Symbol, Conversational, and Societal Grounding with a Toy Robot

Casey Kennington & Sarah Plane
Department of Computer Science
Boise State University
firstnamelastname@boisestate.edu

Abstract

Essential to meaningful interaction is grounding at the symbolic, conversational, and societal levels. We present ongoing work with Anki's Cozmo toy robot as a research platform where we leverage the recent words-as-classifiers model of lexical semantics in interactive reference resolution tasks for language grounding.

Introduction

Grounding is essential in meaningful interaction (Clark, 1996; DeVault, Oved, and Stone, 2006; Schlangen, 2016). Grounding is a term used to denote several distinct aspects of language and communication. We take up three aspects here, though Lücking and Mehler (2014) have identified others: (1) ***symbol grounding*** (Harnad, 1990) where aspects of language are connected with aspects of the things that language denotes, such as visual features (e.g., the word *red* is linked to aspects of visual perception), (2) ***conversational grounding*** (Clark, 1996) where aspects of events that occur between two or more people are recorded for later use and recall, and (3) ***societal grounding*** (DeVault, Oved, and Stone, 2006) which connects symbol and conversational grounding with the accepted uses of language used in a particular language community. These aspects of grounding are summarized in **Figure 1**.

All three types of grounding overlap with each other which allows for meaningful communication. To illustrate, consider a child who sees a pine cone and experiences firsthand its visual and tactile features. A nearby adult says "that's a pine cone" because the adult has already established through *societal grounding* that "pine cone" denotes such an object. By hearing this, the child learns through *symbol grounding* that certain visual and tactile features are linked to the words "pine cone" and both the child and adult establish through *conversational grounding* the event that the child has heard the denotation. Grounding on all three levels in this example occurred through an interactive process which establishes grounding of linguistic meaning between words and the perceived world, between individuals, and between individuals and language communities at large.

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

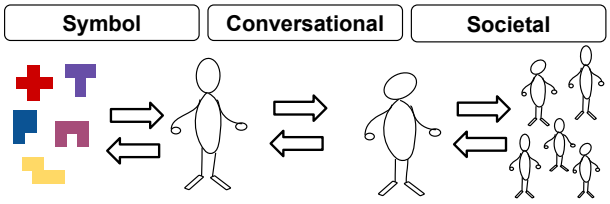

Figure 1: Comparison of grounding types. An individual perceives objects and grounds symbols–conventional denotations for those objects–interactively through conversational grounding with someone else. The conventional denotations are socially grounded through interaction with members of a language community.

It is through this face-to-face spoken conversation setting, the basic and primary setting of language (Fillmore, 1981), where interlocutors can denote objects (often with pointing gestures) in their shared environment which forms the foundation for language acquisition (McCune, 2008), and from which words denoting more abstract concepts are built. A key question is how semantic meaning should be represented and acquired through this co-located grounding process.

We present ongoing work on grounding with a toy robot. We leverage the *words-as-classifiers* (WAC) model of lexical semantics (Kennington and Schlangen, 2015), recently yielding state-of-the-art results in a reference resolution task using real images and deep learning to represent the object regions (Schlangen, Zarriess, and Kennington, 2016). The model is flexible, interpretable, and simple in that each word is treated as its own classifier.

Background & Related Work

This work builds on related work in co-located, language grounding (Roy, 2005) and recent work in grounded language semantic learning in various tasks and settings, notably learning descriptions of the immediate environment (Walter et al.); navigation (Kollar et al., 2010), and verbs (She and Chai, 2016). A common task to evaluate models convincingly is reference resolution to real-world objects. In most cases, the set of candidate objects are simultaneously visible within a scene. This project goes beyond this work: the robot's limited perspective allows it to see one or two objects in front of it at a time. The robot must settle on an object potentially without being able to see all of

the objects–arguably a more realistic task (similar in spirit to navigation tasks such as Han and Schlangen (2017)) and language grounding setting; i.e., the two interlocutors do not share the same perspective. Moreover, previous work has assumed that humans will treat and interact with the robot in such a way that the robot will perform symbolic grounding, but it's not necessarily the same setting where humans acquire their first language: as children. It has been shown that humans treat robots differently depending on how they perceive the robot's gender (Eyssel and Hegel, 2012), social categorization (Eyssel and Kuchenbrandt, 2012), personality (Tay, Jung, and Park, 2014), and intelligence (Novikova et al.). In this work, we take this knowledge into account by using a robot that is more likely to be perceived and treated as a child by humans.

We leverage the recently released *Anki Cozmo* robot as a platform to research spoken language grounding using the WAC model. The Cozmo robot (example in Figure 2) is a small robot that has a well-documented SDK and growing community support.[1] The robot itself has arms that can lift or push small objects, track wheels for movement, a simple text to speech synthesizer (i.e., the robot itself has a small speaker), and a black and white camera which is embedded in a small movable head that has animated eyes. Some built-in capabilities include facial recognition and some basic functionality for detecting specific types of objects (e.g., some blocks that are included with the robot). The hardware that makes up the robot offer enough degrees of freedom to make it a flexible and versatile research platform; the size and affordances of the robot make it manageable for researchers who are not roboticists. The SDK is written in Python making it easily extensible by the myriad of machine learning and natural language processing libraries. Importantly, the robot is affordable (under $180) and very portable. Our group has already acquired two Cozmo robots and we have found them to be accessible, usable, and flexible, even for fairly novice programmers.

Figure 2: Cozmo robot

Language Grounding: Our Approach

We follow a simple strategy for language grounding and acquisition: assuming that the system can *detect* (i.e., *not* recognize) objects–a precondition for learning words that denote objects (Bloom, 2000, p.61)–we apply the WAC model to learn novel words with minimal interactions. We also take into account the essential pragmatic scaffolding that must be in place for language grounding to take place: the Cozmo robot can track a person's face and facial features which we will leverage for positive and negative feedback when the robot performs certain tasks that involve word usage. Learning in real-time interaction is no trivial matter, but here the Cozmo platform is useful: instead of using potentially complicated pointing recognition, we can assume that an object under discussion is the one directly in front of the robot. Evaluation of our model can be done by a reference resolution task similar to a game of fetch where a human player refers to an object and the robot must find that object as soon as possible.

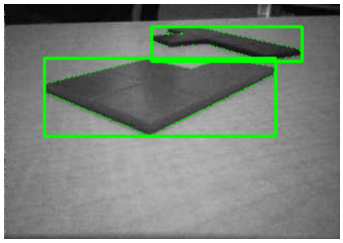

Figure 3: simple object detection from Cozmo's perspective

Our preliminary work using the Cozmo SDK has shown promise. We have applied some of our own object detection to the camera feed using OpenCV (see Figure 3) as well as the YOLO object detection model (Redmon and Farhadi, 2016). Having detected the objects, we can extract low-level object features for the WAC model which does the object recognition and grounds the words in the referring expressions to the objects. In our preliminary experiments, the WAC model selects the correct object about half of the time with minimal training data. Supporting the WAC model will be additional standard dialogue system modules, such as a conversational speech recognizer and a dialogue manager. We build off of our own previous work for evaluating conversational speech recognition (Baumann et al., 2016) to determine the best option, and dialogue management in an interactive setting with a robot (Kennington et al., 2014) using the OpenDial toolkit (Lison and Kennington, 2015).

The outcome of this research will be improved understanding of how lexical semantic meaning is learned and represented through natural interaction. We are exploring a setting where Cozmo interacts with children to perform simple tasks, as Cozmo is marketed as a toy for children to learn procedural 'coding'. In our observations, children find Cozmo aesthetically pleasing and enjoyable to interact with. We anticipate several challenges: for WAC, the robot's integrated camera has a limited, black and white perspective (i.e., the WAC model cannot make direct use of color information in this setting). Verb learning of robot actions will also be challenging (e.g., *move*, *pick up*, *push*, etc.); we will build off of very recent work by She and Chai (2017). The task and setting will also challenge the WAC model due to differences in perspectives (e.g., the word *left* will mean something different depending on the perspective of the users and the robot).

Though we are not roboticists, we feel it important to bring together dialogue systems and robotics researchers to work towards natural, spoken interaction with robots.

[1] https://developer.anki.com/en-us

References

Baumann, T.; Kennington, C.; Hough, J.; and Schlangen, D. 2016. Recognising Conversational Speech: What an Incremental ASR Should Do for a Dialogue System and How to Get There. In *Proceedings of the International Workshop Series on Spoken Dialogue Systems Technology (IWSDS) 2016*.

Bloom, P. 2000. *How children learn the meanings of words.* The MIT Press.

Clark, H. H. 1996. *Using Language*. Cambridge University Press.

DeVault, D.; Oved, I.; and Stone, M. 2006. Societal Grounding Is Essential to Meaningful Language Use. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, 747. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Eyssel, F., and Hegel, F. 2012. (S)he's Got the Look: Gender Stereotyping of Robots1. *Journal of Applied Social Psychology* 42(9):2213–2230.

Eyssel, F., and Kuchenbrandt, D. 2012. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology* 51(4):724–731.

Fillmore, C. J. 1981. Pragmatics and the description of discourse. *Radical pragmatics* 143–166.

Han, T., and Schlangen, D. 2017. Grounding Language by Continuous Observation of Instruction Following. In *Proceedings of EACL: Volume 2, Short Papers*, volume 2, 491–496. Association for Computational Linguistics.

Harnad, S. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42(1-3):335–346.

Kennington, C., and Schlangen, D. 2015. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 292–301. Beijing, China: Association for Computational Linguistics.

Kennington, C.; Funakoshi, K.; Takahashi, Y.; and Nakano, M. 2014. Probabilistic multiparty dialogue management for a game master robot. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14*, 200–201. Bielefeld, Germany: ACM.

Kollar, T.; Tellex, S.; Roy, D.; and Roy, N. 2010. Toward understanding natural language directions. *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on* 259.

Lison, P., and Kennington, C. 2015. A Toolkit for Developing Spoken Dialogue Systems with Probabilistic Rules. In *Proceedings of SEMDial*.

Lücking, A., and Mehler, A. 2014. On Three Notions of Grounding of Artificial Dialog Companions. *Science, Technology & Innovation Studies* 10(1).

McCune, L. 2008. *How Children Learn to Learn Language*. Oxford University Press.

Novikova, J.; Dondrup, C.; Papaioannou, I.; and Lemon, O. Sympathy Begins with a Smile, Intelligence Begins with a Word: Use of Multimodal Features in Spoken Human-Robot Interaction. 86–94.

Redmon, J., and Farhadi, A. 2016. YOLO9000: Better, Faster, Stronger. *arXiv preprint arXiv:1612.08242*.

Roy, D. 2005. Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciences* 9(8):389–396.

Schlangen, D.; Zarriess, S.; and Kennington, C. 2016. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *Acl*, 1213–1223.

Schlangen, D. 2016. Grounding , Justification , Adaptation : Towards Machines That Mean What They Say. In *Proceedings of SemDial*, 35–43.

She, L., and Chai, J. Y. 2016. Incremental Acquisition of Verb Hypothesis Space towards Physical World Interaction. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)* 108–117.

She, L., and Chai, J. Y. 2017. Interactive Learning of Grounded Verb Semantics towards Human-Robot Communication. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1634–1644.

Tay, B.; Jung, Y.; and Park, T. 2014. When stereotypes meet robots: The double-edge sword of robot gender and personality in human-robot interaction. *Computers in Human Behavior*.

Walter, M. R.; Hemachandra, S.; Homberg, B.; Tellex, S.; and Teller, S. A Framework for Learning Semantic Maps from Grounded Natural Language Descriptions.