the 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2017.

[Scheutz *et al.*, 2017] Matthias Scheutz, Evan Krause, Brad Oosterveld, Tyler Frasca, and Robert Platt. Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In *Proceedings of the 16th International Conference on Autoomous Agents and Multiagent Systems*, 2017.

[Scheutz, 2014] Matthias Scheutz. "teach one, teach all" – the explosive combination of instructible robots connected via cyber systems. In *IEEE Cyber*, 2014.