Embodiment and Affordances

Following Generative Lexicon (GL) (Pustejovsky, 1995), lexical entries in the object language are given a feature structure consisting of a word’s basic type, its parameter listing, its event typing, and its quality structure. The semantics of an object will consist of the following:

- Atomic Structure (FORMAL): objects expressed as basic nominal types.
- Subatomic Structure (CONSTR): merontological structure of objects.
- Event Structure (TELIC and AGENTIVE): origin and functions associated with an object.
- Macro Object Structure (FORMAL): objects fit together in space and through coordinated activities.
- Objects can be partially contextualized through their qualia structure: a fixed item has a TELIC value of evil, an instruction for writing, a TELIC of write, a cup a TELIC of hold, so forth. For example, the lexical semantics for the noun chair carries a TELIC value of sit in: chair (SUBJ) chair (OBJ) chair (location) sit (affordance: chair).

The habitat for an object is built by first placing it within an embedding space and then contextualizing it. For example, in order to use a table, the top has to be oriented upward, the surface must be accessible, and so on. A chair must also be oriented up, the seat must be accessible, it must be able to support the user, etc.

Examples of Voxemes

VoxSim: A Visual Platform for Modeling Motion Language

VoxSim (Krishnaswamy and Pustejovsky 2016), is a semantically-inferred visual event simulator built on top of the Unity game engine (Guldstone 2009). VoxSim procedurally composes the properties of voxemes in parallel with the lexemes to which they are linked, input is a simple natural language sentence, which is part-of-speech tagged, dependency-parsed, and transformed into a simple predicate-logic format.

A VoxSim entity’s interpretation at runtime depends on the other entities. It is composed with. In order to contain another object, a cup must be currently situated within a habitat which allows objects to be placed partially or completely made it (represented by partial overlap and tangential or non-tangential proper part relations—PO, TPP, or NTPP according to the Region Connection Calculus (Randell, Cui, and Drinkwater 1992)).

In VoxSim, (CLIPS) is encoded as a concave object with rotational symmetry around the Y axis and reflective symmetry across the X and Z planes, meaning that it opens along the Y axis. Its HABITAT further situates the opening along its positive Y axis, meaning that if the cup’s opening along its Y is currently unobstructed, it affords containment. The contained object must then be transformed to satisfy equivalent constraints, such as turning a spoon to sit in a cup (as shown below). If no such transformation is possible, VoxSim returns an error message.

VoXML: A Language for Concept Visualization

VoXML can also be used to help detect and recognize events and actions in video. In order to do so, our lab is creating a dataset of videos annotated with event subevent relations using ECAF (Do et al., 2016), an internally-developed video annotation tool. This allows us to annotate videos of labeled events with object participants and subevents, and to induce what the common subevent structures are for the labeled superevent.

In our dataset, both human bodies (hgs) and these objects can be tracked and annotated as participants in a recorded motion event. This labeled data can then be used to build a corpus of multimodal semantic simulations of these events that can model object-object, object-agent, and agent-agent interactions through the event duration.

Gesture-Speech Act Mapping

We are also interested in learning the mapping between communicative gestures and their speech acts. In a situation when a human agent has to give directions to a robot in order to achieve a specific task (such as building a structure), using gestures in a multimodal (coverb) manner can be more economical than language alone. However, it turns out that a single human gesture can be interpreted as several different speech acts, depending on local context. Using the same annotation methodology, we can map between (current configuration, verbal command, coverbal gesture) and a speech act of type (target configuration). Given the current configuration in its simulated environment, by receiving a linguistic expression, a coverbal gesture, or the simultaneous articulation of both, the robot can learn to generate the appropriate target configuration.