Towards a Layered Framework for Embodied Language Processing in Situated Human-Robot-Interaction

Matthias A. Priesters, Malte Schilling, Stefan Kopp
Bielefeld University, Faculty of Technology
Center of Excellence "Cognitive Interaction Technology" (CITEC)
Postfach 100 131, 33501 Bielefeld, Germany
{mpriesters,mschilli,skopp}@techfak.uni-bielefeld.de

Abstract

We propose a layered architecture for embodied language processing in a robot, in which the language layer is grounded in a sensorimotor layer via a schematic representation layer, which represents manual action or spatial relations in terms of embodied schemas. The schemas, on the one hand, abstract from the current sensory and motor states of the robot, and, on the other hand, enable mental simulation.

1 Introduction

The goal of the FAMULA project is to develop a bimanual robot torso able to familiarize itself with objects and their affordances.[1] This familiarization shall be scaffolded by situated dialogue grounded in unfolding manual action. Humans usually have no problem following dialogues full of underspecified references to objects or actions (e.g. containing utterances such as "no, the other one" or "yes... further... and now on top of it"), whereas for artificial agents this is no trivial task. We aim to explore the cognitive and linguistic abilities needed for the robot to engage in such situated dialogues before, during, and after action execution. The robot should be aware of its own knowledge gaps and attempt to reduce its uncertainty either by exploring the objects or by soliciting information from the tutor in situated dialogue. In such situated interaction, meaning unfolds dynamically and across language, bodily action and the interactants' environment (Goodwin, 2000), which exceeds current language-based human-robot interaction.

From the perspective of *embodied cognition*, higher-level representations are grounded in lower-level functions and are tightly interconnected (Feldman and Narayanan, 2004; Roy, 2005; Barsalou, 2008). As a consequence, cognition arises from the dynamic sensorimotor interaction with the environment. This leads to a central role of embodied representations grounded in lower-level experiences and sensorimotor behaviors. Our research focuses on modeling how embodied, situated meaning of communicative actions emerges in given interaction contexts.

2 Layered framework concept

Following the embodied cognition stance, the robot is firstly situated in a *sensorimotor* way, i.e inside its own physical form. On the one hand, its possible interactions are constrained by the shape of its hardware. On the other hand, the current states of the robot's 'body', i.e. whether it is currently engaged in an action or registers input through its sensors, influences language understanding and dialogue decision-making. Secondly, the robot is situated in its physical *environment*, i.e. in our context the scenery of objects present on the surface in front of it. These objects and their states and properties constrain the robot's options for actions and influence its needs for guidance and information. Thirdly, language use and meaning is situated in the *interaction* with the human tutor. This includes the dialogue history and common ground, the robot's interactional goals, as well as expectations and constraints for following dialogue acts or bodily actions, "interaction affordances" (Raczaszek-Leonardi et al., 2013), created by the situation and previous dialogue acts.

We devised a three-layer framework (Fig. 1) for embodied language processing. It provides higher-level representations which are embodied in the sense that they are grounded in the sensorimotor layer.

The lowest, **sensorimotor**, layer of our framework comprises the actual control primitives and

[1] http://cit-ec.de/en/content/deep-familiarization-and-learning

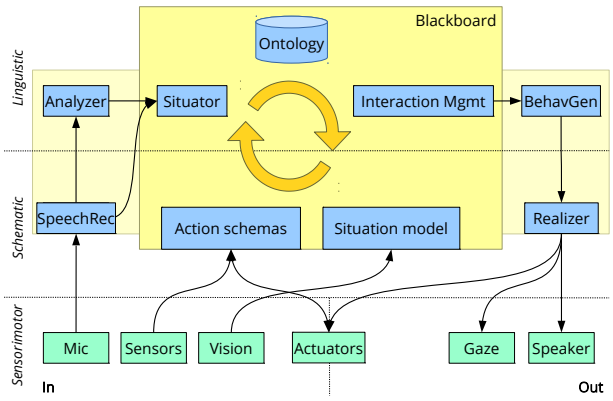

Figure 1: Proposed three-layer system architecture

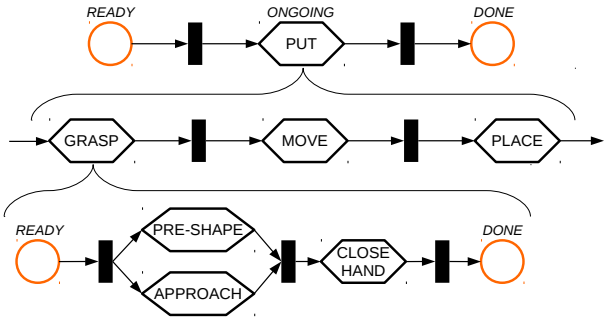

Figure 2: Hierarchical Action Schema representation for PUT (simplified)

sensor readings from the robot hardware. The hardware consists of two robotic arms mounted in a torso-like configuration onto a table. On the arms, five-digit hands are attached, which enable precise exploration and manipulation and are equipped with touch-sensitive fingertips.

To mediate between sensorimotor and linguistic layers and to realize situatedness, we introduce an intermediate **schematic** layer. Actions are represented as executable *action schemas* (Schilling and Narayanan, 2013), a Petri Net-based formalism (Fig. 2), which on the one hand represents sensorimotor states of action execution and on the other hand offers internal simulation capabilities, enabling the system to generate predictions and to represent the unfolding of actions over time. Lower-level sensory input gets accessible as part of the states of the Petri Nets. The scenery the robot is situated in is represented in a *situation model*, which keeps track of the present objects, of their ontological categories and properties.

The highest, **linguistic**, layer deals with speech input and generating answers, as well as with decision-making regarding communicative interaction. The language *analyzer* syntactically and semantically parses the input utterance using an Embodied Construction Grammar parser (Bryant, 2008). The purpose of the *situator* is situated language understanding, i.e. identifying dialogue acts and resolving references to objects or actions based on situational and ontological knowledge and on bodily and interaction states. The *interaction manager* is the decision-making component, which maintains knowledge about the level of certainty or uncertainty of the system in the current situation and decides on appropriate actions.

Implementation of the framework is work in progress and currently focuses on including infomation about the execution status of actions into language processing. For example, when resolving an ambiguous, underspecified reference ("no, the other one"), the system takes into account which object is currently in focus (e.g. the moved object vs. the target location of a PUT action) depending on the state of the ongoing action (GRASP vs. MOVE/PLACE, Fig. 2).

Acknowledgments

This work is supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG).

References

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59(1):617–645.

John Bryant. 2008. *Best-fit constructional analysis*. Ph.D. thesis, University of California, Berkeley.

Jerome Feldman and Srinivas Narayanan. 2004. Embodied meaning in a neural theory of language. *Brain and Language*, 89(2):385–392.

Charles Goodwin. 2000. Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10):1489–1522.

Joanna Raczaszek-Leonardi, Iris Nomikou, and Katharina J. Rohlfing. 2013. Young children's dialogical actions: The beginnings of purposeful intersubjectivity. *IEEE Trans Auton Ment Dev*, 5(3):210–221.

Deb Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1–2):170–205.

Malte Schilling and Srini Narayanan. 2013. Communicating with executable action representations. In *AAAI 2013 Spring Symposium Series*, Stanford.