Embodied Multimodal Agents to Bridge the Understanding Gap

Nikhil Krishnaswamy
Department of Computer Science
Colorado State University
Fort Collins, CO, USA
nkrishna@colostate.edu

Nada Alalyani
Department of Computer Science
Colorado State University
Fort Collins, CO, USA
n.alalyani@colostate.edu

Abstract

In this paper we argue that embodied multimodal agents, i.e., avatars, can play an important role in moving natural language processing toward "deep understanding." Fully-featured interactive agents, model encounters between two "people," but a language-only agent has little environmental and situational awareness. Multimodal agents bring new opportunities for interpreting visuals, locational information, gestures, etc., which are more axes along which to communicate. We propose that multimodal agents, by facilitating an *embodied* form of human-computer interaction, provide additional structure that can be used to train models that move NLP systems closer to genuine "understanding" of grounded language, and we discuss ongoing studies using existing systems.

1 Introduction

As of the 2020s, high-profile NLP successes are being driven by large (and ever-growing) deep neural language models[1]. These models perform impressively according to common metrics on a variety of tasks such as text generation, question answering, summarization, machine translation, and more.

However, deep examinations of large language models show that they may exploit artifacts or heuristics in the data (McCoy et al., 2019; Niven and Kao, 2019). Similar observations have been made about computer vision systems (Zhu et al., 2016; Barbu et al., 2019). The black-box nature of much NLP leads to the argument that they are insufficient at demonstrating understanding of communicative intent and also cannot explain why or when these failures occur. (Bender and Koller, 2020).

[1]As of writing, the largest language model is the Switch Transformer by Google (Fedus et al., 2021), a trillion-parameter language model beating previous record-holder GPT-3 (Brown et al., 2020)

The *neurosymbolic* approach to AI has long argued for *structured representation* (Garcez et al., 2015; Besold et al., 2017), and similar arguments have also been made by deep learning luminaries e.g., Bengio (2017). Introducing preexisting structure into NLP facilitates higher-order reasoning, but a hybrid approach also does so at larger scales than purely symbolic systems, by using flexible deep-learning representations as input channels.

Multimodal NLP systems, particularly those with embodied agents, model encounters between two "people." A fully-featured interaction requires an encoding of structures that make language interpretable and "deeply understandable." This makes embodied multimodal interaction a uniquely useful tool to examine what NLP models in use meaningfully learn and "understand.". If one mode of expression (e.g., gesture) is insufficiently communicative, another (e.g., language) can be used to examine where it went wrong. Each additional modality provides an avenue through which to validate models of other modalities.

In this paper we review common components of embodied multimodal agent systems and their contributions to "deep understanding." We present ongoing experiments in agents who exhibit grounded understanding, using complex multimodal referring expressions as an example task.

2 Multimodal Conversational Systems

As AI systems become more integrated with everyday life, users at times harbor misapprehensions that they behave like humans. This dates as far back as ELIZA (Weizenbaum, 1966), whose users were convinced that ELIZA "understood" them despite Weizenbaum's insistence otherwise. The well-known SHRDLU displayed some situational understanding, but SHRDLU's deterministic perception was not sensor-driven (Winograd, 1972).

Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 41–46

Quek et al. (2002) and Dumas et al. (2009), among many others, define *multimodal* interaction as interactions with multiple distinct input channels, e.g., spoken language and gestures, that complement and reinforce each other. This relates to "communicative intent" in that if one channel is ambiguous, the other(s) may clarify or add to the meaning. Multimodal interactions may increase understanding by making it easier to retrieve a communicative intent from a combination of inputs. This makes it critical in computer systems that display understanding.

Bolt's "Put-that-there" (1980) anticipated some critical issues, including the use of deixis for disambiguation. The subsequent community that evolved around multimodal integration (e.g., Kennington et al. (2013), Turk (2014)) gave rise to a number of *embodied conversation agents* (ECAs) (e.g., Cassel (2000)) deployed in different task domains, including education (Barmaki and Hughes, 2018), negotiations (Mell et al., 2018), and medical practitioner evaluation (Carnell et al., 2019).

Humans communicate contextually; they may gesture and refer to items in their co-perceived space. This co-perception and co-attention is "central to determining the meanings of communicative acts [...] in shared events" (Pustejovsky, 2018). A communicative act, C_a, can be modeled as a tuple of expressions from the modalities available to the agent, which convey complementary or redundant information. For example, if the modalities involved are Speech, Gesture, Facial expression, ga$Z$e, and Action, then $C_a = \langle S, G, F, Z, A \rangle$, of which any may be null or empty. Information in one channel may be supplemented by another channel, and they may disambiguate each other if properly aligned. For instance, if interpreted from the human's point of view, $C_a = \langle S = \text{“}left\text{”}, G = [Point_g \wedge Dir = \text{RIGHT}] \rangle$ (cf. Kendon (2004), Lascarides and Stone (2009)), this may signal a difference in the agents' relative frames of reference. This effect of embodiment is critical for deep understanding of situated language, vis-à-vis something as basic as directional terms.

An *embodied* agent can demonstrate aspects of their worldview, including how they ground and interpret utterances, gestures, or the consequences of actions, by acting themselves in their worlds. "Embodied worlds" may be virtual, physical, or mixed-reality. In the remainder of this paper we focus on virtual worlds.

To align various modalities in the state space accessed by the agent's dialogue system, we assume a "common ground" structure associated with a dialogue state. This state monad (Unger, 2011; Bekki and Masuko, 2014): $\mathbf{M}\alpha = State \rightarrow (\alpha \times State)$, corresponds to those computations that read and modify the state. **M** is a type constructor that consumes a state and returns a pair: $\langle value, modified\_state \rangle$.

A system operating under the aforementioned assumptions is "Diana" (McNeely-White et al., 2019). Diana's semantic knowledge of objects and actions is based on the VoxML modeling language (Pustejovsky and Krishnaswamy, 2016), and she asynchronously interprets spoken language and gesture to collaborate on construction tasks with humans. The human instructs Diana by deictically or linguistically referencing objects and indicating what to do with them multimodally. Diana's responses and *actions* make clear whether she understands what the user intended or not: a clear extraction of communicative intent from utterance, and a measure of "deep understanding."

However, Diana's capabilities are not fully symmetrical. The human may speak verbosely, but Diana's responses are brief: "OK," or the occasional disambiguatory question. To increase Diana's capacity for deep understanding, and symmetric retrieval of intent from utterance on the part of both Diana and the human, we are conducting experiments to fully develop the capabilities afforded by Diana's multimodal embodiment. These are outlined subsequently.

3 Ongoing Experiments

Referring expressions (REs) provide a defined and evaluable case study in the ability of an NLP system to extract communicative intent from an utterance. Either the agent correctly retrieves the object the human referenced or it does not; either the human can correct misunderstanding or they cannot.

Referring expressions exploit information about both object characteristics and locations. Linguistic referencing strategies can use high-level abstractions to distinguish an object in a given location from similar ones elsewhere, yet the location described may still be difficult to ground or interpret. When interacting person-to-person, humans deploy a variety of strategies to clearly indicate objects and locations in a discourse, including mixing and matching modalities on the fly and switching strate-

gies to optimize both economy and clarity. Therefore, in an ongoing study to replicate this same capacity in interactive agents, we are exploring how to generate symmetrically-descriptive references of distinct objects in context.

Krishnaswamy and Pustejovsky (2019) presented the EMRE (Embodied Multimodal Referring Expressions) dataset. It contains 1,500 videos of a version of the Diana agent referencing multiple objects in sequence with a mix of language and gestures. The authors found that human evaluators found multimodal referring strategies most natural and preferred more descriptive language. However, they also found that the data gathered was not enough to train an effective generation model, so we build on this strategy to augment the available data to a sufficient level.

3.1 Multimodal Referring Expression Generation

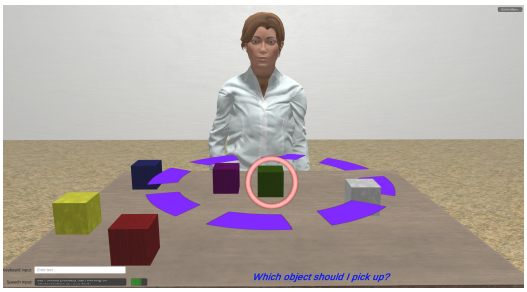

Figure 1: The pink circle indicates the target object. Diana asks "Which object should I pick up?" and waits for her partner to reference an object multimodally. The purple circle shows where the user is pointing.

We are deploying a web-based version of the Diana system, where participants are presented with a scene populated with randomly-placed objects, including objects that are identical in nominal attributes like color, as in the EMRE dataset. The human participant will be prompted analogously to what is shown in Fig. 1. The experimental system indicates one object of focus and Diana prompts the user to generate a referring expression with a question, e.g., "Which object should I pick up?" We then log information about how the human participant responds to the question, and about the scene (which act as contextual parameters). These include including the target object, distance from the agent, object coordinates, relations in the scene, previously referred-to objects; and about how the user responds, including complete utterance, modality(ies) used, attribute (identity) based description, demonstratives used, and relational descriptions used relative to other objects including those referred to at previous points in the dialogue.

Some of these parameters are identical to those in the EMRE dataset; others are included because we hypothesize that they will be useful to our proposed generation models (§3.1.2).

The multimodal inputs our system captures from the user are speech, transcribed via automated speech recognition, and mouse-based deictic gesture as a proxy for live pointing (see more details in §3.1.1). The modes of presentation available to Diana are: speech (generated via text-to-speech), pointing (via animated gesture), and visualized action (acting directly on virtual objects in the scene).

Participants in this study will be given explicit instructions as to the nature of the task, including that they should speak to Diana and can use the mouse to point to objects, to construct the conceit that Diana is a peer, such that the participants treat her like an appropriately collaborative agent. Depending on the crowdsourcing platform, we will also be able to filter the participants we solicit to make sure they have access to appropriate equipment, like a high-quality computer microphone.

Fig. 2 contains an example dialogue, which shows how Diana both extracts distinct objects from multimodal referring expressions and, when those objects do not match the human's intent, the human is able to correct her, also multimodally.

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*]

Figure 2: Sample dialogue.

We expect to deploy this study on a crowdsourcing platform like Prolific or Amazon Mechanical Turk, and aim to solicit interactions from approximately 250 workers using multimodal referring expressions while interacting with Diana in the task describe above. Each worker will view 10 scenes with different configurations in which to refer to up to 10 distinct target objects, resulting in a total of 25,000 samples. These multimodal referring ex-

pressions are expected to fall into three categories; attributive REs, relational REs, and historical REs, as we describe in §3.1.2.

We anticipate taking 3 months to build the web-based version of Diana; 1 month to gather raw data from workers; 1 month to annotate the data; 2 months to perform statistical analysis on the data and investigate good model features; and 3-4 months to design, build, and train our proposed RE generation models.

3.1.1 Input Quality Assurance

For the avatar to respond appropriately, inputs need to be as clean as possible. In our multimodal use case, this includes the speech input, the pointing input, and the parse.

We solicited voice recordings of 30 university graduate students reading scripts containing domain-specific vocabulary. These cover diverse vocal profiles, including accent and timbre, so speech models suitable for these voices should be suitable for a crowd-sourced study. We are using these to assess the quality of speech recognition (e.g., Google Cloud ASR) in the task domain.

Our web-based Diana uses the mouse as a proxy for deixis, instead of CNN-based gesture recogntion. We want to avoid participants "gaming the system" by using accurate mouse deixis alone, so we build some variance into this deixis proxy. The purple circle shown in Fig. 1 fluctuates in size proportionally to the variance in pointing location of the aforementioned gesture recognizer. The location of mouse "deixis" is imprecise by design, forcing participants to also use language to properly describe the target object.

Parsers under investigation are Stanford-CoreNLP and Google's SyntaxNet, and are being evaluated for the the syntactic dependencies they provide and the ability to extract entities and constituent relationships between them.

3.1.2 Proposed Generation Models

When an embodied interactive agent is a partner in a dialogue, this provides a layer of exposure not present in end-to-end systems. The agent's responses can be used to probe where in the pipeline errors occur, be it in the speech recognition, the parse, or the interpretation. As discussed in §2, the existence of multiple modal channels allows investigation into which channel, e.g., the language, the gesture, the gaze/attention, etc. is the likely cause.

Assuming maximally correct inputs, as discussed in §3.1, existing NLP technologies, combined with an embodied multimodal agent, creates a mechanism to examine how multiple modalities can combine to signal aspects of an input that the agent can understand, or extract intent from, and reproduce, or communicate its own intent.

The EMRE dataset contains referring expressions generated via a mix of stochastic sampling and slot filling. From our crowd-sourced study, we plan to extract three particular kinds of additional data to train our multimodal referring expression generation (MREG) models:

- **Attributes** used to denote objects (A-MRE);
- **Relations** to other objects used to describe the target object (R-MRE);
- Dialogue **history** invoked when describing objects (H-MRE), which involves modeling when previous introductions into the common ground lose relevancy.

Attributes, relations, and actions in the history are encoded in VoxML, and serve as the *symbolic* inputs to our neurosymbolic architectures.

We intend our approaches to generate multimodal referring expressions in context that can blend modalities at runtime, for instance using demonstratives with gesture and distance of the target object relative to the agent.

Our proposed MREG models are centered around LSTMs (Hochreiter and Schmidhuber, 1997) for their ability to capture sequential information as occurs in a dialogue. MREG model outputs are likely to take the form $\langle Modality, Utterance, Location, Demonstratives \rangle$, where $M \in [Gesture, Language, Ensemble]$, U is a decoded sentence embedding, L is the location the gesture grounds to, and $D \in [this, that]$. Depending on the value of M, some of the other parameters may be empty by default. Fig. 3 shows schematics of the MREG architectures we are currently exploring.

A-MRE Object attributes are predicted by training an LSTM (**A-LSTM**) over the attributive terms participants use to describe objects (e.g., constituents tagged as $nmod$ in a parse), tracking how they vary terms for the same object over time. The A-LSTM takes a query representing the target object and outputs a descriptor tuple $\langle M, U, L, D \rangle$.

R-MRE Relations are predicted with another LSTM (**R-LSTM**) that takes as input C, pairwise permutations of the target object's configuration

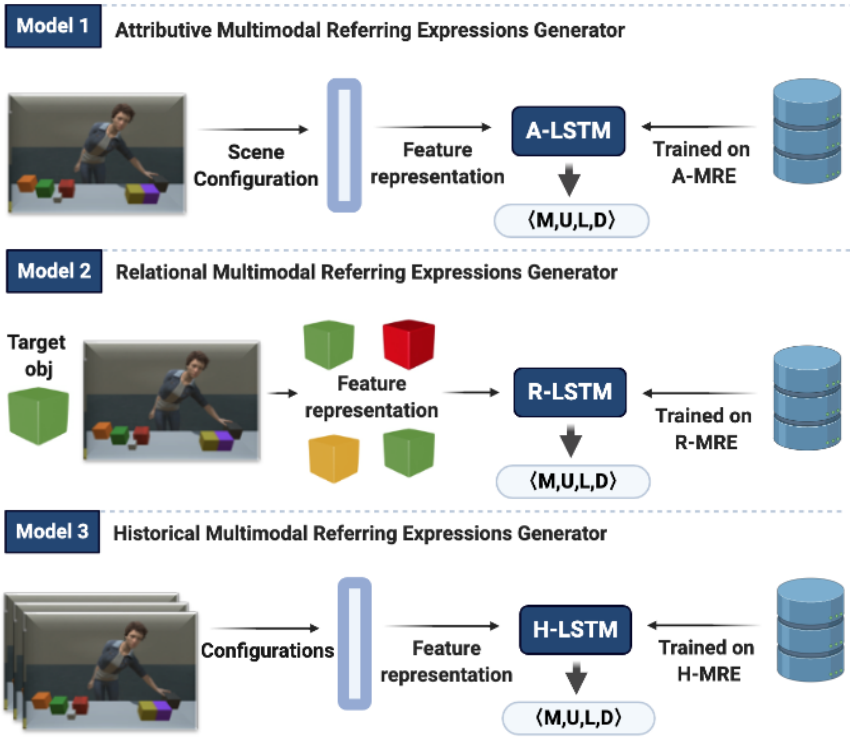

Figure 3: MREG architectures under exploration

relative to objects around it. E.g., in Fig. 3, if the target is "green block," then the proximate objects (orange and red blocks) are used to create binary relation pairs with them and the target (e.g., $left(O, G) \leftrightarrow right(G, O)$). These inputs are then fed to the R-LSTM trained over the R-MRE data with a query consisting of the target, which outputs relational descriptors of the target.

H-MRE Instead of features of a single configuration, the **H-LSTM** is trained over the H-MRE data, or sequences of configurations and the moves the agent makes that led to each of them. The job of the H-LSTM is to encode the features of this list and return the decoded multimodal RE expressions compatible with the given list of configurations.

4 Conclusion

Clark (1996) casts language as a joint activity. We propose that introducing the interactive element into natural language technology enables a number of opportunities to build systems that go beyond *processing* language to *understanding* it. We focus our experiments on embodied systems because they create a conceit of interacting with another "person." If these systems can interact in a way that suggests understanding of their interlocutor's intents, they will have demonstrated a step toward true computational natural language understanding. We are pursuing experiments in multimodal referring expression generation as an illustrative use case.

References

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pages 9448–9458.

Roghayeh Barmaki and Charles E Hughes. 2018. Embodiment analytics of practicing teachers in a virtual immersive environment. *Journal of Computer Assisted Learning*, 34(4):387–396.

Daisuke Bekki and Moe Masuko. 2014. Meta-lambda calculus and linguistic monads. In *Formal Approaches to Semantics and Pragmatics*, pages 31–64. Springer.

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

Yoshua Bengio. 2017. The consciousness prior. *arXiv preprint arXiv:1709.08568*.

Tarek R Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Daniel Lowd, Priscila Machado Vieira Lima, et al. 2017. Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*.

Richard A Bolt. 1980. *"Put-that-there": Voice and gesture at the graphics interface*, volume 14. ACM.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Stephanie Carnell, Benjamin Lok, Melva T James, and Jonathan K Su. 2019. Predicting student success in communication skills learning scenarios with virtual humans. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 436–440.

Justine Cassell. 2000. *Embodied conversational agents*. MIT press.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Bruno Dumas, Denis Lalanne, and Sharon Oviatt. 2009. Multimodal interfaces: A survey of principles, models and frameworks. *Human machine interaction*, pages 3–26.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.

Artur Garcez, Tarek R Besold, L d Raedt, Peter Földiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C Lamb, Risto Miikkulainen, and Daniel L Silver. 2015. Neural-symbolic learning and reasoning: contributions and challenges.

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

Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

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.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, page ffp004.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

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.

Johnathan Mell, Jonathan Gratch, Tim Baarslag, Reyhan Aydoğan, and Catholijn M Jonker. 2018. Results of the first annual human-agent league of the automated negotiating agents competition. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 23–28.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

James Pustejovsky. 2018. From actions to events. *Interaction Studies*, 19(1-2):289–317.

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).

Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E McCullough, and Rashid Ansari. 2002. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 9(3):171–193.

Matthew Turk. 2014. Multimodal interaction: A review. *Pattern Recognition Letters*, 36:189–195.

Christina Unger. 2011. Dynamic semantics as monadic computation. In *JSAI international symposium on artificial intelligence*, pages 68–81. Springer.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Terry Winograd. 1972. Shrdlu: A system for dialog.

Zhuotun Zhu, Lingxi Xie, and Alan L Yuille. 2016. Object recognition with and without objects. *arXiv preprint arXiv:1611.06596*.