Situated Conversational Agents

Will Thompson
Northwestern University
Department of Linguistics
2016 Sheridan Road
Evanston, Illinois 60208
Phone: (847) 864-6858
Email: will.k.t@gmail.com

Abstract

A *Situated Conversational Agent* (SCA) is an agent that engages in dialog about the context within which it is embedded. Situated dialog is characterized by its deep connection to the embedding context, and the precise cross-timing of linguistic and non-linguistic actions. This paper describes initial research into the construction of an SCA that engages in dialog about *collaborative physical tasks*, in which agents engage in dialog with the joint goal of manipulating the physical context in some manner. Constructing an SCA that can interact naturally in such tasks requires an agent with the ability to interleave planning, action, and observation while operating in a partially observable environment. Consequently, I propose to model an SCA as a *Partially Observable Markov Decision Process* (POMDP).

Introduction

A *Situated Conversational Agent* (SCA) is an agent that engages in dialog about the context within which it is embedded. An SCA is distinguished from non-situated conversational agents by an emphasis on the intimate connection of the agent's dialog to its embedding context, and the precisely timed interleaving of linguistic and non-linguistic actions. Situated dialogs can potentially occur in a wide variety of contexts, from a virtually embodied agent in a virtual world, to a physically embodied mobile robot.

This paper focuses on research into the construction of an SCA that engages in dialog in support of *collaborative physical tasks* (Gergle, Kraut, & Fussell 2004), in which the SCA collaborates with a user in order to modify the physical context. This research has two goals. First, it is intended to formalize the claims of psycholinguistic theories that have been proposed to account for properties of such dialogs. Second, an SCA capable of interacting with users in a natural fashion will have a practical benefit for the construction of interfaces to devices that engage in collaborative physical tasks.

Properties of Situated Dialog

Psycholinguistic experiments have revealed a number of interesting properties of situated dialog. For example, Gergle, Kraut, & Fussell (2004) describe an experiment in which pairs of subjects worked together online to solve a virtual jigsaw puzzle, as shown in Figure 1. In this task, the subjects were given the roles of "Helper" and "Worker", with the Helper providing verbal instructions on how to build the puzzle, while the Worker built the puzzle. In a subset of trials the Helper shared a view of the Worker's workspace. In other trials, the Helper had no view of the Worker's workspace, and the interaction was restricted to audio communication. There were marked differences between these two conditions, with shared visual workspace dialogs having the following properties:

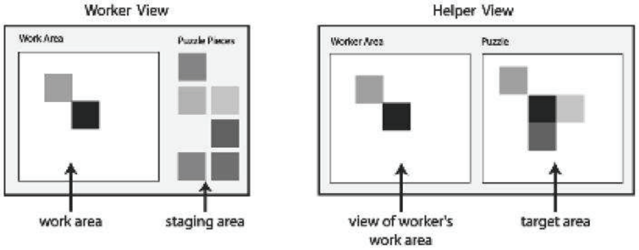

Figure 1: From (Gergle, Kraut, & Fussell 2004)

P1 *Actions substitute for language.* In a collaborative physical task, worker actions have the dual role of communicating information as well as accomplishing task goals (Gergle, Kraut, & Fussell 2004). Actions can even substitute for questions, when a participant initiates an action and then pauses before completion, prompting feedback from another participant (Clark & Krych 2004).

P2 *Actions and language are interleaved in fine increments.* Situated dialog often consist of short speech increments, finely timed with others' actions. The pauses between the speech increments are precisely timed to correspond with the Worker's actions, and these actions can in turn determine the subsequent course of the dialog (Clark & Krych 2004). The participants work together in order to *ground* dialog contribution at a fine time-scale, with dialog often consisting of repeated productions of short sentential fragments, followed by verbal or non-verbal responses.

P3 *Dialog structure is simplified.* A shared visual workspace provides both agents with high fidelity observations about the state of the task, and the state of the ongoing actions of the participants. Agents who are aware their actions are being observed produce fewer explicit reports of their current state of understanding. Agents that can observe others' actions produce fewer explicit requests for grounding feedback (Gergle, Kraut, & Fussell 2004).

Modeling Situated Dialog

Constructing an SCA that can model these three properties requires an agent that can interleave planning, action, and observation. Furthermore, the agent has to model the other participants in the conversation without having direct access to their mental state. Dialog contributions might fail to have their desired effect, due to channel noise, ambiguity, or divergent beliefs about the current state. Therefore, an SCA is planning and acting under uncertainty, making situated dialog a suitable candidate for the use of *Partially Observable Markov Decision Processes* (POMDPs) (Kaelbling, Littman, & Cassandra 1998). POMDPs have recently been applied to the construction of spoken dialog systems (Williams & Young 2007). I propose extending this approach to the case of situated dialog, in order to capture properties P1 - P3.

A POMDP is formally defined in (Kaelbling, Littman, & Cassandra 1998) as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \Omega, \mathcal{O} \rangle$ where $\mathcal{S}$ is a finite set of states of the world, $\mathcal{A}$ is a finite set of actions, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \Pi(\mathcal{S})$ is the state-transition function, giving for each world state and agent action a probability distribution over world states, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function giving the expected immediate reward gained by the agent for taking each action in each state, Ω is a finite set of observations of the world, and $\mathcal{O} : \mathcal{S} \times \mathcal{A} \rightarrow \Pi(\Omega)$ is the *observation function*, giving for each action and resulting state, a probability distribution over possible observations.

Williams & Young (2007) specialize this general framework by constructing a factored architecture of a POMDP-based dialog manager, with state space consisting of a tuple $\langle S_u, S_d, A_u \rangle$, representing the user's goals, the dialog history, and the user's actions, respectively. In their framework, user and system actions are represented as sets of speech acts, and observations consist of speech recognition results.

In order to extend this framework to an SCA, this representation needs to be augmented, minimally with a model of the physical task state, as well as the agent's sensors and actuators with respect to this state. The simplest possible extension is a factored representation where $S = \langle S_u, S_d, A_u, Task \rangle$, $A = \{c_1, ..., c_m, p_1, ..., p_n\}$, $Task = \{t_1, ..., t_o\}$, and $\Omega = \{\overline{c}_1, ..., \overline{c}_{m'}, \overline{p}_1, ..., \overline{p}_{n'}, \overline{t}_1, ..., \overline{t}_{o'}\}$. With this representation, state space has an additional task parameter, consisting of the set of possible task states, the action set consists of a set of communicative and physical actions, and the observation set contains observations of the user communicative and physical actions, as well as task state. The resulting SCA-POMDP gives us the means for modeling properties P1 - P3.

Considering P1, an optimal policy for the SCA-POMDP will take into account a value of information calculation, directing the SCA to choose either a physical or a communicative action based on progress towards goal completion as well as on the basis of how much uncertainty it eliminates. The SCA-POMDP therefore has a principled means for deciding when it is better to act, or to speak. A proper formulation of the reward function will take into account the relative benefits of communicating versus acting, guiding the agent to choose a speaking action only when doing so provides a long-term reward that outweighs the immediate cost of the action. If physical actions are relatively cheap, and can communicate the necessary information by themselves, then an SCA-POMDP will generate fewer dialog contributions.

Modeling P2 requires the action set of the SCA-POMDP to contain a fine-grained set of speech acts, mapping on to functionally useful task actions. Likewise, there must be a set of fine-grained physical actions that these speech acts map to. For example, in the puzzle task shown in Figure 1, there must be speech acts for identifying individual puzzle pieces and for identifying regions of the workspace, and there must be physical actions for selecting a puzzle piece and moving it to a particular region of the workspace. Once such actions are available to the SCA, its reward function must be constructed in order to bias the POMDP towards *early grounding*, causing the system to prefer actions that modify the dialog and task state in the small increments that correspond to the effects of the fine-grained actions.

Handling P3 requires a sufficient model of the user's observational abilities. The SCA-POMDP captures this aspect of situated dialog with (1) its dialog state component S_d, and (2) the state transition function. When the SCA-POMDP shares a visual workspace with the user, the transition function will capture this fact by updating the dialog state more-or-less deterministically after a physical action. Since the visual workspace provides a high fidelity view of the system's action, the dialog state will reflect the result of the system's action with a high degree of certainty after a physical action is performed. Conseqently, in these situations the SCA-POMDP will tend to generate fewer grounding acts.

Once completed, the POMDP implementation of an SCA described here will be evaluated with human subjects using a variant of the puzzle task described in (Gergle, Kraut, & Fussell 2004).

References

Clark, H. H., and Krych, M. A. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language* 50(1):62–81.

Gergle, D.; Kraut, R.; and Fussell, S. 2004. Language efficiency and visual technology: Minimizing collaborative effort with visual information. *Journal of Language and Social Psychology* 23:491–517.

Kaelbling, L.; Littman, M.; and Cassandra, A. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101:99–134.

Williams, J., and Young, S. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language* 21:393–422.