The Challenge of Modeling Dialog Dynamics

Nigel G. Ward
Department of Computer Science
University of Texas at El Paso
El Paso, Texas 79968
nigelward@acm.org

Abstract

This paper advocates the development of general models of dialog dynamics: across observables such as prosody, gesture, facial expression, proxemics, posture and gaze; across functions such as turn-taking, meta-communication, interpersonal relation management, and affect expression; and largely independently of semantic content. The paper also suggests a way to frame dialog dynamics as a machine learning problem, and surveys some issues and complications.

1 Introduction

It is often noted that most dialog systems today use a limited, rather robotic interaction style, and that instead they should be able to respond with subtly appropriate expressions, gesture, tone of voice and so on, adapted moment by moment to be effective and appropriate for the user's state, as revealed by his expression, gesture, speech and so on. Humans often seem able to do this effortlessly. For example, Figure 1 shows a conversation fragment taken from a corpus of dialogs between a university staff member and lower-division students about career paths [1]. Consider utterance C3, *yeah*. Although almost gratuitous from the perspective of the information conveyed, at this point in the conversation it sounds completely natural and appropriate, at many levels. (This is clearer from the audio, which is available at http://www.cs.utep.edu/nigel/abstracts/jaime-specom.html). On the valence level, the student having expressed a slightly positive statement about the TAs, C appropriately echoes that feeling. On the activation level, the student seeming to be losing interest, C shows that this topic is, for her, something important to talk about. On the power dimension, the student is apparently trying to be non-committal and uninvolved, but C's response leads the student to elaborate, while at the same time establishing dominance with respect to who will guide the conversation. In terms of turn-taking, C apparently intends to say little herself about this topic, and is starting the process of closing it out with a few short turns before transitioning to the main topic.

Thus this response is appropriate in many ways. To generate such responses requires the close tracking of many aspects of the interlocutor's state, as revealed by features of prosody and gaze, among others, some of which are fairly subtle. It requires in addition a model of some of the recurrent patterns of actions and interactions by and between dialog participants, moment-by-moment. Detailing these patterns and their underlying mechanisms is a matter of considerable scientific interest [1, 2].

Discovering these patterns of behaviors is also of practical importance, and some recent work has shown how this can be done. For example, Morency [3] showed that the decision of when to produce a back-channel (*uh-huh*, *mm-hmm*, etc.) can be made fairly well based on features of the prosody and gaze of the interlocutor, and Raux and Eskenazi [4] showed similarly how to decide when to initiate a new turn. Others have shown how turn-by-turn responsiveness in attitudinal and emotional dimensions can be accomplished by paying attention to prosodic features in the immediately prior context [1, 5]. The results of experiments with these systems show how such response patterns can make dialogs more effective and more satisfying for users, even when the content of the interactions is not changed.

	Transcription	Emotion	Prosodic Properties
C0	So you're in the 1401 class?	act: 35, val: 10, pow: 35	normal speed, articulating word beginnings
S1	*Yeah.*	*act: 10, val: 5, pow: –5*	*higher pitch*
C1	Yeah? How are you liking it so far?	act: 40, val: 10, pow: 35	unchanged
S2	*Um, it's alright, it's just the labs are kind of difficult sometimes, they can, they give like long stuff.*	*act: 5, val: –10, pow: –15*	*slower speed, falling pitch*
C2	Mm. Are the TAs helping you?	act: 20, val: –10, pow: 10	lower pitch, slower speed
S3	*Yeah.*	*act: 5, val: 5, pow: –15*	*rising pitch*
C3	Yeah.	act: 20, val: 5, pow: –15	rising pitch
S4	*They're doing a good job.*	*act: 10, val: 0, pow: 5*	*normal speed, normal pitch*
C4	Good, that's good, that's good.	act: 35, val: 10, pow: 40	normal pitch, normal speed

Figure 1: Dialog fragment. The emotional values are for activation (involvement), valence (positive or negative feeling), and power (dominance), each on a scale from –100 to +100.

Despite these successes, modeling and exploiting aspects of dialog dynamics has so far been accomplished only for certain specific types of response pattern, and only after labor-intensive analysis and development. Thus we need a general model. By analogy to the field of physics, where dynamics refers to the study of the ways in which physical systems change over time and the causes of those changes, I see the need for a model of "dialog dynamics," explicating the patterns of occurrence of nonverbal dialog phenomena — specifically observables such as prosody, gesture, gaze, proxemics, posture, and facial expression — and the ways in which the interlocutors' dialog states interact and change over time.

2 Framework

The construction of such a model could be based on four working assumptions.

First, it makes sense to work on dialog dynamics in general, rather than to build individual models of the various observables. This is because these generally pattern together and supplement or complement each other, and because all these seem to relate to the same broad set of interrelated communicative functions: turn-taking, including channel control and meta-communication; negotiation and flagging of information status, including recognition, comprehension, grounding, agreement and the interestingness and newness of information; the management of interpersonal relations such as control and affiliation; and the expression of emotion, attitude, and affect.

Second, dialog dynamics can be modeled independently of the propositional content conveyed (but see below). Often dialog seems to have its own "momentum;" and many meaning-independent patterns can be found, as in the systems mentioned above. Further support is seen in the existence of dialogs which, although fairly content-free or treated by one speaker as such, have value to the participants [6]; and from the existence of speakers who are good at rapport but not content, or the opposite, with autism as an extreme case. While leaving on propositional content thus seems like a good initial approach, some immediately accessible information from the lexical items can and should be factored in, for example the affective value of the words spoken.

Third, a model of dialog dynamics should include a learning component. As the patterns of dialog behavior depend on culture, language, situation, and personalities, it seems appropriate to develop a general model with a learning algorithm able to acquire, from a specific body of dialog data, the patterns and parameters of responsiveness in that genre.

Fourth, the primary task for a model of dialog dynamics should be prediction. Given what has happened in a dialog up to time t, if a model can accurately predict the interlocutors' immediately upcoming behaviors, then that model is perforce a good model. Of course, people have free will, so such predictions will only ever be probabilistically correct. Focusing on the prediction problem has several merits: it gives a clear evaluation metric, it is analogous to one of the primary tasks that human dialog participants accomplish, it is useful for dialog systems builders [7], and it is also relevant for tasks in behavior recognition and behavior synthesis. The process of working to improve the predictive power of models will drive the field in the direction of a deeper understanding of dialog dynamics, as it is highly unlikely that a completely general, structure-free, *tabula rasa* model could perform well.

3 Challenges

Building such a model of dialog dynamics will be challenging, for many reasons. This section lists a few, based largely on my own experiences in modeling dialog processes involving non-lexical utterances, gesture, and especially prosody (pitch, energy, and timing) [8, 9, 10, 11, 12, 7, 1, 2].

One set of issues relates to the input and output features for a model of dialog dynamics. First, they are inherently heterogeneous, for example, some are properties of regions of time, such as speech rate, whereas some are more instantaneous, such as the volume at a point in time. Second, some are not always defined (such as pitch, which has no value in periods of silence or unvoiced consonants). Third, many are continuous, not binary; for example a region of low-pitch, functioning as a cue to back-channels, tends to be a stronger cue to the extent that it is lower in pitch or longer in duration. Fourth, features at different "levels" seem to be involved: low-level features such as pitch, "mid-level features" such as the low-pitch cue mentioned above, and higher-level features, the actual dialog-relevant signals, which often are made up of configurations of several mid-level features co-occurring in some way, with semi-flexible temporal constraints on how the component features may appear relative to each other, although the ways in which features are synchronized or semi-synchronized within and among input streams are currently not well understood. Fifth, the relevant mid-level features may be hard to discover automatically, in particular because of the abundance of possibilities. Even simple ones, such as average pitch, pitch slope, maximum pitch, and pitch range, can be computed over various intervals, creating a multitude of possible features. Beyond that, arbitrarily complex features could be involved, such as the number of pitch peaks over the past 500ms, height of highest pitch peak in the last 400ms relative to the baseline computed over the past 2000ms, first coefficient of a second-order approximation to the pitch curve over the last three syllables before a pause of at least 200ms, and so on, where all the feature-defining parameters can range over many values. The abundance becomes breathtaking when one includes features computed from other dimensions, such as energy, voicing type, and gaze.

Another set of issues relates to the mediating variables: the cognitive states and processes that underlie these signals. (To date, most successful models of patterns in dialog dynamics have treated them as largely reflex-like, explicable by identifying surface features in the immediate discourse context that cue surface responses, but this will only take us so far.) First, the relevant cognitive state is complex, and in particular seldom binary-valued; for example, rather than simply intending to speak or not, a speaker may intend with some degree of intensity to speak. Second, intentions may have temporal extent, for example, an interlocutor may be planning to speak at (or up until) some approximate future time. Third, changes in these states are more often continuous than discrete, increasing or decaying at certain rates in the absence of inputs or actions, as a result of mental processes. Fourth, there are many dimensions of relevant state, which probably affect each other in complex ways. For example, turn-taking intentions relate to the information-processing state, to the emotional value of responses, and to interpersonal dimensions.

While the best strategy may be to first model dialog dynamics independently of content, any learning algorithm must be robust to "noise" arising from the lack of true independence. In particular, connections between an observed signal and its dialog significance may be inconsistent and may occur with varying time lags. Noise will also come from the multi-functional nature of each of the signals of interest; prosody, gaze, gesture and so on not only convey dialog-related functions, but also have other uses and other meanings.

4 Prospects

While these challenges are substantial, none seem impossible to overcome, and machine learning research can rise to the challenge. We can take inspiration from the success of Hidden Markov Models, where the development of the associated algorithms was inspired by notions about the nature of human speech, developed by linguists and phoneticians over the centuries, such as that speech is a sequence of phonemes, that phoneme subsequences map to words, and that phonemes map to spectral patterns. Although these are all simplifications of reality, the models they inspired have proven enormously useful, not only as the backbone of all practical speech recognition systems, but for many other problems.

While our understanding of the nature of dialog dynamics today is more tentative (and experimental investigations are certainly needed) some aspects seem reasonably well established, for example: various relevant features and cues can be detected from the input, these are semi-synchronized, configurations of features bear meaning, interlocutors update their internal state based on the information obtained from such features, changes in an interlocutor's internal state happen semi-autonomously according to various processes with various time constants, and these internal states affect the subsequent non-lexical behavior of the speaker. Development of a general, trainable model based on these notions could be of great scientific and practical value, for modeling dialog dynamics and possibly also for other problems.

Acknowledgments

I thank the NSF for support (IIS-0914868) and Olac Fuentes, David Novick, Louis-Philippe Morency and Daniel Gatica-Perez for comments.

References

[1] Jaime C. Acosta and Nigel G. Ward. Achieving rapport with turn-by-turn, user-responsive emotional coloring. *Speech Communication*, 2010. to appear.

[2] Nigel G. Ward, Alejandro Vega, and Timo Baumann. Prosodic and temporal features for language modeling for dialog. *Speech Communication*, 2010. submitted.

[3] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20:70–84, 2010.

[4] Antoine Raux and Maxine Eskenazi. A finite-state turn-taking model for spoken dialog systems. In *NAACL HLT*, 2009.

[5] Nigel G. Ward and Rafael Escalante-Ruiz. Using subtle prosodic variation to acknowledge the user's current state. In *Interspeech*, pages 2431–2434, 2009.

[6] Victor Yngve. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–577, 1970.

[7] Nigel G. Ward, Olac Fuentes, and Alejandro Vega. Dialog prediction for a general model of turn-taking. In *Interspeech*, 2010.

[8] Nigel Ward and Takeshi Kuroda. Requirements for a socially aware free-standing agent. In *Proceedings of the Second International Symposium on Humanoid Robots*, pages 108–114, 1999.

[9] Nigel Ward and Yaffa Al Bayyari. A case study in the identification of prosodic cues to turn-taking: Back-channeling in Arabic. In *Interspeech 2006 Proceedings*, 2006.

[10] Nigel Ward. Non-lexical conversational sounds in American English. *Pragmatics and Cognition*, 14:113–184, 2006.

[11] Nigel G. Ward and S. Kumar Mamidipally. Factors affecting speaking-rate adaptation in task-oriented dialogs. In *Speech Prosody*, 2008.

[12] Nigel G. Ward and Joshua L. McCartney. Visualization to support the discovery of prosodic contours related to turn-taking. Technical Report UTEP-CS-10-24, University of Texas at El Paso, 2010.