Language, Embodiment and Social Intelligence

Matthew Stone
Computer Science and Cognitive Science
Rutgers, The State University of New Jersey
110 Frelinghuysen Road, Piscataway NJ 08854-8019
Matthew.Stone@Rutgers.EDU

Abstract

It is an honor to have this chance to tie together themes from my recent research, and to sketch some challenges and opportunities for NLG in face-to-face conversational interaction.

Communication reflects our general involvement in one anothers' lives. Through the choices we manifest with one another, we share our thoughts and feelings, strengthen our relationships and further our joint projects. We rely not only on words to articulate our perspectives, but also on a heterogeneous array of accompanying efforts: embodied deixis, expressive movement, presentation of iconic imagery and instrumental action in the world. Words showcase the distinctive linguistic knowledge which human communication exploits. But people's diverse choices in conversation in fact come together to reveal multifaceted, interrelated meanings, in which all our actions, verbal and nonverbal, fit the situation and further social purposes. In the best case, they let interlocutors understand not just each other's words, but each other.

As NLG researchers, I argue, we have good reason to work towards models of social cognition that embrace the breadth of conversation. Scientifically, it connects us to an emerging consensus in favor of a general human pragmatic competence, rooted in capacities for engagement, coordination, shared intentionality and extended relationships. Technically, it lets us position ourselves as part of an emerging revolution in integrative Artificial Intelligence, characterized by research challenges like human–robot interaction and the design of virtual humans, and applications in assistive and educational technology and interactive entertainment.

Researchers are already hard at work to place our accounts of embodied action in conversation in contact with pragmatic theories derived from text discourse and spoken dialogue. In my own experience, such work proves both illuminating and exciting. For example, it challenges us to support and refine theories of discourse coherence by accounting for the discourse relations and default inference that determine the joint interpretation of coverbal gesture and its accompanying speech (Lascarides and Stone, 2008). And it challenges us to show how speakers work across modalities to engage with, disambiguate, and (on acceptance) recapitulate each others' communicative actions, to ground their meanings (Lascarides and Stone, In Preparation). The closer we look at conversation, the more we can fit all its behaviors into a unitary framework—inviting us to implement behavioral control for embodied social agents through a pervasive analogy to NLG.

We can already pursue such implementations easily. Computationally, motion is just sequence data, and we can manipulate it in parallel ways to the speech data we already use in spoken language generation (Stone et al., 2004). At a higher level, we can represent an embodied performance through a matrix of discrete actions selected and synchronized to an abstract time-line, as in our RUTH system (DeCarlo et al., 2004; Stone and Oh, 2008). This lets us use any NLG method that manipulates structured selections of discrete actions as an architecture for the production of embodied behavior. Templates, as in (Stone and DeCarlo, 2003; Stone et al., 2004), offer

a good illustration.

Nevertheless, face-to-face dialogue does demand qualitatively new capabilities. In fact, people's choices and meanings in interactive conversation are profoundly informed by their social settings. We are a long way from general models that could allow NLG systems to recognize and exploit these connections in the words and other behaviors they use. In my experience, even the simplest social practices, such as interlocutors' cooperation on an ongoing practical task, require new models of linguistic meaning and discourse context. For example, systems must be creative to evoke the distinctions that matter for their ongoing task, and use meanings that are not programmed or learned but invented on the fly (DeVault and Stone, 2004). They must count on their interlocutors to recognize the background knowledge they presuppose by general inference from the logic of their behavior as a cooperative contribution to the task (Thomason et al., 2006). Such reasoning becomes particularly important in problematic cases, such as when systems must finetune the form and meaning of a clarification request so that the response is more likely to resolve a pending task ambiguity (DeVault and Stone, 2007). I expect many further exciting developments in our understanding of meaning and interpretation as we enrich the social intelligence of NLG.

Modeling efforts will remain crucial to the exploration of these new capabilities. When we build and assemble models of actions and interpretations, we get systems that can plan their own behavior simply by exploiting what they know about communication. These systems give new evidence about the information and problem-solving that's involved. The challenge is that these models must describe semantics and pragmatics, as well as syntax and behavior. My own slow progress (Cassell et al., 2000; Stone et al., 2003; Koller and Stone, 2007) shows that there's still lots of hard work needed to develop suitable techniques. I keep going because of the methodological payoffs I see on the horizon. Modeling lets us take social intelligence seriously as a general implementation principle, and thus to aim for systems whose multimodal behavior matches the flexibility and coordination that distinguishes our own embodied meanings. More generally, modeling replaces programming with data fitting, and a good model of action and interpretation in particular would let an agent's own experience in conversational interaction determine the repertoire of behaviors and meanings it uses to make itself understood.

Acknowledgments

To colleagues and coauthors, especially David DeVault and the organizers of INLG 2008, and to NSF IGERT 0549115, CCF 0541185 and HSD 0624191.

References

J. Cassell, M. Stone, and H. Yan. 2000. Coordination and context-dependence in the generation of embodied conversation. In *INLG*, pages 171–178.

D. DeCarlo, M. Stone, C. Revilla, and J. J. Venditti. 2004. Specifying and animating facial signals for discourse in embodied conversational agents. *Computer Animation and Virtual Worlds*, 15(1):27–38.

D. DeVault and M. Stone. 2004. Interpreting vague utterances in context. In *COLING*, pages 1247–1253.

D. DeVault and M. Stone. 2007. Managing ambiguities across utterances in dialogue. In *DECALOG: Workshop on the Semantics and Pragmatics of Dialogue*.

A. Koller and M. Stone. 2007. Sentence generation as a planning problem. In *ACL*, pages 336–343.

A. Lascarides and M. Stone. 2008. Discourse coherence and gesture interpretation. *Ms, Edinburgh–Rutgers*.

A. Lascarides and M. Stone. In Preparation. Grounding and gesture. *Ms, Edinburgh–Rutgers*.

M. Stone and D. DeCarlo. 2003. Crafting the illusion of meaning: Template-based generation of embodied conversational behavior. In *Computer Animation and Social Agents (CASA)*, pages 11–16.

M. Stone and I. Oh. 2008. Modeling facial expression of uncertainty in conversational animation. In I. Wachsmuth and G. Knoblich, editors, *Modeling Communication with Robots and Virtual Humans*, pages 57–76. Springer.

M. Stone, C. Doran, B. Webber, T. Bleam, and M. Palmer. 2003. Microplanning with communicative intentions: The SPUD system. *Computational Intelligence*, 19(4):311–381.

M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler. 2004. Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics*, 23(3):506–513.

R. Thomason, M. Stone, and D. DeVault. 2006. Enlightened update: a computational architecture for presupposition accommodation. *Ms, Michigan–Rutgers*.