Situated Grounding Facilitates Multimodal Concept Learning for AI

Nikhil Krishnaswamy and James Pustejovsky
nkrishna@brandeis.edu • jamesp@brandeis.edu

Brandeis University

Introduction

Robust communicative interaction between humans and computers requires the following three capabilities:

1. Recognition and generation within multiple modalities, e.g., language, gesture, vision, action;
2. Understanding of contextual grounding and co-situatedness in conversation;
3. Appreciation of consequences of actions taken throughout the dialogue.

Central to these is "semantically grounding" a concept to a situation;

- Certain modalities are better at *grounding* certain types of information
- (e.g., deixis to locations, language to attributives or concept labels).

"Multimodal linking" is insufficient
***Situated grounding* entails knowledge of entities in context ("common ground")**

"What am I pointing at?"

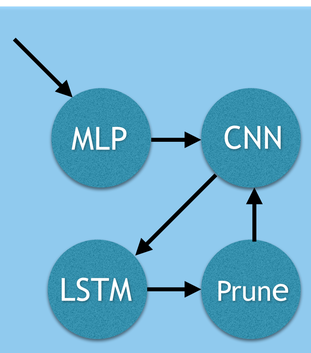
Studying common ground in situated communication and grounding semantic representations to parameters and constraints of situated artifacts allows us to better understand the emergence of linguistic reference in communication without common ground.

Learned Data Sample

Generated examples

Naive users instructed an agent to build a 3-step staircase using language and gesture.

Agent trained over those samples to generate novel examples of the same structure.



Learning framework

Learning inputs were qualitative relations between blocks ([*B6*, *left*, *B3*), (*B3*, *right*, *B6*), …]).

Simulation environment facilitates easy extraction of qualitative relations from raw vectors and coordinates.

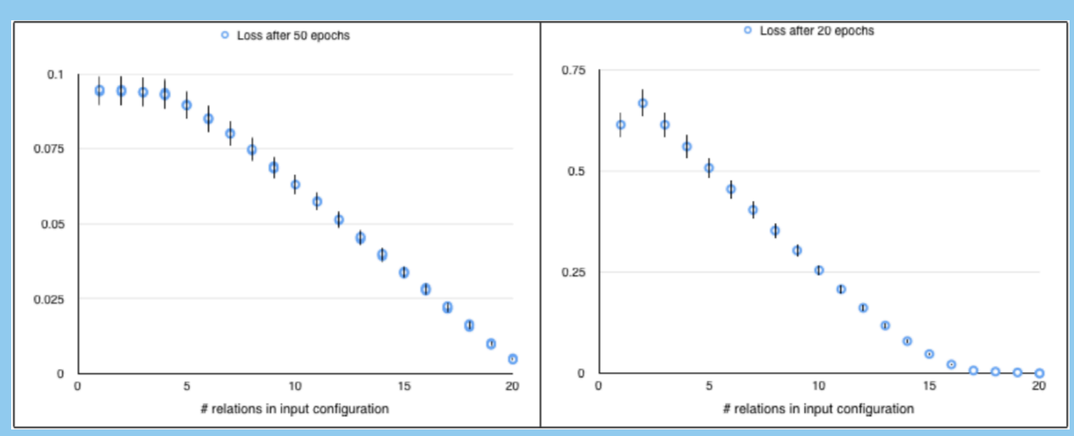
Small dataset allows in-depth assessment of what the model is doing through the learning and generation process, and whether the underlying intuitions and assumptions are backed up by results.

Situated Grounding

- When an agent or user interacts with a simulated world, they adopt a dynamic point of view (or avatar) in that situation.
- When entities in that world can communicate with the user, this creates a correlate to peer-to-peer communication.
- Simulations containing such agents create natural environments for *multimodal learning*, given the right semantic scaffold.
- SItuationally grounding computational behaviors brings up interpretative questions similar to those exhibited by a human.
  - "Which *X*?"
  - "What does *X* mean?"

Validating a Situated Grounding Model

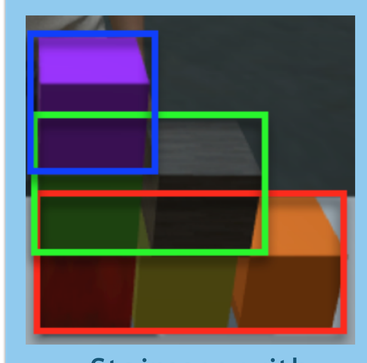
- Model uses a CNN to predict the nearest known sample to the current situation, and an LSTM to generate the most likely sequence of moves to approach it.
- As the structure approaches completion, both these predictions should get less uncertain (lower cross-entropy loss):
  - The closest target example should become clearer, as should the moves needed to get there.
- Validation: measure the training loss while increasing the size of the input to each network.
  - i.e., with 1 relation as input, remaining 19 relations should be very hard to predict; with 19 relations as input, remaining 1 relation needed should be very evident.



CNN training loss

LSTM training loss

Grounding Novel Semantics



Staircase with components marked

- Generating new instances is only part of "grounding";
  - Agents must also be able to recognize and classify learned concepts.
  - We treat this as *constraint satisfaction and inference*.
- Approaches: weighted constraint satisfaction, POMDP, Qualitative Constraint Network
- QCN approach uses combined qualitative spatial relations with interval algebra distinctions;
  - e.g., *E*xternally *C*onnected (touching) vs. *D*isconnected
  - Given structural components, what relations between components satisfy the constraints that appear in the learned samples and generated examples?

$$
\begin{bmatrix}
\textbf{staircase} \\
\text{LEX} = \ldots \\
\text{TYPE} = \begin{bmatrix} \text{HEAD} = \textbf{assembly[1]} \\ \text{COMPONENTS} = \textbf{base[2]},\textbf{step[3]*},\textbf{top[4]} \\ \ldots = \end{bmatrix} \\
\text{HABITAT} = \begin{bmatrix} \text{INTR} = {}_{[5]}\begin{bmatrix} \text{BASE} = align([2], \mathcal{E}_X) \\ \text{UP} = align(vec(loc([4]) - loc([2])), \mathcal{E}_Y) \end{bmatrix} \end{bmatrix} \\
\text{AFFORD\_STR} = \begin{bmatrix} \text{A}_1 = H_{[5]} \rightarrow [put(x, on([1]))]part\_of(x, [1]) \\ \text{A}_2 = H_{[5]} \rightarrow [put(x, on([2]))]part\_of(x, [3]) \\ \text{A}_3 = H_{[5]} \rightarrow [put(x, left \vee right \vee touching([2]) \wedge \neg on([2]))]extend(x, [2]) \\ \text{A}_4 = H_{[5]} \rightarrow [put(x, left \vee right \vee touching([3]) \wedge \neg on([3]))]extend(x, [3]) \end{bmatrix} \\
\text{EMBODIMENT} = \ldots
\end{bmatrix}
$$

Situatedness goes beyond visual grounding. It is a true multimodal approach to demonstrating meaning. Demonstrating knowledge ensures shared understanding.

With a semantic scaffold to transform quantitative values into qualitative values, situated simulation affords tractable language understanding.