An Evaluation Framework for Multimodal Interaction

Nikhil Krishnaswamy and **James Pustejovsky**

nkrishna@brandeis.edu • jamesp@brandeis.edu

Overview

As natural language systems become integrated with everyday use, users will come to expect their interactions to approximate communication with another human, multimodally. With increased interest in multimodal interaction comes a need to evaluate the performance of a multimodal system on all levels with which it engages the user. Evaluation should be modality-agnostic and assess the success of communication between human and computer, within the shared context created by the human-computer interaction.

VoxML (Pustejovsky and Krishnaswamy, 2016) serves as the platform for modeling objects, events, and actions, and and the VoxML-based simulation environment VoxSim (Krishnaswamy and Pustejovsky, 2016a, 2016b) implements a multimodal interaction involving natural language and gesture. This allows us to exercise VoxML object and event semantics to assess conditions on the success or failure of the interaction.

Hallmarks of Communication

A "meaningful" interaction with a computer system should model certain aspects of a similar interaction between two humans. Namely, each interlocutor should have something "interesting" to say, and the interaction enables them to work together to achieve common goals and build off each other's contributions. We therefore build the evaluation scheme off of the following qualitative metrics:

1. Interaction has mechanisms to move the conversation forward (Asher and Gillies, 2003; Johnston, 2009)
2. System makes appropriate use of multiple modalities (Arbib and Rizzolatti, 1996; Arbib, 2008)
3. Each interlocutor can steer the course of the interaction (Hobbs and Evans, 1980)
4. Both parties can clearly reference items in the interaction based on their respective frames of reference (Ligozat, 1993; Zimmermann and Freksa, 1996; Wooldridge and Lomuscio, 1999)
5. Both parties can demonstrate knowledge of the changing situation (Ziemke and Sharkey, 2001)

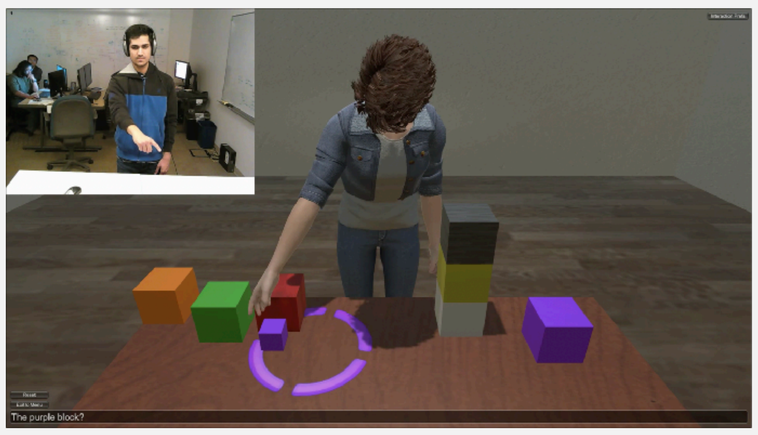
We use distinct semantic properties of specific elements in the interaction to determine what about the interaction enabled or hindered this subjective "shared understanding."

Experimental Setup

- 20 subjects (CS grad students)
- No prior knowledge of avatar's vocabulary
- Told computer could understand language and gesture
- Asked to build 3-stepped staircase with 6 blocks
- Definition of "success" left up to subject
- Logged **interpretable gestures made by human**, **interpretable words spoken by human**, **gesture made by avatar**, **action taken by avatar**, **utterance spoken by avatar**

Human-Avatar-Blocks World (HAB)

A human and an avatar in the VoxSim environment must collaborate to complete a simple construction task using virtual blocks that are manipulated by the avatar.



Example interaction setup showing human (top left) and avatar

- The human must instruct the avatar to reach the goal configuration using a combination of DCNN-recognized gestures and natural language instructions
- The avatar communicates through gestures and natural language output to request clarification of ambiguous instructions or present its interpretation of the human's commands
- The human may indicate (point to) blocks and instruct the avatar to slide and move them relative to other blocks or relative to regions of the virtual table
- The human must also respond to the avatar's questions, when the avatar perceives an ambiguity in the human's instructions.

Data Collection

```
1 HG engage start 1.145281
2 AS "Hello." 1.145281
3 HP r,-0.25,-0.87 4.889832
...
126 HP r,-0.14,-0.62 11.97283
127 HS NO 12.03008
128 AS "Sorry, I don't know what you
  mean." 12.03008
129 HP r,-0.10,-0.41 12.07262
...
```

Example abbreviated log

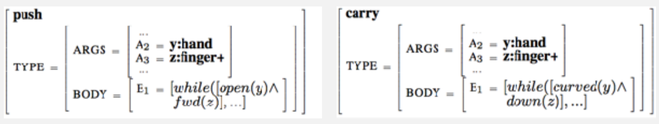
As proxy for the human's understanding of an avatar move, we take the time elapsed between the *first* in a block of avatar moves uninterrupted by a human move, and the human response that follows. Time differences (i.e, human's time to *begin response*) should reflect the clarity or expressiveness of the avatar's move.

The human may make gestures the avatar cannot recognize or interpret, so the human makes multiple moves before the avatar responds. We call this the avatar's time to *recognize content*.

Multimodal Semantics

In the context of shared physical tasks in a common workspace, shared perception creates the context for the conversation between interlocutors (Lascarides and Stone, 2006; Lascarides and Stone, 2009; Clair et al., 2010; Matuszek et al., 2014); it is this shared space that gives many gestures, such as pointing, their meaning:

- **Engage**: Begins and ends the task
- **Positive acknowledge**: Signals agreement or affirmative response to a question
- **Negative acknowledge**: Signals disagreement or negative response to a question
- **Point**: Indicates a region or block(s) in that region.
- **Grab**: Tells the avatar to grasp an indicated block.
- **Carry**: Pick up, move, or put down.
- **Push**: Signals the avatar to push a block in indicated direction



Sample VoxML gesture semantics: [[PUSH]] vs. [[CARRY]]

Effective multimodal systems should support multimodal commands and shared perception, and approximate peer-to-peer conversations. A semantically-informed evaluation scheme, which is intended to be situation- agnostic and relies solely on logging the time and nature of interactions between interlocutors, conditioning on semantic elements during post-processing should scale to domain-agnostic interactions.

Conclusion & Future Directions

We have proposed an evaluation scheme to assess the coverage of multimodal interaction systems and outlined its use evaluating a sample interaction in a system that uses linguistic, gestural, and visual modalities. The example system exploits many advantages of virtual embodiment (Kiela et al., 2016); consistent evaluation is required to discover where the system needs improvement. Our framework can provide this information without very complicated algorithms to process the logged data.

We have presented preliminary results from naive users run through the sample system, which show how we can use simple metrics to assess the ease or difficulty with which specific features communicate information. We believe this type of evaluation will be useful for developing user models and helping researchers assess the gaps in novel computational interaction systems in a variety of modalities, scenarios, and interaction types.

Acknowledgments

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.

Example Results

Response times are charted against the semantic features of the moves that prompted the relevant response. Comparable moves occur in similar semantic contexts: *[mj−n..mj+n]* where *mj* = the move, examined in a window of size *2n + 1*. **P(ti|mj−n..mj..mj+n)** ∝ the probability that response time *t* falls in interval *i* given a move/surrounding context.

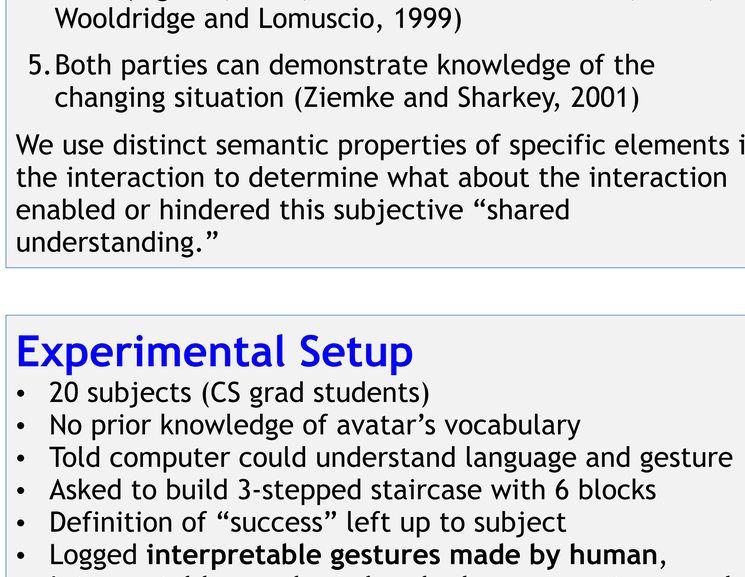


Figure 3: $\mathbf{P}(t_i|M)$, for avatar time to recognize positive/negative acknowledgment gesture (left top/bottom) vs. word "yes"/"no" (right top/bottom)

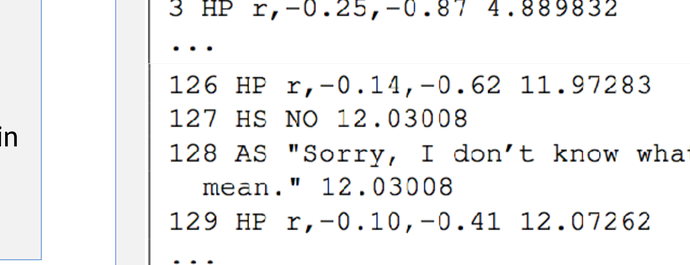


Figure 4: $\mathbf{P}(t_i|M)$ for human time to begin response to [[PUSH]] gesture (left) vs. [[CARRY]] gesture (right)

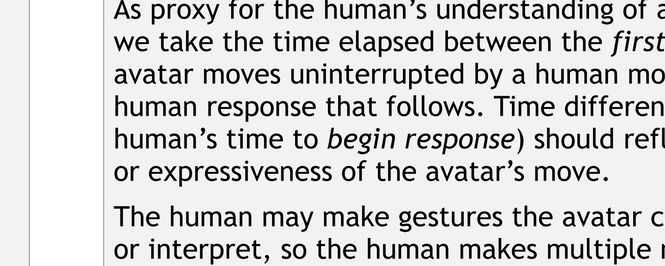


Figure 6: $\mathbf{P}(t_i|M)$ for avatar time to recognize [[POINT]] gesture with the left hand (left) vs. right hand (right)

- Avatar takes more time to recognize spoken "yes" than "no" (human may take longer to communicate a spoken positive acknowledgment than a negative one)
- [[PUSH]] is almost 2x as likely as [[CARRY]] to prompt a "very quick" (first interval) response (may be easier to produce gesture with fingers straight than curved)
- Avatar is quicker to recognize right-handed pointing than left-handed (gesture recognition may have greater variance in detecting left-hand pointing, due to bias in training dataset)

Selected References

Arbib, M. and Rizzolatti, G. (1996). *Neural expectations: A possible evolutionary path from manual skills to language*.

Arbib, M. A. (2008). *From grasp to language: embodied concepts and the challenge of abstraction*.

Asher, N. and Gillies, A. (2003). *Common ground, corrections, and coordination*.

Clair, A. S., Mead, R., Matarić, M. J., et al. (2010). *Monitoring and guiding user attention and intention in human-robot interaction*.

Hobbs, J. R. and Evans, D. A. (1980). *Conversation as planned behavior*.

Johnston, M. (2009). *Building multimodal applications with EMMA*.

Kiela, D., Bulat, L., Vero, A. L., and Clark, S. (2016). *Virtual embodiment: A scalable long-term strategy for artificial intelligence research*.

Krishnaswamy, N. and Pustejovsky, J. (2016a). *Multimodal semantic simulations of linguistically underspecified motion events*.

Krishnaswamy, N. and Pustejovsky, J. (2016b). *VoxSim: A visual platform for modeling motion language*.

Krishnaswamy, N., Narayana, P., Wang, I., Rim, K., Bangar, R., Patil, D., Mulay, G., Ruiz, J., Beveridge, R., Draper, B., and Pustejovsky, J. (2017). *Communicating and acting: Understanding gesture in simulation semantics*.

Lascarides, A. and Stone, M. (2006). *Formal semantics for iconic gesture*.

Lascarides, A. and Stone, M. (2009). *A formal semantic analysis of gesture*.

Ligozat, G. F. (1993). *Qualitative triangulation for spatial reasoning*.

Matuszek, C., Bo, L., Zettlemoyer, L., and Fox, D. (2014). *Learning from unscripted deictic gesture and language for human-robot interactions*.

Pustejovsky, J. and Krishnaswamy, N. (2016). *VoxML: A visualization modeling language*.

Pustejovsky J., Krishnaswamy, N., Draper, B., Narayana, P., and Bangar, R. (2017). *Creating common ground through multimodal simulations*.

Wooldridge, M. and Lomuscio, A. (1999). *Reasoning about visibility, perception, and knowledge*.

Ziemke, T. and Sharkey, N. E. (2001). *A stroll through the worlds of robots and animals: Applying Jakob von Uexkull's theory of meaning to adaptive robots and artificial life*.

Zimmermann, K. and Freksa, C. (1996). *Qualitative spatial reasoning using orientation, distance, and path knowledge*.