Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Grounded Language Learning: Where Robotics and NLP Meet*

Cynthia Matuszek
University of Maryland, Baltimore County, Baltimore, MD
cmat@umbc.edu

Abstract

Grounded language acquisition is concerned with learning the meaning of language as it applies to the physical world. As robots become more capable and ubiquitous, there is an increasing need for non-specialists to interact with and control them, and natural language is an intuitive, flexible, and customizable mechanism for such communication. At the same time, physically embodied agents offer a way to learn to understand natural language in the context of the world to which it refers. This paper gives an overview of the research area, selected recent advances, and some future directions and challenges that remain.

1 Introduction

Advances in robotics are enabling progressively more sophisticated, capable technologies to reach large consumer populations. Such systems offer unprecedented potential for AI to help in a variety of human-centric applications such as elder care and household maintenance. However, natural, easy-to-use interfaces to such systems, such as those employing natural language, are lagging behind. As robots become more prevalent—and as the need for the services they can offer grows—the importance of allowing non-expert users to interact with them naturally and comfortably increases. Natural language is an excellent modality for end users to give instructions and teach robots about their environments.

At the same time, physically grounded agents provide unique opportunities for language learning. Human language does not exist in isolation; it is learned, understood, and applied in the physical world in which people exist. Understanding symbols and symbolic reasoning has been a core element of artificial intelligence throughout the history of the field [Newell and Simon, 1976; Searle, 1980; Harnad, 1990]. Finding the connection between those symbols and their underlying meanings is the *grounded language acquisition* problem: taking linguistic tokens and learning to interpret them by connecting them to real-world percepts and actions [Mooney, 2008].

* This material is based in part upon work supported by the National Science Foundation under Grant No. 1657469.

The core idea underlying this work is that treating language learning as a physically grounded problem can improve the efficiency and efficacy of both natural language processing and robotics. Intuitively, language can be better learned when presented and interpreted in the context of the world it pertains to, and robots can learn to be more useful and more flexible when language is used to describe and disambiguate the noisy, unpredictable world in which they operate. Learning these groundings on the fly from non-specialists allows a deployed robot to learn and continually update a situation-specific model of language and tasks in its environment.

The work presented in this paper centers on the use case of people teaching a robot about objects and tasks in its environment via unconstrained natural language. The research focus is on formulating and using statistical machine learning approaches to allow robots to gain knowledge about the world from interactions with users, while simultaneously acquiring semantic representations of language about objects and tasks.

Rather than considering these problems separately, they are addressed concurrently by employing a joint learning model that treats a combination of language, perception, and task understanding as strongly associated training inputs. This approach allows each of these channels to provide mutually reinforcing inductive bias, constraining an otherwise unmanageable search space and allowing robots to learn from a reasonable number of ongoing interactions.

There are many approaches to symbol grounding, including formal methods that manually define words [Boteanu *et al.*, 2017], cognitive approaches [Mohan and Laird, 2014], and approaches in which meaning is represented as part of a larger-scale knowledge framework [Williams and Scheutz, 2016] or graphical structure [Arumugam *et al.*, 2017]. This paper focuses on using statistical machine learning methods to learn mappings between words and formal representations of the world [Misra *et al.*, 2017]. from paired corpora of language and sensor data. We give brief examples of research achievements and concepts in using machine learning to understand language, then touch on selected future directions and open problems.

2 Language as Classification

One of the key questions for a statistical approach (and, indeed, most approaches) is: What underlying knowledge representation should be used to represent groundings? For a

significant number of grounded language problems, a key insight is that **understanding words and phrases can be conceptualized as a classification problem**. When language references physical objects [Matuszek *et al.*, 2012] or actions [Tucker *et al.*, 2017], words and linguistic structures may be considered to denote a classifier for which those embodied elements are positive data points.

For example, if a human teacher describes multiple objects in some physical context as 'lemons,' the robot's goal is to subsequently perform tasks that rely on knowing what that word denotes, such as "Get a lemon from the basket." In this case, the formal meaning representation of the NL token 'lemon' is a classifier, and the interpretation of something as a referent of the language is the binary output of that classifier; only positive class members are groundings of the utterance. The classifier is trained using every object that has been referred to as 'lemon' as positive examples.

2.1 Training Grounded Language Classifiers

To implement this approach, a joint model of language and sensor data is induced by combining *language feature extraction* from NL utterances with *perception-based context* that captures the physical setting [Pillai and Matuszek, 2018]. The linguistic and perceptual features used are chosen based on the specific problem domain and sensors.

This formalization is shown in Figure 1. In this overview, perceptually-derived world information, or context C, is interpreted by a perceptual model that encodes the kind of knowledge that the system is expected to learn—in this case features representing color and shape. An encoding of some particular context gives a formal, perceptual *world representation* w. Similarly, a language model (here, a learned semantic parser) is applied to a natural language utterance x in order to produce a formal semantic meaning representation, z.

To learn the connection between language and perception, it is necessary to compute the probability of the correct grounding G conditioned on x and C by summing over the latent structures z and w. This joint learning problem can then be expressed as follows. The language model $P(z|x)$ and perception model $P(w|C)$ are trained independently, but then a joint probability is specified by coupling them with a grounding G, given by $P(G|z, w)$, a conditional probability term that holds the models in agreement. When G is observed, a dependency between z and w is introduced. Interpretation of language into world state is then given by maximizing:

$$P(G, z, w|x, C) = P(z|x)P(w|C)P(G|z, w)$$

In this approach, language groundings are learned by treating words and linguistic features as denotations of visual classifiers. Color and shape are well-defined examples for an approach to language grounding in which linguistic concepts denote visual classifiers, but the treatment is appropriate for a wide range of problems. For example, additional modalities of communication such as deictic gesture can be classified as to referent [Matuszek *et al.*, 2014], and those classifications can be used as an additional form of context in the joint groundings model. The type of learning used is equally flexible. For example, the 'words-as-classifiers' approach to image analysis is a special case of this approach, and has been demonstrated to work well for deep learning over large-scale data sets [Schlangen *et al.*, 2016].

3 Learning Groundings Without Predefined Formalisms

The approach described in this section allows **a grounded language learner that can learn novel language describing novel, unanticipated perceptual inputs**. While there is a significant body of research exploring the learning and use of language in physical agents, much of that work is focused on learning how novel natural language can be interpreted into a predefined formal representation. The formal semantic representation is usually defined with respect to the robot's perceptual and actuator capabilities. As a result, natural language inputs may be learned during training, but the formal output is limited to a fixed grammar, which is inconsistent with the goal of building adaptable robots that interact with people in unpredictable environments.

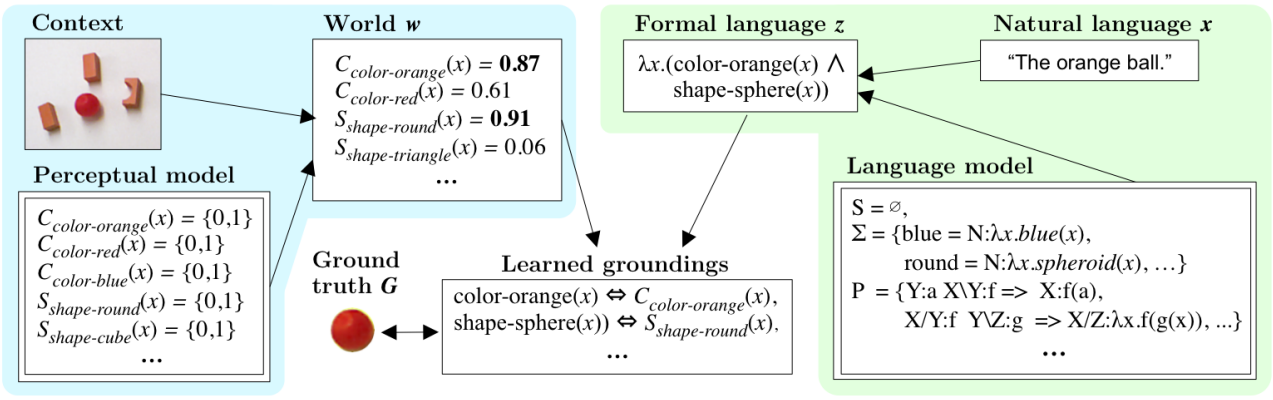

Figure 1: A joint modeling approach to classifier-based language grounding. *Blue, left:* The system's perceived world-state, or *context* (C), is interpreted in terms of a perceptual model that translates from raw percepts to a meaningful world description w. *Green, right:* natural language input x is interpreted into a computer-meaningful formal representation z using a language model. In this example, the system has learned that the parse $\lambda x.color\text{-}orange$ is applicable to percepts that produce a high value from the color classifier $C_{color\text{-}orange}(x)$.

3.1 Learning About Novel Physical Attributes

In the language acquisition example in Figure 1, the system learns to connect natural language terms x to formal meaning representations z that can be interpreted, through perception, into a world state w. To provide a concrete example, on encountering the new word x = 'ball,' the robot learns that things described in that way can be formally interpreted as $z = \lambda \texttt{x.shape-round(x)}$, which is associated with the visual classifier $w = C_{shape\text{-}sphere}$. If there is no formal symbol `shape-round(·)` with an associated classifier, a physical grounding for the shape word 'ball' cannot be learned.

There are two components to addressing this limitation. First, an embodied agent learning from language must add new formal language tokens z on the fly as language is encountered. Second, new mechanisms of interpreting the world w must be created and associated with the new tokens—that is, when novel language is encountered, one or more new classifiers must be created. Broadly, symbols are generated as needed in the formal representation of the world, and classifiers are created to represent *hypotheses* about possible semantic meanings of the token.

In [Matuszek *et al.*, 2012], a CCG-based semantic parser was used to learn novel language about completely novel concepts. When new English tokens were encountered, a set of classifiers was created that represent hypotheses of the category type of the word (such as shape, a color, a nonvisual concept, or a synonym for another term). When one classifier reached a certain level of descriptive strength, the others were pruned and a new terminal linked to that classifier was inserted into the formal meaning representation language. (Following the example above, on encountering x = 'ball' with new percepts, $z = \lambda \texttt{x.shape-NEW-1(x)}$ and $z = \lambda \texttt{x.color-NEW-1(x)}$ would be added to the language model and associated with the new visual classifiers $w = C_{shape\text{-}new\text{-}1}$ and $w = C_{color\text{-}new\text{-}1}$). An example is shown in Figure 2.

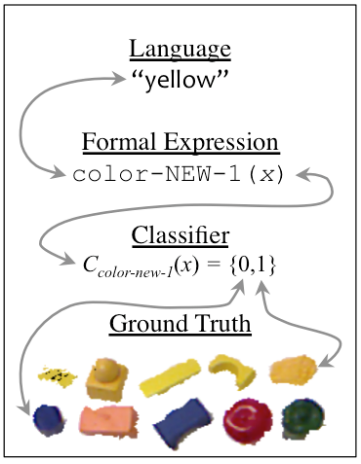

Figure 2: An example of the learning process when a novel natural language term is encountered. In practice, usually each term would cause the creation of a large number of competing classifiers, each of which represents a hypothesized grounding in perceptual space.

This approach is similar to research on learning object characteristics from exploratory robot interactions [Sinapov *et al.*, 2016; Thomason *et al.*, 2018] and, particularly, attempts to extend language models [She and Chai, 2016], but with the addition of learning more generalized concepts of perception. However, we build a joint model of language and observation directly rather than learning over and populating an underlying representation.

This approach allows for understanding of new lexical items in language and new objects and characteristics in the environment. The models learned are still subject to limitations; most notably, the categories of types that can be learned, such as 'color,' are still predefined and have appropriate classifier features designed beforehand (see Section 5 for a discussion of relaxing this constraint). Nonetheless, this work represents significant progress towards the goal of autonomous grounded language learning.

4 Semantic Analysis of Physical Space

Much of the existing work on grounded language performs some form of learning over paired corpora of language and context. As the learning problem becomes less constrained, mappings must be found between complex utterances and a complex physical space, leading to a large search space.

This is a bootstrapping problem: knowing the meaning of language can help narrow the search in the grounding space and vice versa. However, **even before a grounding is established, it is possible to use semantic analysis of language to learn more about the physical world**.

The intuition underlying this claim comes from our ability to perform extensive analysis of completely ungrounded language. Documents can be clustered into topics, salient terms can be selected from utterances, and distance metrics can be evaluated with no understanding of how that language connects to the physical world. If some language does have embodied meaning, relationships within that language may also hold in the physical context.

4.1 Negative Visual Examples from Language

In Section 3.1, the positive examples used in classifier training are objects that are described using a particular language label. Everything described as "a lemon" to a robot is treated as a positive example for a perceptual classifier. However, most learning approaches depend on having negative examples, which are notoriously difficult to obtain from unconstrained speech. Unless prompted explicitly, it is unusual for someone to describe an lemon as "not an apple." In addition, positive labels cannot be assumed to be exhaustive—if a person says "This is a lemon," that object is not necessarily a negative example for the word 'yellow,' even though it was not described as yellow in that sentence.

However, while a single label may not provide sufficient information to assume that something is a negative example of a word or concept, a set of descriptions sufficient for learning embodied language is likely to be more complete. This makes it possible to find negative *perceptual* examples for classifier training by comparing the *linguistic* descriptions of those objects, even when the meaning of that language has not yet been learned.

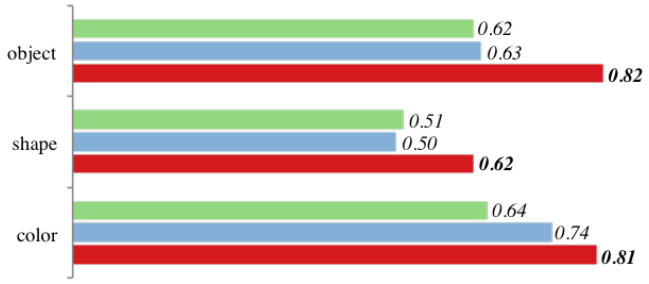

Figure 3: Average accuracy of classifiers for three different categories of embodied object attributes. Negative data is selected by using randomly chosen non-positively-labeled objects (green, top), all non-positive objects (blue, middle), or using semantic similarity (red, bottom). Using an informed distance metric to select negative examples improves performance in every category.

To test this intuition, a set of descriptions of common food objects was collected. Descriptions of each object were concatenated into an unordered document, which was encoded into the well-known Paragraph Vector representation [Mikolov *et al.*, 2013]. Cosine similarity between these vectors was used to find semantically dissimilar objects, which were then used as negative examples for classifier training [Pillai and Matuszek, 2018]. This is consistent with the fully unsupervised label identification of [Roy, 2002], but uses document similarity metrics rather than term clustering.

This similarity metric was evaluated both directly by annotators, where it was found to be generally consistent with human judgments, and indirectly by using semantic distance to select negative examples for visual classifier training, which improved classifier accuracy significantly (Figure 3).

Overall, this work bears out the insight that linguistic similarity is closely tied to perceptual similarity in the shared grounding space. This allows for unsupervised selection of negative examples for learning, but also implies there may be other ways inputs to a grounded language learning system may be evaluated before mappings between language and world context are determined. The same reasoning may apply equally well to analyzing sensor data in order to discover relationships within elements of language.

5 Open Challenges

To be customizable and flexible in unpredictable environments, robots will need be able to learn new concepts from their users efficiently. Current grounded language learning often requires hundreds of utterances to learn concepts. Such a data set is reasonable in machine learning and natural language processing, especially given the complexity of the task. However, end users are unlikely to provide that many labels for a single task or learning target. User training will therefore need to be supplemented by other mechanisms from the rich existing corpus of research on learning. The first two suggested research directions directly address this question, while the third depends on solving it.

Human-Robot Dialog for Active Learning One way of improving efficiency is to make use of active learning. Asking questions about specific topics often leads to faster learning than receiving data sequentially, in part because queries can be selected using a variety of information-theoretic methods [Settles, 2012]. To go beyond learning passively from a human teacher, it is necessary for a robot to ask questions and otherwise direct its own learning, which will ultimately require incorporation of models of dialog and discourse.

This is particularly relevant for problems which require local or customized labeling. For example, to learn what a lemon is, asking Mechanical Turk or doing an image search is almost always correct. By comparison, if a person is trying to ask a robot to retrieve "my mug," the correct language grounding is very specific to the environment.

Sharing Learned Models among Agents In a collaborative setting with multiple robots, users should not be expected to teach the same things repeatedly. Learned models of language grounding should, as much as possible, be shared among all robots in the environment. Similarly, robots that are not physically co-located should be able to share and combine learned models when applicable. Sharing data among robots with different sensors further increases complexity, since learned models cannot be transferred directly.

Transferring learned models of grounded language between heterogeneous agents is understudied, with the exception of systems that use shared pre-existing knowledge into which concepts can be grounded [Bozcuoğlu *et al.*, 2018]. The question of how to transfer purely learned models draws from transfer learning, domain and feature adaptation [Long *et al.*, 2015], sensor difference modeling, shared language evolution [Spranger and Steels, 2015], and other active areas of research.

Beyond Objects and Tasks In Section 3, the inability of the learner to generalize to entirely novel visual categories is mentioned. Color, shape, and object type are only a few of the many different categories people may use to describe objects. While this specific example may be addressed in part by using generic visual features [Donahue *et al.*, 2014] and more sophisticated learning, it is representative of a greater problem, which is that concrete language about objects and actions is only a small part of embodied language use. The biggest challenge that lies ahead is moving beyond descriptive language to more general vocabulary and higher-level representations of meanings, including more abstract, pragmatic, and intent-driven language interpretations. This will require richer representations of contextual world information, including models of time [Paul *et al.*, 2017], people and intent.

6 Discussion

Language-using robots must learn how words are grounded in the noisy, perceptual world in which a robot operates, and natural language systems can benefit from the rich contextual information provided by sensor data about the world. Despite an extensive and growing body of research, a significant number of challenges still need to be addressed before we see language-controlled robot assistants deployed in human spaces. This goal will continue to drive technical advances in robotics, natural language processing, machine learning, cognitive science, and other areas.

References

[Arumugam *et al.*, 2017] Dilip Arumugam, Siddharth Karamcheti, Nakul Gopalan, Lawson L.S. Wong, and Stefanie Tellex. Accurately and efficiently interpreting human-robot instructions of varying granularities. In *Robotics: Science and Systems (RSS)*, 2017.

[Boteanu *et al.*, 2017] Adrian Boteanu, Jacob Arkin, Siddharth Patki, Thomas Howard, and Hadas Kress-Gazit. Robot-initiated specification repair through grounded language interaction. In *AAAI Fall Symposium on Natural Communication for Human-Robot Collaboration*, 2017.

[Bozcuoğlu *et al.*, 2018] Asil Kaan Bozcuoğlu, Gayane Kazhoyan, Yuki Furuta, Simon Stelter, Michael Beetz, Kei Okada, and Masayuki Inaba. The exchange of knowledge using cloud robotics. *IEEE Robotics and Automation Letters (RA-L)*, 3(2):1072–1079, 2018.

[Donahue *et al.*, 2014] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proc. of the 31st International Conference on Machine Learning*, 2014.

[Harnad, 1990] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1), 1990.

[Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *Proc. of the 32nd International Conference on Machine Learning*, 2015.

[Matuszek *et al.*, 2012] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proc. of the 29th International Conference on Machine Learning (ICML)*, 2012.

[Matuszek *et al.*, 2014] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proc. of the 28th National Conference on Artificial Intelligence (AAAI)*, March 2014.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[Misra *et al.*, 2017] Dipendra Misra, John Langford, and Yoav Artzi. Mapping instructions and visual observations to actions with reinforcement learning. In *Proc. of the 34th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[Mohan and Laird, 2014] Shiwali Mohan and John E Laird. Learning goal-oriented hierarchical tasks from situated interactive instruction. In *Proc. of the 28th National Conference on Artificial Intelligence (AAAI)*, 2014.

[Mooney, 2008] Raymond J. Mooney. Learning to connect language and perception. In Dieter Fox and Carla P. Gomes, editors, *Proc. of the 23rd Conference on Artificial Intelligence (AAAI)*, pages 1598–1601, 2008.

[Newell and Simon, 1976] Allen Newell and Herbert A Simon. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 1976.

[Paul *et al.*, 2017] Rohan Paul, Andrei Barbu, Sue Felshin, Boris Katz, and Nicholas Roy. Temporal grounding graphs for language understanding with accrued visual-linguistic context. In *Proc. of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[Pillai and Matuszek, 2018] Nisha Pillai and Cynthia Matuszek. Unsupervised end-to-end data selection for grounded language learning. In *Proc. of the 32nd National Conference on Artificial Intelligence (AAAI)*, New Orleans, USA, February 2018.

[Roy, 2002] Deb K Roy. Learning visually grounded words and syntax for a scene description task. *Computer Speech & Language*, 16(3):353–385, 2002.

[Schlangen *et al.*, 2016] David Schlangen, Sina Zarrieß, and Casey Kennington. Resolving references to objects in photographs using the words-as-classifiers model. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016.

[Searle, 1980] John R Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(03):417–424, 1980.

[Settles, 2012] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[She and Chai, 2016] Lanbo She and Joyce Y Chai. Incremental acquisition of verb hypothesis space towards physical world interaction. In *Proc. of the 54th Meeting of the Association for Computational Linguistics (ACL)*, 2016.

[Sinapov *et al.*, 2016] Jivko Sinapov, Priyanka Khante, Maxwell Svetlik, and Peter Stone. Learning to order objects using haptic and proprioceptive exploratory behaviors. In *Proc. of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[Spranger and Steels, 2015] Michael Spranger and Luc Steels. Co-acquisition of syntax and semantics—an investigation in spatial language. In *Proc. of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, Buenos Aires, Argentina, 2015.

[Thomason *et al.*, 2018] Jesse Thomason, Jivko Sinapov, Raymond J Mooney, and Peter Stone. Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In *Proc. of the 32nd National Conference on Artificial Intelligence (AAAI)*, 2018.

[Tucker *et al.*, 2017] Mycal Tucker, Derya Aksaray, Rohan Paul, Gregory J Stein, and Nicholas Roy. Learning unknown groundings for natural language interaction with mobile robots. In *International Symposium on Robotics Research (ISRR)*, 2017.

[Williams and Scheutz, 2016] Tom Williams and Matthias Scheutz. A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. In *Proc. of the 30th Conference on Artificial Intelligence (AAAI)*, 2016.