Ideas on Multi-layer Dialogue Management for Multi-party, Multi-conversation, Multi-modal Communication

Extended Abstract of Invited Talk

David R Traum

University of Southern California, Institute for Creative Technologies

1 Overview

Most current dialogue systems concern only a short dialogue between a single system and single user, focused on a single task. On the other hand, the full spectrum of communication between interacting agents includes cases in which multiple segments of conversation can be interleaved with other, sometimes unrelated actions and events (e.g., a cocktail party). Language use in the Mission Rehearsal Exercise Project at ICT (Swartout, Hill, Gratch, Johnson, Kyriakakis, Labore, Lindheim, Marsella, Miraglia, Moore, Morie, Rickel, Thiebaux, Tuch, Whitney and Douglas 2001) falls between these two extremes, having one main purpose (Army platoon-level leadership training using virtual reality and virtual humans), but multiple characters, each with its own goals, interests, and capabilities. In this scenario, multiple characters must engage in dialogue, both with each other and with the human trainee. Moreover, multiple conversations are involved, each with a distinct context for interpretation. The conversations are also multimodal in two senses. First, communication can occur not just with speech, but also with visual media including gesture and gaze of artificial characters, and secondly different media sets must be used for different communications. E.g., face-to-face communication for some characters and radio communication for others who are not physically co-present.

We present here a multiple layer approach towards modelling and managing these complexities, including who is accessible for conversation, paying attention, involved in a conversation, as well as turn-taking, initiative, grounding, and higher-level dialogue functions. The method will follow that used in the Trindi project, where one specifies an information state, and ”dialogue moves” representing input and output, as well as associated updates to information state.

2 Context: The Mission Rehearsal Exercise Project

The test bed for our dialogue model is the Mission Rehearsal Exercise project at the University of Southern California’s Institute for Creative Technologies. The project is exploring the integration of high-end virtual reality with Hollywood storytelling techniques to create engaging, memorable training experiences. The setting for the project is a virtual reality theatre, including a visual scene projected onto an 8 foot tall screen that wraps around the viewer in a 150 degree arc (12 foot radius). Immersive audio software provides multiple tracks of spatialized sounds, played through ten speakers located around the user and two subwoofers. Within this setting, a virtual environment has been constructed representing a small village

in Bosnia, complete with buildings, vehicles, and virtual characters. This environment provides an opportunity for Army personnel to gain experience in handling peacekeeping situations.

The first prototype implementation of a training scenario within this environment was completed in September 2000 (Swartout et al. 2001). To guide the development, a Hollywood writer, in consultation with Army training experts, created a script providing an overall story line and representative interactions between a human user (Army lieutenant) and the virtual characters. In the scenario, the lieutenant finds himself in the passenger seat of a simulated Army vehicle speeding towards the Bosnian village to help a platoon in trouble. Suddenly, he rounds a corner to find that one of his platoon's vehicles has crashed into a civilian vehicle, injuring a local boy The boy's mother and an Army medic are hunched over him, and a sergeant approaches the lieutenant to brief him on the situation. Urgent radio calls from the other platoon, as well as occasional explosions and weapons fire from that direction, suggest that the lieutenant send his troops to help them. Emotional pleas from the boy's mother, as well as a grim assessment by the medic that the boy needs a medevac immediately, suggest that the lieutenant instead use his troops to secure a landing zone for the medevac helicopter. Figure 1 shows a small excerpt from the original script.

The interaction in Figure 1 illustrates a number of issues that arise for embodied agents, some going beyond capabilities of current implemented systems. First, at a broad level, the agents must concern themselves with multiple characters and multiple conversations. The main conversation is between the lieutenant and sergeant, but the medic is also brought in, and the mother is an important overhearer. Other platoon members and townspeople may also be potential overhearers. There is also a separate conversation between the sergeant and the squad leaders, at the end of the excerpt given here. In other parts of the scenario, the lieutenant engages in radio conversations with his home base, another platoon, and sometimes a medevac helicopter. Some of these conversations have explicit beginning and ending points (especially the radio conversations), while others, which are more focused on specific tasks, end without remark as the local purpose of the interaction is established and resolved and attention of the conversants shifts to other matters. In all cases, agents must reason about who is speaking, who is being spoken to, who is listening, and whether they need to speak to another agent.

In this immersive virtual world, the agents must also coordinate speech with other communicative modalities. In many cases, gestures and other nonverbal cues are important in carrying some of the communicative function. Some examples here are the way the lieutenant approaches the sergeant to initiate conversation, the way that the sergeant glances at the medic to signal that he should take the turn and respond to the lieutenant's question, and the way the medic glances at the mother while formulating a less direct answer about the boy's health — focusing on the consequence of his condition rather than directly stating what might be upsetting to her.

actor	communication
LT	(Drive up, exit vehicle, approach SGT)
SGT	(Look at LT)
LT	Sergeant, what happened here?
SGT	They just shot out from the side street sir. (Gesturing towards the civilian vehicle) The driver couldn't see 'em coming.
LT	How many people are hurt?
SGT	The boy and one of our drivers. (Gesturing toward the boy)
LT	Are the injuries serious?
SGT	(Makes eye contact with medic and nods)
MEDIC	Driver's got a cracked rib but the kid's – (Glancing at the mother) Sir, we gotta get a medevac in here ASAP.
LT	We'll get it.
LT	Platoon Sergeant, secure the area.
SGT	Yes Sir!
SGT	(Shouting) Squad leaders, listen up! (Raises arm, looks around at squad leaders) I want 360 degree security here now! First squad 12-4 (Looks at 1st squad leader and gestures) Second squad 4-8 (Looks at 2nd squad leader and gestures) Third squad 8-12 (Looks at 3rd squad leader and gestures)
Leaders	(move squads into place)

Figure 1: Example of MRE Multi-modal, multi-character interaction.

3 Approach: Multi-layer dialogue management

We are designing agents that are capable of engaging in dialogue such as that exemplified in the previous section (Rickel, Marsella, Gratch, Hill, Traum and Swartout 2002). This section summarizes the dialogue modelling and dialogue management aspects that support such complex interactions.

The dialogue management component of a spoken dialogue system (or artificial agent) is concerned with several key tasks:

- Updating the dialogue context as new contributions to the dialogue are observed,
- Deciding when to speak, and what to say
- Interfacing with the back-end or task model, interpreting general language utterances within the context of the abilities of the system, and converting natural language utterances into commands that a back-end system can understand.
- Provide expectations for interpretation given the current dialogue context.

We follow Trindi project approach to dialogue manangement (Larsson and Traum 2000). The part of the context deemed relevant for dialogue modelling, termed *information state*, is maintained as a snapshot of the dialogue state. This state is then updated by dialogue moves, seen as abstract input and output descriptions for the dialogue modeling component. There are *update rules* which govern how the information state is modified, both for observation of dialogue moves and other dialogue inference. There are also *selection rules* that allow the system to choose dialogue moves to perform given a configuration of information state.

Depending on the type of dialogue and theory of dialogue processing, many different views of the specifics of information state and dialogue moves are possible. A complex environment such as the MRE situation presented in the previous section obviously requires a fairly elaborate information state to achieve fairly general performance within such a domain. We try to manage this complexity by partitioning the information state and dialogue moves into a set of *layers*[1], each dealing with a coherent aspect of dialogue that is somewhat distinct from other aspects.

Each layer is defined by information state components, a set of relevant dialogue acts, and then several classes of rules relating the two and enabling dialogue performance:

recognition rules that decide when acts have been performed, given observations of language and non-linguistic behavior in combination with the current information state

[1] The term *layer* is used informally, avoiding the issue of technical differences between ranks and levels (Halliday 1961), in a manner similar to (Allen and Core Draft, 1997)

- contact
- attention
- conversation
  - participants
  - turn
  - initiative
  - grounding
  - topic
  - rhetorical
- social commitments (obligations)
- negotiation

Figure 2: Multi-party, Multi-conversation Dialogue Layers

update rules that modify the information state components with information from the inferred dialogue acts

selection rules that decide which dialogue acts the system should perform

realization rules that indicate how to perform the dialogue acts by some combination of linguistic expression (e.g., natural language generation), non-verbal communication, and other behavior.

The layers used in the current system are summarized in Figure 2. The *contact* layer (Allwood, Nivre and Ahlsen 1992, Clark 1996, Dillenbourg, Traum and Schneider 1996) concerns whether and how other individuals can be accessible for communication. Modalities include visual, voice (shout, normal, whisper), and radio. The *attention* layer concerns the object or process that agents attend to (Novick 1988). Contact is a prerequisite for attention. The *Conversation* layer models the separate dialogue episodes that go on during an interaction. A conversation is a reified process entity, consisting of a number of sub-layers. Each of these layers may have a different information content conversations happening at the same time. The *participants* may be active speakers, addressees, or overhearers (Clark 1996). The *turn* indicates the (active) participant with the right to communicate (using the primary channel) (Novick 1988, Traum and Hinkelman 1992). The *initiative* indicates the participant who is controlling the direction of the conversation (Walker and Whittaker 1990). The *grounding* component of a conversation tracks how information is added to the common ground of the participants (Traum 1994). The conversation structure also includes a *topic* that governs relevance, and *rhetorical* connections between individual content units. Once material is grounded, even as it still relates to the topic and rhetorical structure of an ongoing conversation, it is also added to the social fabric linking agents, which is

not part of any individual conversation. This includes *social commitments* — both obligations to act or restrictions on action, as well as commitments to factual information (Traum and Allen 1994, Matheson, Poesio and Traum 2000). There is also a *negotiation* layer, modeling how agents come to agree on these commitments (Baker 1994, Sidner 1994). More details on these layers, with a focus on how the acts can be realized using verbal and non-verbal means, can be found in (Traum and Rickel 2002).

4 Dialogue Processing in MRE

We have built a preliminary implementation of a dialogue manager using the dialogue layers described in the previous section to allow artificial agents to communicate in the domain described in Section 2. This is built on top of the Steve agent (Johnson and Rickel 1998), making use of the existing task-model and action selection mechanisms. Language processing occurs in two distinct and interleavable "cycles", one for understanding language and updating the information state, and a second for producing language. This separation of input and output processing cycles allows the agent to have an arbitrary interleaving of contributions by itself and others rather than enforcing a rigid turn-alternation. Each communitive contribution is simultaneously interpreted at each layer, and may correspond to a number of acts at different layers. Generation usually starts from an intention to perform a main act, however any realized utterance will also correspond to a number of acts, some of which (e.g., turn-taking) may be as much a result of the timing of the performance with respect to other events as to the planned behavior.

Acknowledgements

The author would like to thank members of the MRE project for interesting discussions about the complexities of interaction in this scenario, as well as our many colleagues involved in building the artificial world, simulation environment, and agents. Larry Tuch wrote the script with creative input from Richard Lindheim and technical input on Army procedures from Elke Hutto and General Pat O'Neal. The work described in this paper was supported by the Department of the Army under contract number DAAD 19-99-D-0046. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Department of the Army.

References

Allen, J. and Core, M.(Draft, 1997), Draft of DAMSL: dialog act markup in several layers, available through the WWW at: http://www.cs.rochester.edu/research/trains/annotation.

Allwood, J., Nivre, J. and Ahlsen, E.(1992), On the semantics and pragmatics of linguistic feedback, *Journal of Semantics*.

Baker, M.(1994), A model for negotiation in teaching-learning dialogues, *Journal of Artificial Intelligence in Education* **5**(2), 199–254.

Clark, H. H.(1996), *Using Language*, Cambridge University Press, Cambridge, England.

Dillenbourg, P., Traum, D. and Schneider, D.(1996), Grounding in multi-modal task-oriented collaboration, *Proceedings of the European Conference on AI in Education*.

Halliday, M. A. K.(1961), Categories of the theory of grammar, *Word* **17**, 241–92.

Johnson, W. L. and Rickel, J.(1998), Steve: An animated pedagogical agent for procedural training in virtual environments, *SIGART Bulletin* **8**, 16–21.

Larsson, S. and Traum, D.(2000), Information state and dialogue management in the TRINDI dialogue move engine toolkit, *Natural Language Engineering* **6**, 323–340.

Matheson, C., Poesio, M. and Traum, D.(2000), Modelling grounding and discourse obligations using update rules, *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics*.

Novick, D.(1988), *Control of Mixed-Initiative Discourse Through Meta-Locutionary Acts: A Computational Model*, PhD thesis, University of Oregon. also available as U. Oregon Computer and Information Science Tech Report CIS-TR-88-18.

Rickel, J., Marsella, S., Gratch, J., Hill, R., Traum, D. and Swartout, W.(2002), Toward a new generation of virtual humans for interactive experiences, *IEEE Intelligent Systems*.

Sidner, C. L.(1994), An artificial discourse language for collaborative negotiation, *Proceedings of the Fourteenth National Conference of the American Association for Artificial Intelligence (AAAI-94)*, pp. 814–819.

Swartout, W., Hill, R., Gratch, J., Johnson, W., Kyriakakis, C., Labore, K., Lindheim, R., Marsella, S., Miraglia, D., Moore, B., Morie, J., Rickel, J., Thiebaux, M., Tuch, L., Whitney, R. and Douglas, J.(2001), Toward the holodeck: Integrating graphics, sound, character and story, *Proceedings of 5th International Conference on Autonomous Agents*.

Traum, D. R.(1994), *A Computational Theory of Grounding in Natural Language Conversation*, PhD thesis, Department of Computer Science, University of Rochester. Also available as TR 545, Department of Computer Science, University of Rochester.

Traum, D. R. and Allen, J. F.(1994), Discourse obligations in dialogue processing, *Proceedings of the 32^{nd} Annual Meeting of the Association for Computational Linguistics*, pp. 1–8.

Traum, D. R. and Hinkelman, E. A.(1992), Conversation acts in task-oriented spoken dialogue, *Computational Intelligence* **8**(3), 575–599.

Traum, D. R. and Rickel, J.(2002), Embodied agents for multi-party dialogue in immersive virtual worlds, *to appear in Proceedings of the first International Joint conference on Autonomous Agents and Multiagent systems*.

Walker, M. A. and Whittaker, S.(1990), Mixed initiative in dialogue: An investigation into discourse segmentation, *Proceedings ACL-90*, pp. 70–78.