Multimodal Communication with Robots and Computational Agents

Instructors:
James Pustejovsky (jamesp@brandeis.edu)
Nikhil Krishnaswamy (nkrishna@brandeis.edu)

This tutorial introduces the requirements and challenges involved in developing a multimodal semantics for human-robot interactions. Unlike human-computer (HC) interactive agents (e.g., chatbots or personal digital assistants), human-robot interaction (HRI) inherently requires multiple modalities. Robotic agents are embodied and situated which affords robots the ability to affect the real world, but also requires them to have accurate and robust interpretive capabilities for multiple input modalities, which must run in real time. In addition, a robot must be able to communicate with its human interlocutors using all communicative modalities humans may use, including natural language, body language, gesture, demonstrated action, etc.

The very near future of communicative interactions with computational agents and robots will be not be limited to speech. It will be inherently multimodal, drawing on spoken and typed language, haptic input from pads, keys, and sensors, head and hand gestures from image and video RGBD captures, and contextualized and situational awareness from registering local actions in the environments. Researchers in each of these modalities have traditionally needed their own separate control language or manager for navigating a conversation or dialogue. When communications become multimodal in nature, however, such dialogue managers are no longer rich enough to model the context and interaction. Each modality in operation provides an orthogonal angle through which to probe the computational model of the other modalities, including the behaviors and communicative capabilities afforded by each. Through studying and modeling the semantics of the communication, conversational capabilities with computers will be enormously enhanced by providing representations for the common ground between human and computer.

This tutorial focuses on the semantics of actions and object affordances and the impact such knowledge has on reasoning in HRI. While the dynamic semantics of epistemic updating in discourse has been extensively modeled, there has been less development of integrated models of the dynamics of actions and affordances in cooperative or goal-directed discourse. We present a dynamic semantics of the language, VoxML, to model both HC and HR interactions by creating multimodal simulations of both the communicative content and the agents' common ground. A multimodal simulation is an embodied 3D virtual realization of both the situational environment and the co-situated agents, as well as the most salient content denoted by communicative acts in a discourse. VoxML provides a representation for the situated grounding of expressions between individuals involved in a communicative exchange. It does this by encoding objects with rich semantic typing and action affordances, and actions themselves as multimodal programs, enabling contextually salient inferences and decisions in the environment. These affordances provide the semantic scaffold on which to build abstract control and dialogue management strategies for multiple modalities.

We will demonstrate strategies and provide examples for composing multimodal information in real time using a *Multimodal Semantic Grammar.* Multimodal Semantic Grammar (MSG) is a continuation-passing style grammar for multimodal communicative sequencing, integrating language (speech/text), gesture, gaze, and action. It is designed as a language for multimodal interpretation and control for human-computer and human-robot interaction, integrating three separate control structures for communication: a discourse sequence grammar; gesture grammar; and an action grammar.

Additional course material, including but not limited to slides, videos, and readings, will be posted at http://www.voxicon.net in advance of LREC 2020.

Topics of Interest:
Human Computer Interaction, multimodal communication, semantic grounding, situated and embodied communication, Human Robot Interaction, Language, perception, theory of action.