Colorado State University

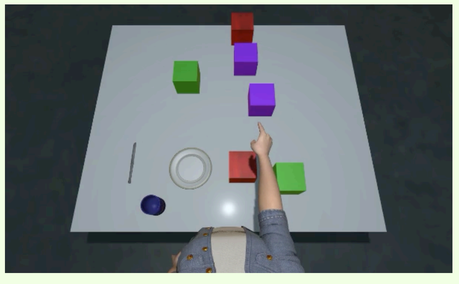
Embodied Multimodal Agents to Bridge the Understanding Gap

Nikhil Krishnaswamy and **Nada Alalyani**

nkrishna@colostate.edu • n.alalyani@colostate.edu

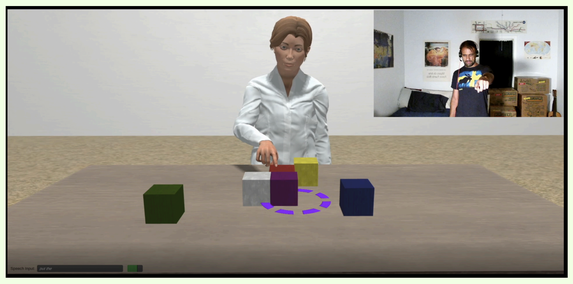
Introduction

- Embodied multimodal agents (avatars) model encounters between two "people," with environmental awareness
- Provide additional structure that can move NLP systems closer to genuine understanding of grounded language
- Where large language models and computer vision systems are difficult to probe, embodied agents have multiple avenues to demonstrate their understanding
- If one modality is insufficiently communicative, then another may supplement it
- "Understanding"~ retrieval of communicative intent from an utterance (Bender and Koller, 2020)
- We present ongoing experiments in multimodal agents exhibiting environmentally-grounded understanding

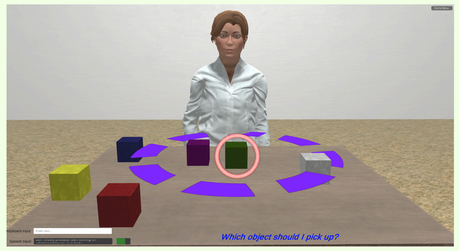


Diana System

- A co-perceptive, co-attentive agent
- A communicative act C_a = <***S***peech, ***G***esture, ***F***acial expression, ga***Z***e, and ***A***ction>
  - e.g., C_a = <***S*** = "left", ***G*** = [$Point_g$ ^ Dir = RIGHT]> - say "left" and point to the right; this signals *a*'s frame of reference
- Diana's semantic knowledge of objects and actions is based on VoxML (Pustejovsky and Krishnaswamy, 2016)
- Interprets language and gestures to collaborate on object movement tasks with humans
- Demonstrates understanding: if the human refers to "the purple block," Diana directs her attention there
- Diana's capabilities are not fully symmetric: the human may talk a lot, but Diana doesn't say much
- To increase "deep understanding" for both Diana and the human, we are conducting experiments on multimodal referring expressions



Ongoing Experiments



- Building a web-deployed version of Diana to study how people mix modalities in REs
- Purposely coarse mouse deixis + automated speech recognition
- Assessed quality of Google ASR with 20+ college students reading 5 pre-defined scripts
  - Assessed open vocabulary and syntactically-adaptive domain vocabulary recognition

	Open Vocabulary	Restricted Vocabulary
Accuracy	81.882%	84.345%
WER	18.002%	15.519%
Std. Dev. (WER)	0.20332	0.19087

- Log target object, coordinates, distance to agent, relations in scene, modality used, attributes of objects, relations between objects, previously-referenced objects

HUMAN: Take that purple block. [*points to two purple blocks near each other and far from Diana*]
DIANA: This one? [*points to a purple block*]
HUMAN: No. [*"thumbs down" gesture*]
DIANA: How about this one? [*points to the other purple block*]
HUMAN: Yes. Put it on the green block you just moved. [*points to a red block that* DIANA *recently put down*]
DIANA: Do you mean the red block I just put down?
HUMAN: [*"thumbs up" gesture*]

- Deploy on Prolific in 3 months: 250 workers, 10 scenes, 10 distinct target objects

Proposed Models

Outputs

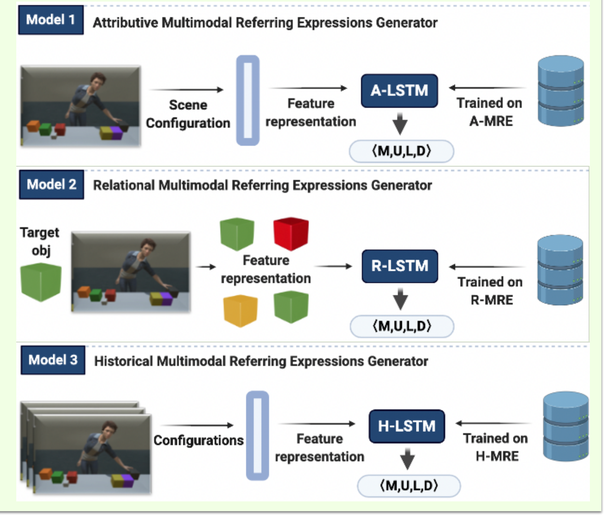
< **M**odality, M ∈ [Gesture, Language, Ensemble]
Utterance, U: decoded sentence embedding
Location, L: location gesture grounds to
Demonstratives > D ∈ [this, that]

Based on LSTMs

A-LSTM: takes target object as query, outputs descriptor tuple where M = L ∨ E and attribute ∈ U

R-LSTM: takes pairwise relations between target object and others, outputs relational descriptors satisfied by target

H-LSTM: takes past states, output actions previously taken with target



Selected Related Work

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proc. of ACL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Casey Kennington, Spyridon Kousidis, and David Schlangen. 2013. Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. *Proceedings of SIGdial 2013*.

Nikhil Krishnaswamy and James Pustejovsky. 2019. Generating a novel dataset of multimodal referring expressions. In *Proceedings of the 13th International Conference on Computational Semantics- Short Papers*, pages 44–51.

David G McNeely-White, Francisco R Ortega, J Ross Beveridge, Bruce A Draper, Rahul Bangar, Dhruva Patil, James Pustejovsky, Nikhil Krishnaswamy, Kyeongmin Rim, Jaime Ruiz, et al. 2019. User- aware shared perception for embodied agents. In *2019 IEEE International Conference on Humanized Computing and Communication (HCC)*, pages 46– 51. IEEE.

James Pustejovsky and Nikhil Krishnaswamy. 2016. VoxML: A visualization modeling language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).