RIPLEY, HAND ME THE CUP!
(SENSORIMOTOR REPRESENTATIONS FOR GROUNDING WORD MEANING)

Deb Roy, Kai-Yuh Hsiao, Nikolaos Mavridis, and Peter Gorniak

Cognitive Machines Group
MIT Media Laboratory
www.media.mit.edu/cogmac

ABSTRACT

People leverage situational context when using language. Rather than convey all information through words, listeners can infer speakers' meanings due to shared common ground [1, 2]. For machines to engage fully in conversation with humans, they must also link words to the world. We present a sensorimotor representation for physically grounding action verbs, modifiers, and spatial relations. We demonstrate an implementation of this framework in an interactive robot that uses the grounded lexicon to translate spoken commands into situationally appropriate actions.

1. SITUATED SPOKEN LANGUAGE

Speakers use spoken language to convey meaning to listeners by leveraging situational context. Context includes many levels of knowledge ranging from fine grain details of shared physical environments to shared cultural norms. As the degree of shared context decreases between communication partners, the efficiency of language also decreases since the speaker is forced to explicate increasing quantities of information that could otherwise be left unsaid. A sufficient lack of common ground can lead to communication failures.

If machines are to engage in meaningful, fluent, situated spoken dialog, they must be aware of their situational context. As a starting point, we focus our attention on physical context. A machine that is aware of where it is, what it is doing, the presence and activities of other objects and people which are in its vicinity, and salient aspects of recent history, can use these contextual factors to understand spoken language in a context-dependent manner.

A concrete example helps illustrate how a machine can make use of situational context. Consider a speech interface to the lights in a room[1]. If a person simply says, "Lights!", the appropriate action will depend on the current state of the light. If it is already on, the command means *turn off*, but if it is already off, it means the opposite. In this simple example, the language understander needs access to a single bit of situational context, the current state of the light. Consider a slightly richer problem, still in the domain of the light controller. How should the spoken command *softer* be interpreted by the light? Perhaps the simplest solution would be to decrease the intensity of the light by a fixed amount. Although this solution might be functional, it is not necessarily the most natural. In contrast to a fixed-interval solution, a person responding to this request would be likely to decrease the intensity by an amount that is a function of the intensity of light in the room at the time of the request. In general, many sources of light (e.g., from a setting sun) may contribute to the total ambient light in the room. For a machine to leverage this situational information, we could add a light sensor to the controller that is able to monitor ambient lighting conditions. A *context-dependent* interpretation of "softer" could then be defined.

1.1. Language Grounding

A necessary step towards creating situated speech processing systems is to develop representations and procedures that enable machines to *ground* the meaning of words in their physical environments. In contrast to dictionary definitions that represent words in terms of other words (leading, inevitably, to circular definitions for all words), grounded definitions anchor word meanings in non-linguistic primitives. Assuming that a machine has access to its environment through appropriate sensory channels, language grounding enables machines to link linguistic meanings to elements of the machine's environment.

From environmentally aware light controllers to car navigation systems that see the same visual landmarks as the driver, the idea of a context-grounded speech processing is the tip of a very large iceberg. We believe that a large class of spoken language understanding applications may benefit from language grounding. We will refer to this class of systems as having *grounded* semantics in light of the explicit links of semantic representations to the machine's physical

[1] Ignoring, for the moment, the difficult issue of microphone placement and background noise that would also need attention.

world.

To create grounded systems, questions of representation and learning arise. How should the physical environment of the machine be represented to facilitate semantic grounding? How can a machine automatically acquire such knowledge structures? In this paper, we present our approach to the grounded representation of object properties (*blue, heavy, small, soft*), action verbs (*lift, move*), and spatial phrases (*in front of me, to your left*). These representations have been implemented for a small vocabulary speech understanding system embodied in an manipulator robot that is able to engage in "face to face" speech mediated interactions with a human communication partner.

2. RIPLEY: EMBODIMENT OF A GROUNDED DIALOG SYSTEM

For a light controller, it might seem sufficient to represent the world through a single scalar value that measures ambient light. But as suggested above, perhaps the number of people in the room also matters, including their locations, what they are doing, what other objects are in the space, and so forth. A car may need to know even more about its city if it is to contextualize language. To drive our research, we have chosen to build conversational service robots, which gives rise to a wide range of words and speech acts. In particular, actuated robots force us to confront motor representations and active perception, which are crucial to our definition of virtually all lexical semantics including seemingly non-motor terms such as *red* (as we shall see, color terms are with respect to motor procedures used to look at objects, a precursor to measuring the color of an object). With a growing interest in domestic robots, conversational interfaces in this domain may be of practical value. Beyond robots, we believe that the underlying principles and methods that we are developing will transfer to other domains of immediate practical interest as well.

Our current work is based on a robotic manipulator called Ripley (Figure 1). Ripley has 7 degrees of freedom (DOFs), enabling it to manipulate objects in a 3 foot radius workspace. The robot may be thought of as an articulated torso terminating with a head that consists of a "mouth" (a one DOF gripper), stereo color cameras, microphones, and an inertial sensor. Ripley's physical structure provides a foundation for grounding verbs related to manipulation. The placement of cameras on the head provides the means for effecting shifts of visual perspective. Ripley's perceptual channels enable grounding of various concepts related to objects, their properties, and spatial relations. Other concepts, such as those related to body movement through space, do not arise since Ripley is not a mobile robot, and are thus not addressed.

Motor control is performed through trajectories of target joint configurations. An elastic force model is used to provide compliant motion [3]. Image processing relies on color based separation of objects from the background table on which all objects must lie [4]. The robot's proprioception system includes touch sensors that line the tips of the gripper as well as position and force sensors embedded in each actuated joint.

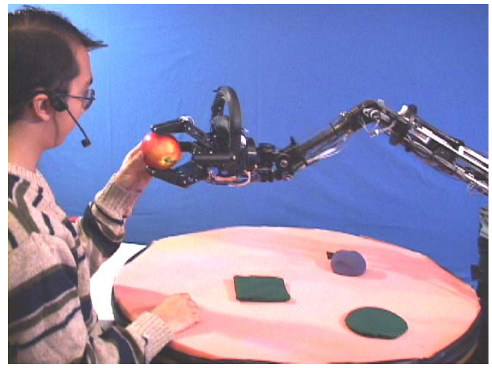

Fig. 1. Ripley hands an apple to its human communication partner in response to the phrase "Hand me the thing on your left".

3. A MENTAL MODEL FOR OBJECT PERMANENCE AND PERSPECTIVE SHIFTS

As we move around our direction of gaze, objects come in and out of sight, but our conception of objects stays stable. The same is true for Ripley as it moves, since its cameras are placed on either side of its gripper. A mental model provides a stable representation of the physical environment that factors out shifts of perspective (Figure 2) [3]. The model consists of a Newtonian physics rigid body simulator. As Ripley moves about its work space, the location of objects and their properties (currently just size and color) are relayed to the mental model. A hysteresis function is used to smooth sensory data. Persistent evidence for the presence, movement, or disappearance of objects drives updates in the mental model. Ripley's own body is also simulated in the mental model. A face tracker [5] detects and tracks the location of the human communication partner, whose location is represented in the simulator using a simple rigid body model.

By constructing a 3-D model of the environment, Ripley is able to "imagine" its environment from any point of view, including the human's point of view, by moving a synthetic camera that uses projective geometry to construct an image of the world from the camera's perspective. Figure 2 shows a view using synthetic vision from the human's perspective as Ripley looks at two objects on a table.

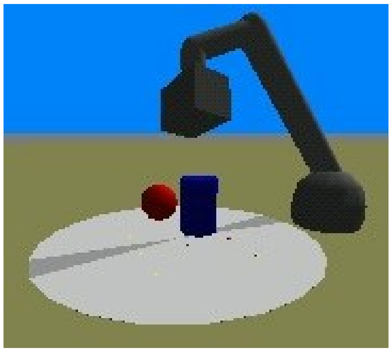

Fig. 2. Ripley's mental model from the perspective of its human communication partner.

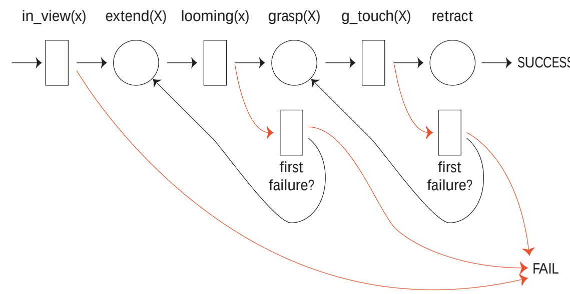

Fig. 3. A sensorimotor network that encodes the semantics of *pickup*.

4. A GROUNDED LEXICON

We have developed a set of sensorimotor representations that ground the meaning of a small lexicon. Ripley uses a standard chart parser to parse spoken input based on this lexicon and take actions. This section presents the structure of the lexicon, while the next section describes parsing and semantic composition.

4.1. Verbs = Sensorimotor Networks

The meaning of manipulation verbs (*lift*, *pick up*, *touch*) are grounded in *sensorimotor networks* (SN). SNs can be used to execute actions on the robot (in that sense, they may be thought of as plan fragments), but they also serve as a representational substrate for the semantics of verbs, and as we shall see, modifiers that are linked to verbs.

A SN is defined by a linked set of *perceptual conditions* and *motor primitives*. Figure 3 shows the SN for *pickup*. Perceptual conditions are indicated by rectangles, motor primitives by circles. Verbs expect a single argument x, the patient of the verb [2]. The main execution path of this SN is a single alternating sequence of perceptual conditions and motor primitives. The *pickup* SN may be interpreted as (1) ensure x is in view, (2) extend head until x is visually looming (recall that Ripley's cameras are mounted next to the gripper), (3) grasp with the gripper until the gripper touch sensors are activated, and finally, (4) retract. Errors can be sensed at each perceptual condition. The default behavior on all errors is to retry the previous motor action once, and then give up. All SNs terminate in either a **success** or **failure** final state.

[2] In ongoing work, we are expanding our formalism to accept agents, instruments, and manner arguments.

4.2. Modifiers = Sensorimotor Expectations

Modifiers, such as color, shape, and weight, are defined with respect to an underlying SN. Figure 4 illustrates the representation of *heavy* and *light*. This structure captures the commonsense notion that something is heavy if it is difficult to lift. The SN (bottom) grounds the meaning of *lift*. The dashed line indicates a *projection function* that projects the execution of an SN into a low dimensional feature space. In this case, the projection function accumulates joint forces during the execution of the **retract** motor primitive, effectively weighing the patient of *lift*. The meaning of *heavy* and *light* are grounded as distributions of expected values with respect to this projection of the underlying SN. These distributions are referred to as *activation functions*. To determine how well a word fits an object, the SN underlying that word must be executed and projected using the associated projection function. The activation function associated with the word is evaluated at the projected the point to determine how well the word fits the object. Since activation functions are continuous, all scores are continuously graded.

Categorical distinctions (e.g., determining whether an object is blue or not, as a binary decision) are made using a simple voting mechanism. Within a feature space, the most activated function determines the category label of the object. This rigid treatment of categorical boundaries is problematic since in natural language use, boundaries shift as a function of contextual factors such as other objects present, the kind of object, etc. In future work, we plan to refine this aspect of the representation.

The grounding of color terms closely parallels weight terms (Figure 5). In place of *lift*, color terms are defined in terms of the SN associated with *lookat*, which, when executed, causes Ripley to center the object x in the robot's visual field. The projection function computes the average value of color in all pixels of the visual region corresponding to the object. Color terms such as *green* and *orange* are defined as two-dimensional Gaussian distributions within this projected feature space.

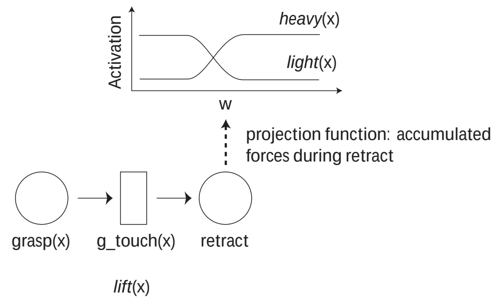

Fig. 4. The meaning of *heavy* and *light* are grounded in expected resistance measuring while lifting an object.

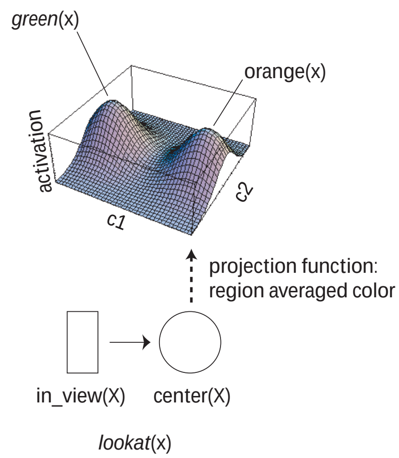

Fig. 5. The meaning of *green* and *orange* are grounded in expected distributions of context-normalized color space measured by looking at an object.

The representation of color clearly demonstrates a crucial difference between our approach and other representations that have been proposed for linking vision and language. Rather than treat color as a passive perceptual association, our approach explicitly links color to the active methods by which the machine can measure the property. As we discuss below, these links create the basis for not only language understanding, but also planning in order to resolve ambiguities and acquire additional non-linguistic information to understand language in context.

Shape descriptors are grounded using histograms of local geometric feature, described in [6]. The histograms are generated using a projection function defined in terms of the same SN as color terms (*lookat*).

4.3. Spatial Relations and Perspective Shifting

To ground spatial words (e.g., *above, to the left of*) in our past work with two-dimensional virtual worlds (cf. [7]), we have used Regier's spatial features [8], which take into account the relative shape and size of objects. Since Ripley's mental model is three dimensional, we use projective transforms to capture 2-D views of the mental model (using synthetic vision). Regier's features are then computed on the 2-D image. In Regier's models, and our previous work, the perspective of the viewer has always remained fixed, assuming a first person perspective. In Ripley's mental model, the synthetic camera can be moved to any 3-D location and orientation. Using this perspective shift operation, the semantics of *my left* versus *your left* can be differentiated by using the word *my*, in this linguistic context, as a trigger for positioning the synthetic camera. Ripley's proprioceptive system guides the placement of the camera for first person perspectives, and the face-tracker driven human model enables shifting to the human's point of view.

This arrangement of perspective shifting enables many interesting behaviors. For example, Ripley is able to detect objects that are not in view from the human's perspective (due to occlusion from other objects, for example). Although we have yet to make use of this knowledge, one possible use would be to generate appropriate forms of reference taking into account points of view.

5. IMPLEMENTATION OF A GROUNDED SPEECH UNDERSTANDING SYSTEM

Using the SN and projection function representation, we have encoded a small vocabulary of words that cover verbs (*pickup, touch*, etc.), names of objects (*apple, beanbag, cup*, etc.), and terms for color, weight, and spatial relations. A speech recognizer, parser, and semantic composition system work together to convert commands into robot actions. Most aspects of the lexical structures are hand coded. Only the activation functions (pdf's) are trained from examples using standard statistical estimation techniques [3].

5.1. Speech Parsing and Semantic Composition

Front end speech recognition is performed using a HMM-based decoder [9]. The single best word sequence is passed to a chart parser which serves as the first step of a semantic composition procedure. The composition process is presented in detail in [10]. In summary, each lexical entry has a function interface that specifies how it performs semantic composition. Currently, the interface definition consists of the number and arrangement of arguments the entry is willing to accept. Semantic type mismatches are handled during composition rather than being enforced through the interface. Each entry can contain a *semantic composer* that encapsulates the actual function to combine this entry with other constituents during a parse.

[3] Our future plans are to develop structured learning algorithms to acquire both SNs and projection functions.

The system is able to resolve the referent of utterances with multiple modifiers. To achieve this, virtual objects consisting of one or more actual objects are internally generated during semantic composition. Consider the spoken command, "Pick up the large green cup to the left of the blue plate". To resolve the reference of *large green cup*, the innermost term, *cup* is first bound to objects in the robot's environment based on the visual shape models associated with the word. If multiple cups are found, they are grouped into a virtual object. This virtual object is then composed with the representation of *green*, which will threshold and sort the contents of the virtual object based on greenness, and pass along the new virtual object to *large*. The landmark phrase *blue plate* is processed in the same way, resulting in a second virtual object. The spatial phrase *to the left of* is used to find the best pair of objects, one drawn from each of the virtual objects. Finally, the best referent is passed as an argument to the *pickup* SN, which actually executes the action and picks up the target object.

The words *my* and *your* are given special treatment when adjacent to spatial terms, each triggering an appropriate shift of visual perspective within Ripley's mental model. Subsequent spatial terms are evaluated in the shifted frame of reference.

In situations where no referent matches one or more word meaning, (recall that a voting scheme is used to determine categorical boundaries), a null virtual object results. Ripley's default response in such cases is to look up at the person and report, "sorry, not found" using a speech synthesizer.

In situations where multiple referents satisfy a request, the robot uses a simple template to generate a clarifying question (e.g., "which red one?" in response to a request for a red object when multiple red objects are found). The subsequent description from the human partner (e.g., "The one on the right") is parsed and combined with the ambiguous request. If an ambiguity persists, the best matching object is selected (for a related approach that addresses dialog uncertainty in an decision-theoretic framework, see [11]).

A request for objects based on weight (e.g., *Hand me the heavy one*) can lead to particularly interesting behavior. If Ripley has not lifted an object, it has no way to know the object's weight. In order to interpret the meaning of the request, missing information about the environment must first be acquired. Weight terms are defined in terms of a projection of the *lift* SN. Thus, to acquire the weight of the object, Ripley executes this SN and makes the requisite measurements (accumulated joint forces) to determine the best referent for the request.

In summary, Ripley is able to respond to spoken requests using a limited vocabulary of sensorimotor grounded words. The interaction is fluid and responsive. When insufficient information is available to act, Ripley may either ask clarifying questions or actively acquire missing information in order to take appropriate actions.

6. RELATED WORK

On the surface, our work closely resembles the pioneering work of Winograd's SHRDLU system [12]. SHRDLU was a simulated robot that would accept natural language commands and translate them into actions in a simulated blocks world. The system demonstrated an impressive range of language understanding capabilities through rich integration of language processing with a world model. This work established the importance of domain knowledge in language processing, and introduced procedural semantic representations. A fundamental difference between Winograd's work (and the body of related work that arose from those early ideas) and ours is that SHRDLU existed in a constructed reality, represented purely through a symbolic representation of blocks and their properties. In contrast, our interest is in building language processing systems that are physically embodied, and sense the actual context within which the human partner is co-situated. By pushing semantics into the real world, many of the assumptions built into SHRDLU and the underlying semantics of the model, such as noise and ambiguity free world models, are invalidated. A second significant departure from Winograd's work, reflected in related threads of our work, is that we are developing grounded systems that learn (cf. [13, 4, 7]).

In the past decade, there have been several significant advances in linking natural language semantics to sensorimotor representations. Siskind [14] has suggested that the meaning of verbs can be represented in terms of force dynamics [15] grounded in video. In his approach, Allen relations [16] are used to capture temporal structure. Several researchers from the Neural Theory of Language group at Berkeley have proposed visually grounded representations of spatial relations [8], and sensorimotor inspired procedural representations of verbs [17, 18].

In our own previous work, we have constructed several systems that explore the nature of visually-grounded semantics. In [13], we used a mutual information based clustering technique to acquire a visually-grounded vocabulary from unannotated speech and video data, resulting in a plausible cognitive model of infant word learning [19]. We have also investigated visually-guided grammar acquisition [7] and speech understanding systems that connect referring expressions to objects in visual scenes [4, 10].

The approach presented in this paper makes new contributions along three dimensions. First, the use of sensorimotor networks provides a link from perceptual concepts to motor-grounded structures. Second, the introduction of a mental model enables shifts in visual perspective. In previous work, the visual point of view of the machine has always been fixed at a first person perspective. Third, these representations have been integrated and implemented on a real-time, interactive, robotic platform.

7. LOOKING AHEAD

We have presented an approach to representing word meanings that enables an interactive robot to respond to a range of spoken commands. Situational context including the presence and location of the human speaker, and the presence, properties, and configuration of objects in a shared space are used to interpret the meaning of spoken language. The deep structure underlying word meanings enables the robot to plan actions such as generating clarifying questions and triggering active perception to resolve ambiguities. We believe the underlying principles and methods can be transferred to a broad range of applications in which situational context is essential to understanding the intentions of the speaker.

There are many directions in which we plan to take this work forward. The vision system is overly simplistic and relies heavily on controlled backgrounds. We are in the process of designing a new vision system that is better suited to complex visual environments, and also better suited to capturing action-related features of environments. The speech recognizer in Ripley is capable of producing word lattices, but currently only the best path is used. For robust performance, we plan to parse multiple paths with semantic constraints based on the environment. To enlarge the range of language that Ripley can process, we will explore learning algorithms that are able to acquire grounded word meanings within our representational framework.

8. REFERENCES

[1] Jon Barwise and John Perry, *Situations and Attitudes*, MIT-Bradford, 1983.

[2] Herbert Clark, *Using Language*, Cambridge University Press, 1996.

[3] Kai-yuh Hsiao, Nikolaos Mavridis, and Deb Roy, "Coupling perception and simulation: Steps towards conversational robotics," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003.

[4] D. Roy, P. Gorniak, N. Mukherjee, and J. Juster, "A trainable spoken language understanding system for visual object selection," in *International Conference of Spoken Language Processing*, 2002.

[5] Intel, "Open source computer vision library (http://www.intel.com/research/mrl/research/opencv)," .

[6] Deb Roy, Bernt Schiele, and Alex Pentland, "Learning audio-visual associations from sensory input," in *Proceedings of the International Conference of Computer Vision Workshop on the Integration of Speech and Image Understanding*, Corfu, Greece, 1999.

[7] Deb Roy, "Learning visually-grounded words and syntax for a scene description task," *Computer Speech and Language*, vol. 16(3), 2002.

[8] Terry Regier, *The human semantic potential*, MIT Press, Cambridge, MA, 1996.

[9] Benjamin Yoder, "Spontaneous speech recognition using hidden markov models," M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2001.

[10] Peter Gorniak and Deb Roy, "Grounded semantic composition for visual scenes," submitted to JAIR.

[11] T. Paek and E. Horvitz, "Conversation as action under uncertainty," in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, 2000.

[12] T. Winograd, *A Process model of Language Understanding*, pp. 152–186, Freeman, 1973.

[13] Deb Roy, "Integration of speech and vision using mutual information," in *Proc. of ICASSP*, Istanbul, Turkey, 2000.

[14] Jeffrey Siskind, "Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic," *Journal of Artificial Intelligence Research*, vol. 15, pp. 31–90, 2001.

[15] L. Talmy, "Force dynamics in language and cognition," *Cognitive Science*, vol. 12, pp. 49–100, 1988.

[16] J. Allen, "Maintaining knowledge about temporal intervals," *Communications of the ACM*, vol. 26, pp. 832–843, 1983.

[17] D. Bailey, *When push comes to shove: A computational model of the role of motor control in the acquisition of action verbs*, Ph.D. thesis, Computer science division, EECS Department, University of California at Berkeley, 1997.

[18] Srinivas Narayanan, *KARMA: Knowledge-based active representations for metaphor and aspect*, Ph.D. thesis, University of California Berkeley, 1997.

[19] Deb Roy and Alex Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.