Multimodal Simulations and Situational Grounding

James Pustejovsky

Brandeis University
jamesp@brandeis.edu

There has been growing interest in providing "semantic grounding" for linguistic expressions in natural language text. Typically, this involves anchoring a referring expression or verbal predicate to the appropriate segments in an image that denote an object or an action, respectively. While useful, such cross-modal linking is far from the situated grounding required to understand utterances, deixis, and actions, in the context of multimodal communication. Two humans engaged in dialogue share a common ground for their beliefs, perception, and situatedness. This is not the case, however, when communicating with computers or robots.

I argue that situated grounding is a requirement for more natural communication between humans and computers (or robots), and is therefore critical to the success of both natural language understanding and AI. In order to achieve this goal, human-computer/robot interactions will require at least the following capabilities and competencies: robust recognition and generation of multiple modalities (language, gesture, vision, action); understanding of contextual grounding and co-situatedness; and appreciation of the consequences of behavior and actions.

In this paper, I describe an approach to modeling human-computer interactions based on creating multimodal simulations. 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. It is built on the modeling language VoxML, which encodes objects with rich semantic typing and action affordances, and actions themselves as multimodal programs, enabling contextually salient inferences and decisions in the environment.

Since a simulation reveals the elements of the common ground in discourse between speakers, it offers a rich platform for studying the generation and interpretation of expressions, as conveyed through multiple modalities, including: language, gesture, and the visualization of objects moving and agents acting in their environment. I will present current research demonstrating multimodal human-computer interactions in a collaborative task, and discuss ongoing lines of research in using a multimodal simulation context for introducing novel concepts and situations to a computational agent.