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

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

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/