The VoxWorld Platform for Multimodal Embodied Agents

Nikhil Krishnaswamy, William Pickard, Brittany Cates, Nathaniel Blanchard, and James Pustejovsky • nkrishna@colostate.edu

VoxWorld

A platform for multimodal agent behaviors, presented as a resource to the AI/NLP community

What Makes an Agent?

- Perceives through sensors and acts through actuators
- Epistemic point of view from which it observes the world
- Virtual world is mode of presentation, allows observer to see what agent does
- Embodied agents add new dimensions to human/agent interactions
- Must recognize and interpret inputs in multiple modalities (e.g., gesture, speech, gaze action)
- Solving these problems has driven development of VoxWorld

Theory

- **VoxML** modeling language and **VoxSim** event simulator
- Events composed of subevent semantics that decompose into minimal primitive set
- Objects provide minimal encoding of properties, e.g., **habitats** and **affordances**
- Relations sample from distributions under constraints
- Event, relations, and objects composed at runtime

$$\left[\begin{array}{l}\textbf{put}\\ \text{BODY} = \left[\begin{array}{l}\text{E}_1 = grasp(x,y)\\ \text{E}_2 = while(hold(x,y) \wedge \neg at(y,z)) \\ \qquad \rightarrow move(x,y,z,\mathbf{P}(),[loc(y),z,y])\\ \text{E}_3 = if(at(y,z) \rightarrow ungrasp(x,y)\end{array}\right]\end{array}\right]$$

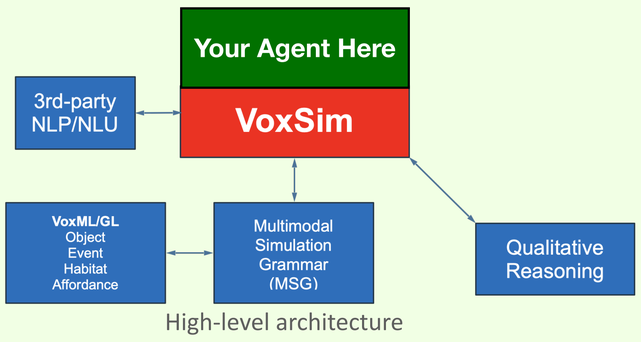
$$\left[\begin{array}{l}\textbf{in\_front}\\ \text{TYPE} = \left[\begin{array}{l}\text{CLASS} = \textbf{config}\\ \text{VALUE} = RCC8.EC\\ \text{ARGS} = \left[\begin{array}{l}\text{A}_1 = \textbf{x:physobj}\\ \text{A}_2 = \textbf{y:physobj}\end{array}\right]\\ \text{CONSTR} = Z(x) > Z(y)\end{array}\right]\end{array}\right]$$

$$\left[\begin{array}{l}\textbf{block}\\ \text{AFFORD\_STR} = \left[\begin{array}{l}\text{A}_1 = H_{[2]} \rightarrow [put(x,y,on([1]))]\\ \qquad support([1],y)\\ \text{A}_2 = H_{[2]} \rightarrow [grasp(x,[1])]\\ \qquad hold(x,[1])\\ \text{A}_3 = H_{[2]} \rightarrow [lift(x,[1])]\\ \qquad hold(x,[1])\\ \text{A}_4 = H_{[2]} \rightarrow [ungrasp(x,[1])]\\ \qquad release(x,[1])\end{array}\right]\end{array}\right]$$

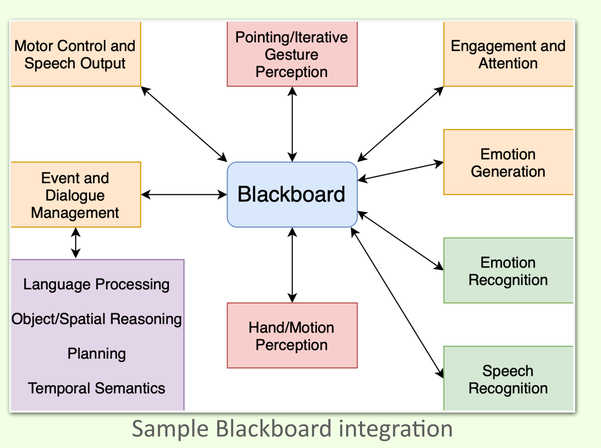
Sample voxemes

Implementation

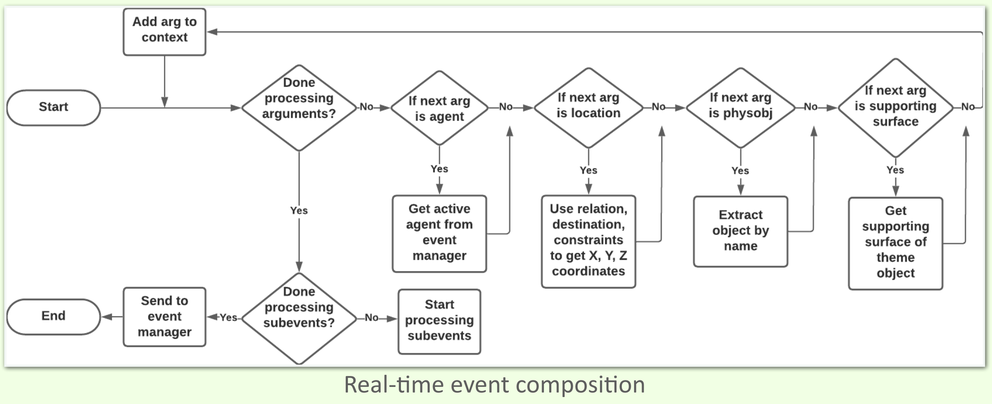
- Built on Unity game engine
- Accommodates qualitative calculi, machine learning inputs
- Interaction management via blackboard and pushdown automata
- Integrated with functional programming semantics
- Support arbitrary inputs and web deployment



High-level architecture



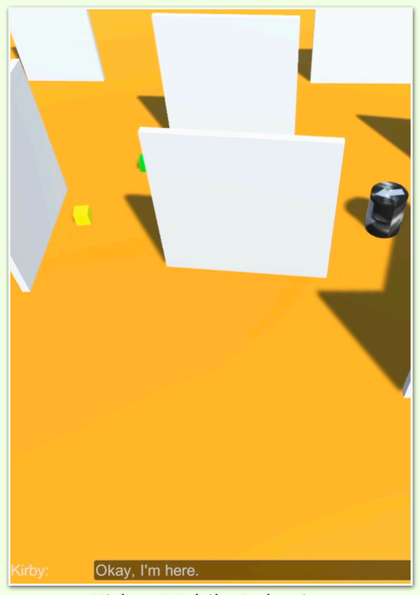
Sample Blackboard integration



Real-time event composition

Agent Implementations

Diana: Multimodal Interaction

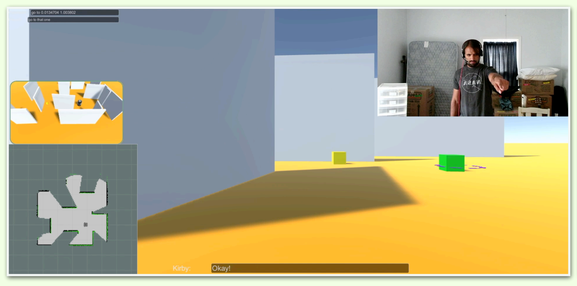


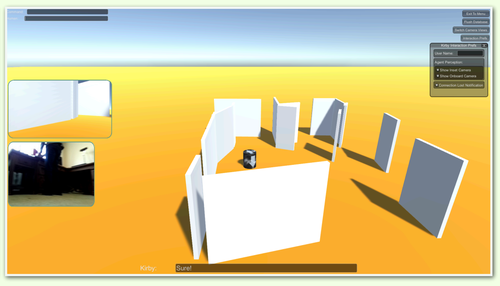
Kirby: Mobile Robotics

BabyBAW: Exploration with RL

Diana: interactive multimodal agent for peer-to-peer human-computer communication

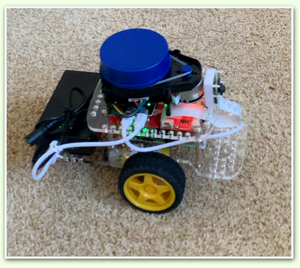
- Diana interprets asynchronous speech and gesture
- Received mean 74.3 System Usability Score in User Study





Kirby: Navigating robot

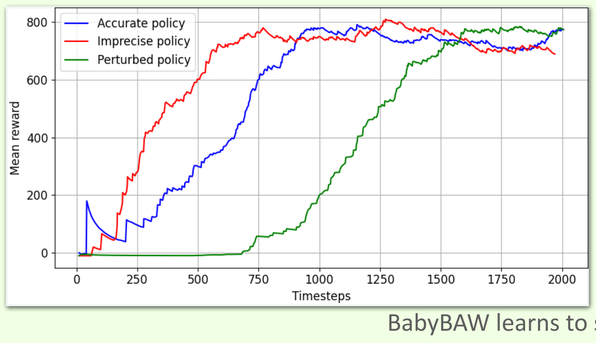
- Same multimodal interface as Diana
- Integrates Robot Operating System, LIDAR, live camera feed
- Fiducial and object detection

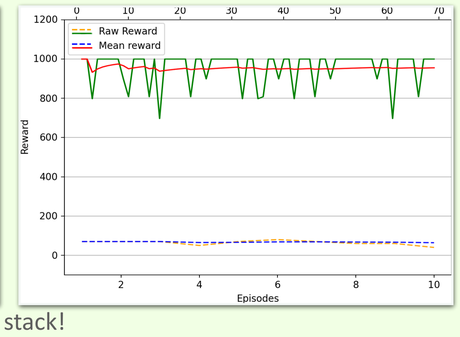


GoPiGo3 with LIDAR

BabyBAW: Environmental exploration with RL

- Self-guided learner intended to approximate aspects of infant/toddler learning
- Integrates Unity ML-Agents, OpenAI Gym
- Ground actions and objects in world to learned labels





BabyBAW learns to stack!

Resources

- Asset package and bleeding-edge source code: https://github.com/VoxML/VoxSim
- Sample project: https://github.com/VoxML/VoxWorld-QS
- Documentation: https://www.voxicon.net/api/

This work was supported in part by the U.S. Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under contract W911NF-15-C-0238 at Brandeis University and contract W911NF-15-1-0459 at Colorado State University, and by the National Science Foundation (NSF) under grant numbers CNS 2033932 and DRL 2019805 to Brandeis University and DRL 1559731 to Colorado State University.