Embodied Human-Computer Interactions through Situated Grounding

James Pustejovsky and Nikhil Krishnaswamy

IVA '20: ACM International Conference
on Intelligent Virtual Agents
October 19–23, 2020
Glasgow, UK

Situated Semantic Grounding and Embodiment

- Task-oriented dialogues are embodied interactions between agents, where language, gesture, gaze, and actions are situated within a common ground shared by all agents in the communication.
- Situated semantic grounding assumes shared perception of agents with co-attention over objects in a situated context, with co-intention towards a common goal.
- VoxWorld : a multimodal simulation framework for modeling Embodied Human-Computer Interactions and communication between agents engaged in a shared goal or task.
- Embodied HCI and robot control in action.

Situated Meaning

Mother and son interacting in a shared task of icing cupcakes

Situated Meaning in a Joint Activity

- Son: *Put it there (gesturing with co-attention)?*
- Mother: *Yes, go down for about two inches.*
- Mother: *OK, stop there. (co-attentional gaze)*
- Son: *Okay. (stops action)*
- Mother: *Now, start this one (pointing to another cupcake).*

Situated Meaning

Elements from the Common Ground

Agents	mother, son
Shared goals Beliefs, desires, intentions	baking, icing Mother knows how to ice, bake, etc. Mother is teaching son
Objects	Mother, son, cupcakes, plate, knives, pastry bag, icing, gloves
Shared perception	the objects on the table
Shared Space	kitchen

Embodied Human-Computer Interaction

- Elements of Situated Meaning
  - Identifying the *actions and consequences* associated with objects in the environment.
  - Encoding a multimodal expression contextualized to the *dynamics of the discourse*
  - *Situated grounding*: Capturing how multimodal expressions are anchored, contextualized, and situated in context
- Modalities Deployed
  - gesture recognition and generation
  - language recognition and generation
  - affect, facial recognition, and gaze
  - action generation

IVA in Embodied Environment

An encounter between two “people” with multimodal dialogue: language, gesture, gaze, action.

Figure: IVA Diana engaging in an embodied HCI with a human user.

Link

Affordance and Goal Recognition

1. Perceived purpose is an integral component of how we interpret situations and reason about utterances in communicative contexts.

- Events are purposeful and directed;
- Places are functional;
- Objects are usable and manipulable.

2. Affordances are latent action structures of how an agent interacts with objects in the environment, in different modalities:

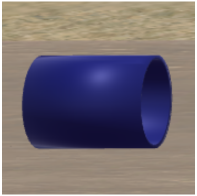
- language, gesture, vision, action;

3. Qualia Structure provides a link to such latent actions structures associated with objects in utterances and the context.

Focus on Objects

- Context of objects is described by their properties.
- Object properties cannot be decoupled from the events they facilitate.
  - *Affordances* (Gibson, 1979)
  - *Qualia* (Pustejovsky, 1995)

"He **slid** the cup across the table. Liquid spilled out."
"He **rolled** the cup across the table. Liquid spilled out."



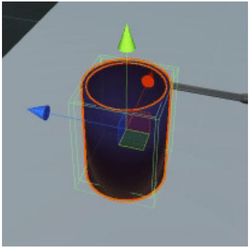
Visual Object Concept Modeling Language (VoxML)

Pustejovsky and Krishnaswamy (2016)

- Encodes afforded behaviors for each object
  - Gibsonian: afforded by object structure (Gibson,1977,1979)
    - grasp, move, lift, etc.
  - Telic: goal-directed, purpose-driven (Pustejovsky, 1995, 2013)
    - drink from, read, etc.
- Voxeme
  - Object Geometry: Formal object characteristics in R3 space
  - Habitat: Conditioning environment affecting object **affordances** (behaviors attached due to object structure or purpose);
  - Affordance Structure:
    - What can one do to it
    - What can one do with it
    - What does it enable

VoxML - cup

$$
\left[\begin{array}{l}
\textbf{cup} \\
\text{LEX} = \left[\begin{array}{l} \text{PRED} = \textbf{cup} \\ \text{TYPE} = \textbf{physobj, artifact} \end{array}\right] \\
\text{TYPE} = \left[\begin{array}{l} \text{HEAD} = \textbf{cylindroid[1]} \\ \text{COMPONENTS} = \textbf{surface, interior} \\ \text{CONCAVITY} = \textbf{concave} \\ \text{ROTATSYM} = \{Y\} \\ \text{REFLECTSYM} = \{XY, YZ\} \end{array}\right] \\
\text{HABITAT} = \left[\begin{array}{l} \text{INTR} = {\scriptstyle [2]} \left[\begin{array}{l} \text{CONSTR} = \{Y > X, Y > Z\} \\ \text{UP} = align(Y, \mathcal{E}_Y) \\ \text{TOP} = top(+Y) \end{array}\right] \\ \text{EXTR} = {\scriptstyle [3]} \left[\text{UP} = align(Y, \mathcal{E}_{\perp Y}) \right] \end{array}\right] \\
\text{AFFORD\_STR} = \left[\begin{array}{l} A_1 = H_{[2]} \rightarrow [put(x, on([1]))]support([1], x) \\ A_2 = H_{[2]} \rightarrow [put(x, in([1]))]contain([1], x) \\ A_3 = H_{[2]} \rightarrow [grasp(x, [1])] \\ A_4 = H_{[3]} \rightarrow [roll(x, [1])] \end{array}\right] \\
\text{EMBODIMENT} = \left[\begin{array}{l} \text{SCALE} = \textbf{<agent} \\ \text{MOVABLE} = \textbf{true} \end{array}\right]
\end{array}\right]
$$



VoxML

VoxML for Actions and Relations

$$
\left[\begin{array}{l}
\textbf{put} \\
\text{LEX} = \left[\begin{array}{lcl} \text{PRED} & = & \textbf{put} \\ \text{TYPE} & = & \textbf{transition\_event} \end{array}\right] \\
\text{TYPE} = \left[\begin{array}{lcl}
\text{HEAD} & = & \textbf{transition} \\
\text{ARGS} & = & \left[\begin{array}{lcl} \text{A}_1 & = & \textbf{x:agent} \\ \text{A}_2 & = & \textbf{y:physobj} \\ \text{A}_3 & = & \textbf{z:location} \end{array}\right] \\
\text{BODY} & = & \left[\begin{array}{lcl} \text{E}_1 & = & grasp(x,y) \\ \text{E}_2 & = & [while(hold(x,y), move(x,y))] \\ \text{E}_3 & = & [at(y,z) \rightarrow ungrasp(x,y)] \end{array}\right]
\end{array}\right]
\end{array}\right]
$$

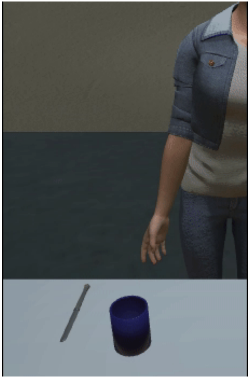
$$
\left[\begin{array}{l}
\textbf{on} \\
\text{LEX} = \left[\begin{array}{lcl} \text{PRED} & = & \textbf{on} \end{array}\right] \\
\text{TYPE} = \left[\begin{array}{lcl}
\text{CLASS} & = & \textbf{config} \\
\text{VALUE} & = & \textbf{EC} \\
\text{ARGS} & = & \left[\begin{array}{lcl} \text{A}_1 & = & \textbf{x:3D} \\ \text{A}_2 & = & \textbf{y:3D} \end{array}\right] \\
\text{CONSTR} & = & \textbf{y}{\rightarrow}\text{HABITAT}{\rightarrow}\text{INTR}[align]
\end{array}\right]
\end{array}\right]
$$

VoxML - grasp

$$\left[\begin{array}{l}\textbf{grasp}\\ \text{LEX} = \left[\begin{array}{l}\text{PRED} = \textbf{grasp}\\ \text{TYPE} = \textbf{transition\_event}\end{array}\right]\\ \text{TYPE} = \left[\begin{array}{l}\text{HEAD} = \textbf{transition}\\ \text{ARGS} = \left[\begin{array}{l}\text{A}_1 = \textbf{x:agent}\\ \text{A}_2 = \textbf{y:physobj}\end{array}\right]\\ \text{BODY} = \left[\,\text{E}_1 = grasp(x, y)\,\right]\end{array}\right]\end{array}\right]$$

VoxML - grasp cup

- Continuation-passing style semantics for composition
- Used within conventional sentence structures and between sentences in discourse in MSG

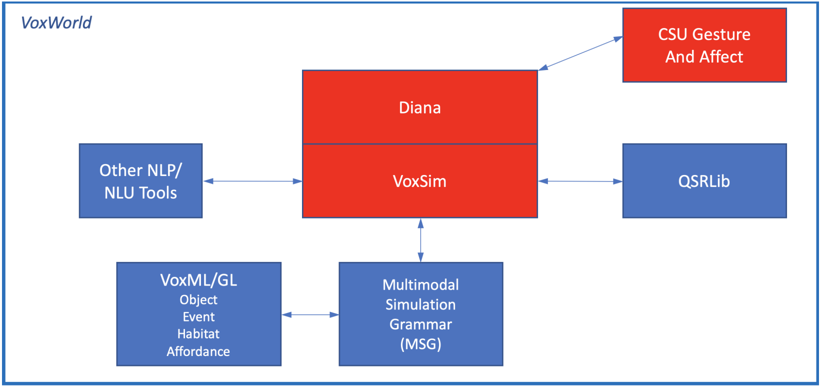


Multimodal Simulations

- Human understanding depends on a wealth of common-sense knowledge; humans perform much reasoning qualitatively.
- To simulate events, every parameter must have a value
  - "Roll the ball." How fast? In which direction?
  - "Roll the block." Can this be done?
  - "Roll the cup." Only possible in a certain orientation.
- VoxML: Formal semantic encoding of properties of objects, events, attributes, relations, functions.
- VoxSim: What can situated grounding do? (Krishnaswamy, 2017)
  - Exploit numerical information demanded by 3D visualization;
  - Perform qualitative reasoning about objects and events;
  - Capture semantic context often overlooked by unimodal language processing.

VoxWorld: A Platform for Multimodal Simulations

Interfacing Diana to CSU Gesture and Affect Systems



Dynamic Discourse Interpretation

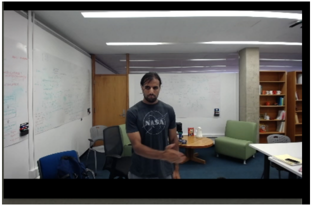
- Common Ground Structure
    - Co-belief
    - Co-perception
    - Co-situatedness
- Multimodal communication act:
    - language
    - gesture
    - action
- Dynamic tracking and updating of dialogue with:
    - Discourse Sequence Grammar
    - Gesture Grammar
    - Action Grammar

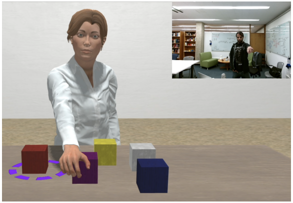
Co-belief and Co-perception in the Common Ground

- *Public announcement logic* (PAL)
  - $[\alpha]\varphi$ denotes that an agent "α *knows* φ".
  - Public Announcement: $[!\varphi_1]\varphi_2$
  - Any proposition, φ, in the common knowledge held by two agents, α and β, is computed as: $[(\alpha \cup \beta)^*]\varphi$.
- *Public perception logic* (PPL)
  - $[\alpha]_\sigma\varphi$ denotes that agent "α perceives that φ".
  - $[\alpha]_\sigma\hat{x}$ denotes that agent "α perceives that there is an x."
  - Public Display: $[!\varphi_1]_\sigma\varphi_2$
  - The co-perception by two agents, α and β includes φ : $[(\alpha \cup \beta)^*]_\sigma\varphi$

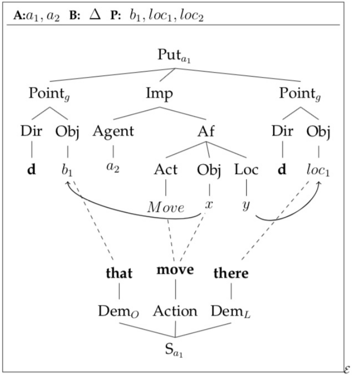
Situated Meaning

Gesture and co-gestural speech imperative





a_1: "**That** object b_1 **move** b_1 to **there**, location loc_1."



$$\lambda k_s' \otimes k_g'.\overline{(\langle \textbf{that}, Point_1\rangle\langle \textbf{move}, Move\rangle)}(\lambda r_s \otimes r_g.\overline{\langle \textbf{that}, Point_2\rangle}$$
$$(\lambda k_s \otimes k_g.k_s' \otimes k_g'(k_s \otimes k_g r_s \otimes r_g)))$$

Transfer Learning of Object Affordances

- Gibsonian/Telic affordances are associated with abstract properties:
  - spheres roll, sphere-like entities probably do too;
  - small cups are graspable, small cylindroid-shaped objects probably are too.
- Similar objects have similar habitats/affordances:
- This informs the way you can talk about items in context:
  - Q: "What am I pointing at?"
  - A: "I don't know, but it looks like {a ball/a container/etc.}

Transfer Learning of Object Affordances

Exploits the linkages between affordances and objects in VoxML

- Train over a sample of 17 different objects: blocks, KitchenWorld objects (apple, grape, banana, book, etc.)
- Trained 200 dimensional affordance and habitat embeddings using a Skip-Gram model, for 50,000 epochs with a window size of 3:
  - These embeddings serve as the inputs to the object prediction architectures
- Using the affordance embeddings in vector space, predict which object they belong to: using a 7-layer MLP; a 4-layer CNN with 1D convolutions

Transfer Learning of Object Affordances

- The architectures:

MLP	**CNN**
Input	Input
Dense (32 $\times$ *tanh*)	**Conv1D (64** $\times$ *ReLU*)
20% Dropout	*ReLU*
Dense (196 $\times$ *ReLU*)	20% Dropout
20% Dropout	**Conv1D (250** $\times$ *ReLU*)
Dense (92 $\times$ *tanh*)	Global Max Pooling 1D
20% Dropout	20% Dropout
Dense (196 $\times$ *tanh*)	**Dense (196)**
Dense (92 $\times$ *ReLU*)	20% Dropout
Dense (32 $\times$ *tanh*)	*ReLU*
Output ($softmax$)	**Output** ($softmax$)
70,913 params	100,923 params

- Ground truth clusters generated by k-means clustering over human-annotated object similarity. Sample aggregate results:

Model	**% predictions in correct cluster**	**% predictions always in correct cluster**
MLP (Habitats)	78.82	27.06
MLP (Affordances)	**84.71**	38.82
CNN (Habitats)	78.82	27.06
CNN (Affordances)	81.18	**40.00**

- Object specific results (input: vectorized affordances for plate)

MLP (Habitats)	**MLP (Affordances)**	**CNN (Habitats)**	**CNN (Affordances)**
book, cup, bowl, bottle	cup, bottle, apple	book	cup, bottle

Transfer Learning of Object Affordances

Play!

Refactoring VoxWorld for Robot Navigation and Control

Kirby's World

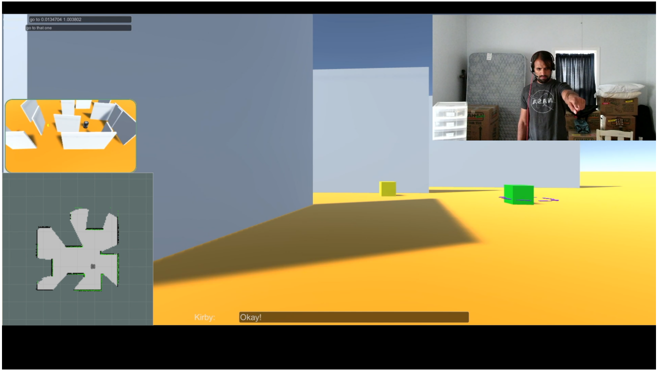
- Gesture and language communication with a Turtlebot-3:

- Fiducials represent registered proxies for object sorts in the environment:

Figure 2: Two fiducials.

Refactoring VoxWorld for Robot Navigation and Control

Kirby's World



▸ Link

Conclusion - Embodied HCI

- VoxWorld facilitates experimentation with IVAs in embodied HCI contexts, using multiple modalities in diverse settings.
- An embodied HCI, such as that enabled by the simulation environment VoxWorld, provides a venue for the human and computer or robot to share an epistemic space,
- Any communicative modality that can be expressed within that space (e.g., linguistic, visual, gestural) enriches the ways in which a human and a computer or robot can communicate regarding objects, actions, and situation-based tasks.

Thank You ☺

- Brandeis LLC lab members: Nikhil Krishnaswamy, Kyeongmin Rim, Mark Hutchens, Ken Lai, Katherine Krajovic, Daeja Showers, Eli Goldner, Kelley Lynch
- CSU Vision lab members: Ross Beveridge, Bruce Draper, Rahul Bangar, David White, Pradyumna Narayana, Dhruva Patil
- University of Florida lab members: Jaime Ruiz, Isaac Wang
- Funded by a grant from DARPA within the CwC Program