A Simulation-based Framework for Spoken Language Understanding and Action Selection in Situated Interaction

David Cohen
Carnegie Mellon University
Nasa Research Park
Moffett Field, CA
david.cohen@sv.cmu.edu

Ian Lane
Carnegie Mellon University
Nasa Research Park
Moffett Field, CA
lane@cs.cmu.edu

Abstract

This paper introduces a simulation-based framework for performing action selection and understanding for interactive agents. By simulating the objects and actions relevant to an interaction, an agent can semantically ground natural language and interact considerately and on its own initiative in situated environments. The framework proposed in this paper leverages models of the environment, user and system to predict possible future world states via simulation. It leverages understanding of spoken language and multimodal input to estimate the state of the ongoing interaction and select actions based on the utility of future outcomes in the simulated world. In this paper we introduce this framework and demonstrate its effectiveness for in-car navigation.

1 Introduction

Speech and multimodal interactive systems have many challenges to overcome before they can effectively interact with users in the real world. These challenges include semantically grounding vague and ambiguous natural language utterances, understanding the user's knowledge and capabilities, and acting on their own initiative to plan and take appropriate actions in complex environments. To overcome these challenges, interactive agents require more than just models of the environment, user goals, and attention, they need the ability to infer the consequences of both their and the users' actions – a capability which simulation provides.

For each given task, an agent must plan the best way to carry it out. In many cases, a simple set of context-dependent behavior templates will not be sufficient. For example, if an in-car navigation assistant is trying to direct a driver to his destination, it should probably not give directions within the driver's own neighborhood, with which he is already familiar. However, it should inform the driver if there is road construction in the area of which he/she is unaware. Alternatively, if the driver is having an important conversation and the cost of the detour is outweighed by the cost of interrupting the conversation, perhaps the system should remain quiet. Understanding all the contexts that affect interaction is difficult and defining a set of heuristics to choose the appropriate behavior will quickly become unmanageable. An agent in the real world will be faced with complicated situations that will require planning and an understanding of the effects its actions will have.

To capture the full context necessary to perform understanding and planning in situated interaction, this paper argues for a unified model of the environment, the user's knowledge, attention and goals, and the simulated consequences of different courses of action.

2 Related Work

Early work in deep natural language understanding (Schank and Abelson, 1977; Wilensky, 1983) formed cognitive theories and developed software to reinforce the idea that understanding an agent's words requires an understanding of that agent's plans, goals, and planning mechanisms. Other work (Allen and Perrault, 1980) focused on identifying these plans and goals from the partial information available; interpreting speech acts as primitive actions in a STRIPS planner (Fikes and Nilsson, 1971), and using heuristics to determine an agent's plan based on their speech acts. Traum (1994) adopted a similar definition of speech acts, and developed a computational theory of grounding whereby multiple agents come to understand each other's plans and meaning.

NAACL-HLT 2012 Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data, pages 33–36,
Montréal, Canada, June 7, 2012. ©2012 Association for Computational Linguistics

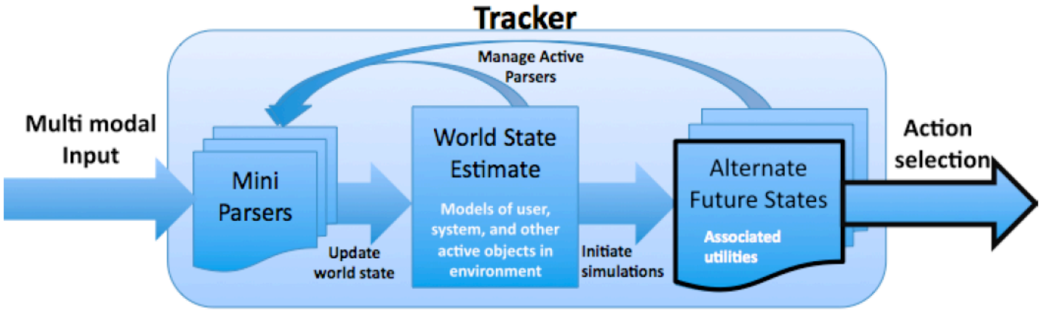


Figure 1: Overview of the proposed simulation-based understanding and action selection framework.

Previous work on considerate mixed-initiative systems has placed an emphasis on modeling the user's mental state, particularly attention and cognitive loading. Horvitz et al. (2003) treat attention as critical to reasoning about the value of taking action and potentially disrupting users. Multiple modalities such as speech and gesture recognition, as well as mouse and keyboard behavior all contribute to their models of attention. Their work also stressed the importance of attention cues in effective collaborative communication. Other work from the same author (Horvitz, 1999) probabilistically tracked a belief in the user's goal based on attentional cues, specifically trying to determine if a behavior from the system was desired. This work all reinforces the idea that close attention to the user's mental state must be paid to act considerately with mixed initiative, but never attempts to endow a system with the ability to reason about the consequences of its actions.

There are several existing paradigms for spoken dialogue systems. RavenClaw (Bohus and Rudnicky, 2009) uses a human-engineered task tree to guide the logic of an interaction, which allows for well-understood behavior, but does not permit the flexibility of planning needed for complex, dynamic interaction. The collaborative agent framework, COLLAGEN (Rich and Sidner, 1996), specifies the data structures for recipes and attention models based off the SharedPlan collaborative discourse framework. The framework proposed by Allen and et al. (2002) is built on a collaborative discourse framework similar to SharedPlan, and is similar to our work in its situation theoretic world model and focus on user goal and plan modeling. However, to the best of our knowledge, these frameworks have never been successfully applied to a situated agent in a dynamic environment with many interacting objects and a wealth of multi-modal input as is available within an in-car assistant environment. It is in these situations that we believe our framework will demonstrate its applicability compared to prior approaches.

3 A Simulation-based Framework for Understanding Situated Interaction[1]

In this paper, we propose a framework in which an interactive agent leverages a model of the ongoing situated interaction and simulations of possible future scenarios to perform understanding and decision-making (Figure 1). The model supports complex inference about natural language as well as other modalities of input, and provides a suitable environment for the system to evaluate possible courses of action. As an example, we evaluated the effectiveness of this framework for planning and interacting in an in-car navigation assistant.

Simulated Interaction and Environment

The system models its environment in terms of an object-oriented probabilistic model that allows for multiple simultaneous actions. It is assumed that the model is an incomplete view of the world, and there are objects that the model is unaware of. Included in this model is the set of primitive actions all the objects in the world can take, defined by their pre-conditions and post-conditions. Through simulation, the system can project the current world state forward in time in an attempt to predict possible futures. Within each simulated scenario, the system, user, and any number of other actors will interact. At each time step, every object selects a primitive action, which is applied to the world if its pre-conditions have been met.

[1] An initial version of the simulator used in the work can be download from: http://speech.sv.cmu.edu/SimInteraction

Programs for Modeling High-Level Actions

In order to make inferences about the long-term behavior of objects in the simulator, plans and high-level actions need a representation within the simulator. To do this, programs are defined for several realistic behaviors for each actor. These programs are a specific form of options (Sutton et al., 1999), which in the context of a Markov Decision Process are closed-loop policies for choosing action over an extended period of time.

In the current implementation, programs are finite state machines, which are resumed at each time interval, changing state based on the actor's internal state until a primitive action is selected.

Modeling User Knowledge and Awareness

An actor carrying out a program will choose a different sequence of actions depending on their internal mental state. That is why the world model must contain this information to make accurate predictions. In particular, a user's knowledge and attention play critical roles in their decision-making, and thus must be modeled.

Tracking and Parsing

The tracker maintains the current world model including the set of objects that are relevant for simulation and estimated distributions over uncertain variables such as the user's mental state and the programs being run by all relevant objects. The tracker is responsible for initiating simulations to project the situation model into the future. The tracker also manages and interfaces to a set of mini-parsers which interpret input across multiple modalities in various ways.

In the proposed framework, the tracker also uses information from the parsers to add new objects to the world model, and modify the parameters of the objects already in the model. Additional parsers can be spawned based on simulation results. For example, if a simulated scenario predicts the car running out of gas, the tracker might spawn a new parser to interpret the driver's awareness of their gas level based on gaze.

Utility Estimation and Action Selection

The desirability of every simulated scenario is determined by a utility score, defined by the system designer to maximize the system's usefulness. The system includes itself and its own possible programs in each simulation it runs, and picks the program that gives the best expected utility.

Table 1: Description of three evaluation tasks.

TaskID	Task Description
1	Destination is a business in downtown area, mostly a straight path as a warm-up task.
2	Destination is a residence in Palo Alto, insufficient gas to get to destination.
3	Destination is a residence in Mountain View, retrace much of the path from Task 2.

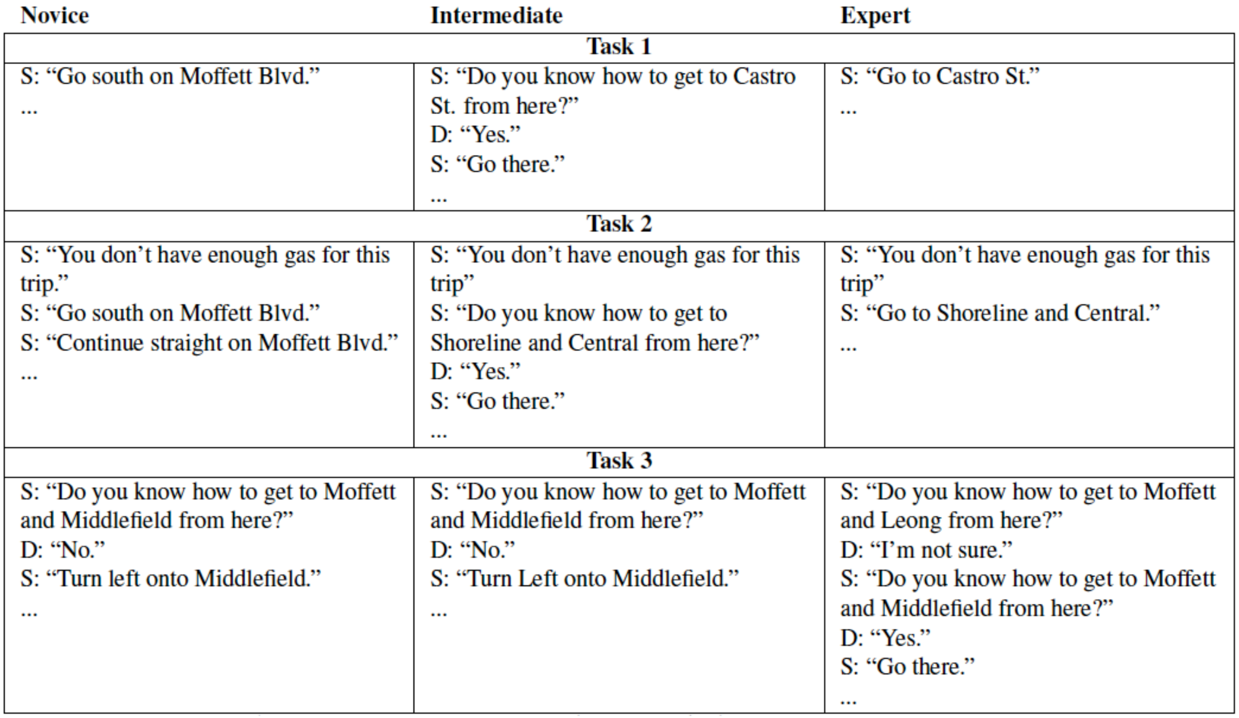
Table 2: Average number of system turns for baseline and the proposed system. System turns include questions, notifications, and instructions.

TaskID	Novice	Intermediate	Expert
1	7.0	7.0	3.0
2	13.0	15.8	9.0
3	12.0	12.8	9.0

4 Demonstration Example

We demonstrate the effectiveness of the proposed framework for an in-car navigation assistant. We tested this demonstration with ten test subjects each navigating through three the tasks listed in Table 1. The subjects navigated through Mountain View and Palo Alto, California in Google Earth$^{\text{TM}}$ while a supervisor observed their progress, entered it into the system and relayed messages between the subject and system. Some subjects had been in the area only a few times and some were current residents of Mountain View and neighboring cities. Based off the subjects' initial self-assessment, the system was given one of three different starting familiarity map estimates - novice, medium, and expert. These initial estimates reflected our intuitive assessment of the likelihood that a driver would know major streets and neighborhoods.

For users with different levels of familiarity we counted the number of system turns, which include questions, notifications, and instructions required to complete the task. These counts are shown in Table 2 show a decrease in the number of system turns across all tasks for users who were more familiar with the area. This is a direct result of the system's ability to direct these users to waypoints they were familiar with along the route, saving unnecessary directions. Example interactions obtained from the experiments are shown in Figure 1.



Novice	Intermediate	Expert
Task 1		
S: "Go south on Moffett Blvd." ...	S: "Do you know how to get to Castro St. from here?" D: "Yes." S: "Go there." ...	S: "Go to Castro St." ...
Task 2		
S: "You don't have enough gas for this trip." S: "Go south on Moffett Blvd." S: "Continue straight on Moffett Blvd." ...	S: "You don't have enough gas for this trip" S: "Do you know how to get to Shoreline and Central from here?" D: "Yes." S: "Go there." ...	S: "You don't have enough gas for this trip" S: "Go to Shoreline and Central." ...
Task 3		
S: "Do you know how to get to Moffett and Middlefield from here?" D: "No." S: "Turn left onto Middlefield." ...	S: "Do you know how to get to Moffett and Middlefield from here?" D: "No." S: "Turn Left onto Middlefield." ...	S: "Do you know how to get to Moffett and Leong from here?" D: "I'm not sure." S: "Do you know how to get to Moffett and Middlefield from here?" D: "Yes." S: "Go there." ...

Figure 2: Sample interactions from subjects with different starting familiarity estimates.

5 Conclusions

This paper introduces a simulation-based framework for performing action selection and understanding in an interactive agent. The framework uses a simulator to predict possible future world states incorporating and updating models of the environment, user and system based on observed input. Understanding of spoken language and multimodal input is performed leveraging the past, current and future world states in the simulator. Action selection is performed based on the utility of future world states and the expected user goal. In this paper we introduce this framework and demonstrate its effectiveness for in-car navigation.

References

James Allen, Nate Blaylock, George Ferguson 2002. A Problem Solving Model for Collaborative Agents. Proc. AAMAS.

James F. Allen and C. Raymond Perrault. 1980. Analyzing Intention in Utterances. Artificial Intelligence.

Dan Bohus and Eric Horvitz. 2011. Multiparty Turn Taking in Situated Dialog: Study, Lessons, Directions Proc. SIGdial.

Dan Bohus and Alexander I. Rudnicky. 2009. The RavenClaw Dialog Management Framework: Architecture and Systems. Computer Speech and Language.

Richard E. Fikes and Nils J. Nilsson 1971. STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. IJCAI.

Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces Proc. SIGCHI.

Eric Horvitz, Carl Kadie, Tim Paek, David Hovel. 2003. Models of Attention in Computing and Communication: From Principles to Applications. Communications of ACM.

Charles Rich and Candace L. Sidner 1996. COLLAGEN: When Agents Collaborate with People. Mitsubishi Electric Research Laboratories Inc.

Roger Schank and Robert Abelson. 1977. Scripts Plans Goals and Understanding: an Inquiry into Human Knowledge Structures. Lawrence Erlbaum Associates, Inc., Publishers.

Richard S. Sutton, Doina Precup, Satinder Singh. 1999. Between MDPs and semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. Artificial Intelligence.

David R. Traum. 1994. A Computational Theory of Grounding in Natural Language Conversation Ph.D. Thesis

Robert Wilensky. 1983. Planning and Understanding: A Computational Approach to Human Reasoning. The Addison-Wesley series in artificial intelligence.