An Explainable Collaborative Dialogue System using a Theory of Mind

Philip R. Cohen, Lucian Galescu, Maayan Shvo

Openstream.ai

{phil.cohen, lucian, maayan}@openstream.com

Abstract

Eva is a neuro-symbolic domain-independent multimodal collaborative dialogue system that takes seriously that the purpose of task-oriented dialogue is to assist the user. To do this, the system collaborates by inferring their intentions and plans, detects obstacles to success, finds plans to overcome them or to achieve higher-level goals, and plans its actions, including speech acts, to help users accomplish those goals. In doing so, the system maintains and reasons with its own declaratively-specified beliefs, goals and intentions, and explicitly reasons about those of its user. Because Eva can track different users' mental states, it can engage multiple agents in multi-party dialogues. Reasoning is accomplished with a modal Horn-clause meta-interpreter that enables computable inference within the subset of logic implemented. The system employs both hierarchical and backward-chaining planning, operating over a rich modal logic-based knowledge and action representation. The planning and reasoning subsystems obey the principles of persistent goals and intentions including: 1) The formation and decomposition of intentions to perform complex actions, 2) the conditions under which persistent goals and intentions can be given up, and 3) persistent goal and intention revision using the relativizing formulas that are created during the planning process. The system treats its speech acts just like its other actions. This general approach enables Eva to plan a variety of speech acts, including requests, informs, questions, confirmations, offers, acceptances, and emotive expressions. Because the dialogue engine is a planner, as the dialogue proceeds, the system can flexibly generate, execute, and potentially repair its plans using physical, digital, and speech actions. Importantly, Eva can explain its utterances because it has created a plan that caused it to utter them.[1]

1 Introduction

In this paper we describe Eva, a fully-functional neuro-symbolic domain-independent collaborative dialogue system that takes seriously the tenet that the purpose of task-oriented dialogue is to *assist* the user. Eva attempts to collaborate with its users by inferring and debugging their plans, then planning to overcome obstacles to achieving their higher-level goals. In order to do so, Eva represents and reasons with beliefs, goals and intentions ("BDI")[2] of the user and the system itself. Because the dialogue engine is a planner, as the dialogue proceeds, the system is able to go beyond scripted, slot-filling, or finite state dialogue behavior to flexibly generate, execute, and potentially repair its plans using both non-communicative actions and speech-acts. As part of its reasoning, Eva performs plan/goal recognition on the user's mental state. Importantly, the system itself decides what to say, not the developer, by obeying the well-studied principles of persistent goals and intentions (see Cohen and Levesque [1]). Importantly, thanks to the BDI underlying machinery, Eva is able to explain its actions and its plans, thus achieving more trustworthy interactions.

A useful and meaningful dialog system in a rich task-oriented natural language conversational setting must be collaborative. Indeed, collaboration is so essential to society that we teach our children to be collaborative at a very early age [2]. True collaboration is more than just being "helpful", in that one could help someone else by setting up the "environment" such that the other agent succeeds. For example, we might be helpful with children in such a way that they do not know what we have done to help them. However, most conversational systems, even those dubbed as "assistants," do not know how to be helpful, much less to collaborate. At the dialogue level, they are generally incapable of inferring and responding to the intention that motivated the utterance. We and others have argued that deep collaboration involves agents' (mutual) beliefs and joint intentions to ensure that the joint goals are achieved [3–6]. Whereas physical actions are planned to alter the physical world, communicative acts are planned to alter the (joint)

---

[1] ACM Class: I.2.7; I.2.8; I.2.4; I.2.3; I.2.11. Additional Keywords: Dialogue, planning, reasoning, multiagent systems, modal logic, intention, plan recognition, theory of mind

[2] We will use the expression "BDI" even though the system deals with (persistent) goals rather than desires.