Situated Grounding Facilitates Multimodal Concept Learning for AI

Nikhil Krishnaswamy
Department of Computer Science
Brandeis University
Waltham, MA 02453
nkrishna@brandeis.edu

James Pustejovsky
Department of Computer Science
Brandeis University
Waltham, MA 02453
jamesp@brandeis.edu

Abstract

Peer-to-peer human-computer interactions require a minimum level of capability that remains beyond current unimodal approaches. Computers must recognize and generate communicative acts within multiple modalities, understand the grounding of communicative acts within the shared context and situation of both interlocutors, and appreciate the consequences of behavior and actions within the interaction. In this short paper, we discuss an approach to interactive concept learning using *multimodal simulations* that situate and contextualize the interaction, thereby visually demonstrating what the computer believes and understands. We examine an example of situated grounding in a collaborative task, and its uses in probing learned models and interactive learning.

1 Introduction

Any robust communicative interaction between humans and computers or robots will require at least the following three capabilities: (a) robust recognition and generation within multiple modalities, including language, gesture, vision, and action; (b) an understanding of contextual grounding and co-situatedness in the conversation; and (c) an appreciation of the consequences of behavior and actions taking place throughout the dialogue. Central to all of these capabilities is the notion of "semantically grounding" a concept to the current situation. Language use may reflect only a subset of all properties of the current situation, where a full description may be impossible or at least unwieldy. Some kinds of information may in fact be more efficiently communicated using other modalities, such as gesture (e.g., deixis for pointing), demonstration or action, images, relative configurations, or some other *visual* modality.

Work on "multimodal semantic grounding" in the natural language and image processing communities has produced various large corpora linking lexemes or captions with images. Some of these corpora augment the multimodal linkages with other information, like semantic roles [23], bounding boxes [22], or visual attention heatmaps [4], however, in this paper, we argue that language understanding and linking to abstract instances of concepts in other modalities is insufficient; *situated grounding* entails knowledge of situational and contextual entities beyond multimodal linking.

Imagine interacting with an iPhone, Google Assistant, or Amazon Echo. A question that would be completely ordinary in a person-to-person interaction, such as "What am I pointing at?" results in the agent dodging the question (Try it with Siri: if an answer is provided it is usually something like "Interesting question."). It does this because the agent must provide an answer, but is unable to interpret the question because it lacks key information, such as what is present in the situation, how the asker is situated relative to it, where the asker is pointing, etc.; i.e., because it lacks a live vision feed and appropriate machinery for it, it cannot fully communicate using language alone.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

What we propose instead is that the future of computational language understanding and intelligent agents lies in a framework for studying interactions and communication between agents engaged in a shared goal or task (peer-to-peer communication). When two or more people are engaged in dialogue during a shared experience, they share a common ground, which facilitates situated communication. By studying the constitution and configuration of common ground in situated communication, we can ground semantic representations to the parameters and constraints of actual artifacts in the discourse and situation, and we can better understand the emergence of decontextualized linguistic reference in communicative acts, where there is no common ground.

2 Situated Grounding

When an agent or user interacts with entities in a virtual or simulated world, the agent adopts a dynamic point of view or avatar in that proxy situation. When entities in the virtual world can communicate with the user, this creates a ready correlate with peer-to-peer communication, as between humans, albeit one mediated by a computational, rather than biological, platform. Such situations are often depicted in video games, where the AI driving non-user agents (a.k.a. non-player characters or NPCs) is likely rudimentary or underdeveloped relative to real human language faculties. Unlike multimodal grounding systems over static images [5], it is only recently that computational agents have begun to learn using continuous simulated data [11], and learning through instructions or dialogue while situated in an environment is more nascent still [2]. Nonetheless, we argue that the virtual simulations within which such agents reside creates a natural environment for a kind of *multimodal* learning, given the right semantic scaffold.

There has been some discussion in the Human-Robot Interaction literature on how to resolve ambiguities that may arise from utterances in situated dialogues. For example, depending on the situation, the definite description in the command “Open the box" may uniquely refer or not, depending on how many boxes (if any) are in the context. These and similar miscommunications or the need for clarification in dialogue are called *situated grounding problems* [15], and can be viewed as problematic only in a model that appeals to and encodes both a visual modality and situational information into the dialogue state. What the occurrence of these issues makes apparent is the complexity underlying the interpretation of referential expressions in actual situated dialogues. The richness provided by situationally grounding computer or robot behaviors brings to the surface interpretive questions similar to those exhibited by a human in the same scenario, e.g. “Which *X*?" or “What does *X* mean?".

Figure 1: Interactive situated environment (taken from [13])

We have previously explored one such scenario, shown in Fig. 1 [13]. Using data acquired from naive users interacting with an avatar in a mixed-reality environment, where the avatar residing in a virtual world with virtual objects can see, hear, and respond to instructions given by a human using real spoken language and gestures, the avatar was trained to produce novel examples of a structure previously not in its vocabulary (in this sample, a 6-block staircase). Interestingly, inputs to the training, which is a combined CNN-LSTM method over a graph-matching heuristic function, are not vector representations of the block coordinates or overall position in the completed structure, but qualitative relation sets between pairs of blocks, e.g., $[(B6,\ left,\ B3), (B3,\ right,\ B6),\ ...]$, where each block and relation are given numerical indices. We argued that the simulated environment facilitates the easy extraction of qualitative relations from raw object vectors and coordinates, and

that visualizing the avatar building examples of the newly-learned structure allows a human to easily validate or reject the new sample generated by the agent's learned model.

This sample, from a small dataset in a straightforward Blocks World domain, that outputs qualitative relations, allows one to assess in some depth what the model is doing at each step and whether the intuitions behind its training pipeline, that depend heavily on direct visual and situated grounding, are backed up by the results.

3 Validating a Situated Grounding Model

As described in [13], the intuition behind this regime for training the model is that as the avatar starts placing blocks, the number of possible alternative paths to completion of the target structure (i.e., a satisfactory staircase), decreases. We used a CNN to predict a most likely target example at each step and an LSTM to generate the most likely sequence of remaining moves to get there. As the structure gets closer and closer to completion, both of these predictions should get less and less uncertain (i.e., lower and lower cross-entropy loss).

To validate this, we subsequently reran the same training data from [13] through the same model, but where that study ran all combinations of input relations to predict the remaining relations (i.e., input 1 relation to predict remaining 19, input 2 relations to predict remaining 18, etc.), we measured the loss across epochs (n=50 for the model's CNN and n=20 for the model's LSTM) while steadily increasing the size of the input by 1 relation each time. We predicted that, should the intuition in [13] hold, we would observe a steady attenuation in cross-entropy loss across both networks until, when the size of the input reaches $max - 1$, the loss should be very close to 0 (as it should be almost certain which example structure the agent is approaching, or which particular relations remain to create it).

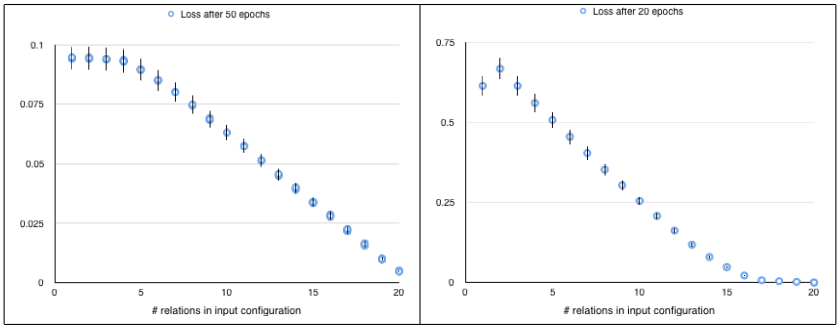

Figure 2: Cross-entropy loss vs. size of input to training for CNN after 50 epochs (L) and LSTM after 20 epochs (R)

Fig. 2 shows results over 5 trials with the input relations randomized each time. As predicted, loss steadily drops across both networks as input window size increases. In addition, despite randomizing the particular inputs selected each time, the loss value per window size remained very close across all 5 trials, with an average σ per window size of 2.586×10^{-4} over the CNN and 1.974×10^{-4} over the LSTM. This suggests that features drawn from situated grounding in interactions can serve as effective and regular model biases to train a model of a novel vocabulary item.

4 Grounding Novel Semantics

Generating new instances of a concept is only part of the "grounding" problem involved. An agent must also be able to classify and recognize instances of the new concept, and not just produce them. This is a far more difficult problem than the procedural building task, but we propose a solution once again facilitated by situated multimodal grounding.

We recast the problem as one of *constraint satisfaction* instead of regression or similarity. Fig. 3 shows a sample agent-constructed staircase generated by its model. After delineating and labeling the components of the structure, in this case, the top, step, and base, we can begin to automatically infer the constraints that inhere between those components in this sample, and across all samples generated by the agent and validated by a human interlocutor.

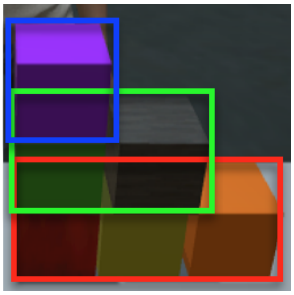

Figure 3: Staircase with base, step, and top components marked

This can be approached from many angles, such as a weighted constraint satisfaction problem (as some constraints, such as the requirement that all blocks be flush with another, can be relaxed in some samples), or as a partially-observable Markov decision process, or POMDP (where the agent begins with the known components—here, base, step, and top—as observables, and then reasons by adding constraints or constraint conjuncts/disjuncts to its belief MDP and querying the belief against the known state of the world, thus assigning block-relation rules to the structural components; cf. [14]).

One approach in particular that we are interested in pursuing is a qualitative constraint network (QCN) [17, 21]. We propose a form of the algorithm outlined in [17] where Allen Temporal Relations [1] are replaced with the relations used in [13] (from the QSRLib relation library [8]). As above, we allow conjuncts and disjuncts, and also keep interval algebra distinctions for "flush" vs. "separated" (cf. "Externally Connected vs. $D$is$C$onnected from the Region Connection Calculus [20]).

From initial trial runs over the sample data from [13] and other subsequently generated positive samples, we encoded the results of some sample outputs using the habitats [16, 18] and affordances [9, 10] of the VoxML semantic modeling language [19] on which the avatar-interaction system from which we sourced the original training examples is built; an example is shown below.

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

This illustrates that we can successfully extract certain constraints that describe not just the staircase shown, but an abstract staircase. These constraints include: the steps ascending to either the left *or* right (A_3, A_4), that placing an object on the base or step creates a new (non-base or -step) tier (A_3, A_4), or that putting something on the base makes it part of the step (A_2). This provides at least some of the components of a semantic model for the new object. When asked about a "staircase," the agent now has semantics for a decontextualized reference that can be reproduced and adjusted without having to retrain the purely numerical model. This allows us to probe a learned model in a tractable way by examining qualitative relations and derived constraints that are provided by situated grounding.

5 Conclusion

Situatedness goes beyond visual grounding; it is a true multimodal approach to demonstrating meaning and understanding. We believe that simulation can play a crucial role in human-computer

communication; it creates a shared epistemic model of the environment inhabited by a human and an artificial agent, and demonstrates the knowledge held by the agent publicly. Demonstrating knowledge is needed to ensure a shared understanding with its human interlocutor. In a situated environment like a simulation, the human and computer share an epistemic space, and any modality of communication that can be expressed within that space (e.g., linguistic, visual, gestural) enriches the number of ways that a human and a computer can communicate within object and situation-based tasks, such as those investigated by Hsiao et al. [12], Dzifcak et al. [7], and Cangelosi [3], among others. If an agent is able to receive information from a human and interpret that relative to its current physical circumstances, it can create an epistemic representation of that same information. However, without a modality to express that representation independently, the human is unable to verify or query what the agent is perceiving or how that perception is being interpreted. A situated representation natively provides these modalities, such as visualized qualitative relations, differences in frame of reference, and the consequences of actions in satisfying constraints.

Davis and Marcus [6] make a strong argument against the efficacy of simulation in explaining natural language understanding, particularly regarding linguistic phenomena involving continuous ranges or underspecified values. We argue that with the addition of a semantic scaffold to render raw quantitative values into qualitative values, situated simulation can provide facilitate tractable computational language understanding. Existing computational approaches to semantic processing, when taken together, provide a framework on which to implement a simulator as an extension of a model builder. Situated simulation, when used to model a dynamic qualitative spatial and temporal semantics, can provide a robust environment for examining the interpretation of linguistic behaviors, including those described multimodally.

Acknowledgments

We would like to thank the reviewers for their helpful comments. This work is supported by a contract with the US Defense Advanced Research Projects Agency (DARPA), Contract W911NF-15-C-0238. Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

[1] Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM 26*(11), 832–843.

[2] Anderson, P., Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683.

[3] Cangelosi, A. (2010). Grounding language in action and perception: from cognitive agents to humanoid robots. *Physics of life reviews 7*(2), 139–151.

[4] Das, A., H. Agrawal, L. Zitnick, D. Parikh, and D. Batra (2017). Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding 163*, 90–100.

[5] Das, A., S. Kottur, J. M. Moura, S. Lee, and D. Batra (2017). Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2951–2960.

[6] Davis, E. and G. Marcus (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence 233*, 60–72.

[7] Dzifcak, J., M. Scheutz, C. Baral, and P. Schermerhorn (2009). What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pp. 4163–4168. IEEE.

[8] Gatsoulis, Y., M. Alomari, C. Burbridge, C. Dondrup, P. Duckworth, P. Lightbody, M. Hanheide, N. Hawes, D. Hogg, A. Cohn, et al. (2016). Qsrlib: a software library for online acquisition of qualitative spatial relations from video.

[9] Gibson, J. J. (1977). The theory of affordances. *Perceiving, Acting, and Knowing: Toward an ecological psychology*, 67–82.

[10] Gibson, J. J. (1979). *The Ecology Approach to Visual Perception: Classic Edition*. Psychology Press.

[11] Hermann, K. M., F. Hill, S. Green, F. Wang, R. Faulkner, H. Soyer, D. Szepesvari, W. Czarnecki, M. Jaderberg, D. Teplyashin, et al. (2017). Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*.

[12] Hsiao, K.-Y., S. Tellex, S. Vosoughi, R. Kubat, and D. Roy (2008). Object schemas for grounding language in a responsive robot. *Connection Science 20*(4), 253–276.

[13] Krishnaswamy, N., S. Friedman, and J. Pustejovsky (2019). Combining Deep Learning and Qualitative Spatial Reasoning to Learn Complex Structures from Sparse Examples with Noise. In *AAAI Conference on Artificial Intelligence (AAAI)*. AAAI.

[14] Lee, J., G.-H. Kim, P. Poupart, and K.-E. Kim (2018). Monte-carlo tree search for constrained pomdps. In *Advances in Neural Information Processing Systems*, pp. 7923–7932.

[15] Marge, M. and A. I. Rudnicky (2013). Towards evaluating recovery strategies for situated grounding problems in human-robot dialogue. In *2013 IEEE RO-MAN*, pp. 340–341. IEEE.

[16] McDonald, D. and J. Pustejovsky (2013). On the representation of inferences and their lexicalization. In *Proceedings of the Second Annual Conference on Advances in Cognitive Systems ACS*, Volume 135, pp. 152. Citeseer.

[17] Mouhoub, M., H. Al Marri, and E. Alanazi (2018). Learning qualitative constraint networks. In *25th International Symposium on Temporal Representation and Reasoning (TIME 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[18] Pustejovsky, J. (2013). Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pp. 1–10. ACL.

[19] Pustejovsky, J. and N. Krishnaswamy (2016, May). VoxML: A visualization modeling language. In N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

[20] Randell, D., Z. Cui, A. Cohn, B. Nebel, C. Rich, and W. Swartout (1992). A spatial logic based on regions and connection. In *KR'92. Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference*, San Mateo, pp. 165–176. Morgan Kaufmann.

[21] Tong, Y. and Q. Ji (2008). Learning bayesian networks with qualitative constraints. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE.

[22] Yang, S., Q. Gao, C. Liu, C. Xiong, S.-C. Zhu, and J. Y. Chai (2016). Grounded semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 149–159.

[23] Yatskar, M., L. Zettlemoyer, and A. Farhadi (2016). Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5534–5542.