Socially-Aware Animated Intelligent Personal Assistant Agent

Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao,
Oscar J. Romero, Sushma Anand Akoju and Justine Cassell
ArticuLab, Carnegie Mellon University, Pittsburgh, PA 15213 USA
`{yoichim, ranzhao, justine}@cs.cmu.edu`
`{arjunb1, oscarr, sakoju}@andrew.cmu.edu`

Abstract

SARA (Socially-Aware Robot Assistant) is an embodied intelligent personal assistant that analyses the user's visual (head and face movement), vocal (acoustic features) and verbal (conversational strategies) behaviours to estimate its rapport level with the user, and uses its own appropriate visual, vocal and verbal behaviors to achieve task and social goals. The presented agent aids conference attendees by eliciting their preferences through building rapport, and then making informed personalized recommendations about sessions to attend and people to meet.

1 Introduction

Currently major tech companies envision intelligent personal assistants, such as Apple Siri, Microsoft Cortana, and Amazon Alexa as the front ends to their services. However those assistants really play little other role than to query, with voice input and output - they fulfill very few of the functions that a human assistant might. In this demo, we present SARA, the Socially-Aware Robot Assistant, represented by a humanoid animated character on a computer screen, which achieves similar functionality, but through multimodal interaction, and with a focus on building a social relationship with users. Currently SARA is the front end to an event app. The animated character engages its users in a conversation to elicit their goals and preferences and uses them to recommend relevant conference sessions to attend and people to meet. During this process, the system monitors the use of specific conversational strategies (such as self-disclosure, praise, reference to shared experience, etc.) by the human user and uses this input, as well as acoustic and nonverbal input, to estimate the level of rapport between the user and system. The system then employs conversational strategies shown in our prior work to raise the level of rapport with the human user, or to maintain it at the same level if it is already high (Zhao et al., 2014), (Zhao et al., 2016b). The goal is to use rapport to elicit personal information from the user that can be used to improve the helpfulness and personalization of system responses.

2 SARA's Computational Architecture

SARA is therefore designed to build interpersonal closeness over the course of a conversation through understanding and generation visual, vocal, and verbal behaviors. The current system leverages prior work on the dynamics of rapport (Zhao et al., 2014), and the initial consideration of the computational architecture of a rapport building agent (Papangelis et al., 2014). Figure 1 shows the overview of the architecture. All modules of the system are built on top of the Virtual Human Toolkit (Hartholt et al., 2013). Main modules of our architecture are described below.

2.1 Visual and Vocal Input Analysis

Microsoft's Cognitive Services API converts speech to text, which is then fed to Microsoft's LUIS (Language Understanding Intelligent Service) to identify user intents. In the demo, as the train data of this specific domain is still limited, a Wizard of Oz GUI will be served as backup in the case of speech recognition and natural language understanding errors. OpenSmile (Eyben et al., 2010) extracts acoustic features from the audio signal, including fundamental frequency (F0), loudness (SMA), jitter and shimmer, which then serve as input to the rapport estimator and the conversational strategy classifier modules. Open-Face (Baltrušaitis et al., 2016)) detects 3D facial landmarks, head pose, gaze and Action Units, and
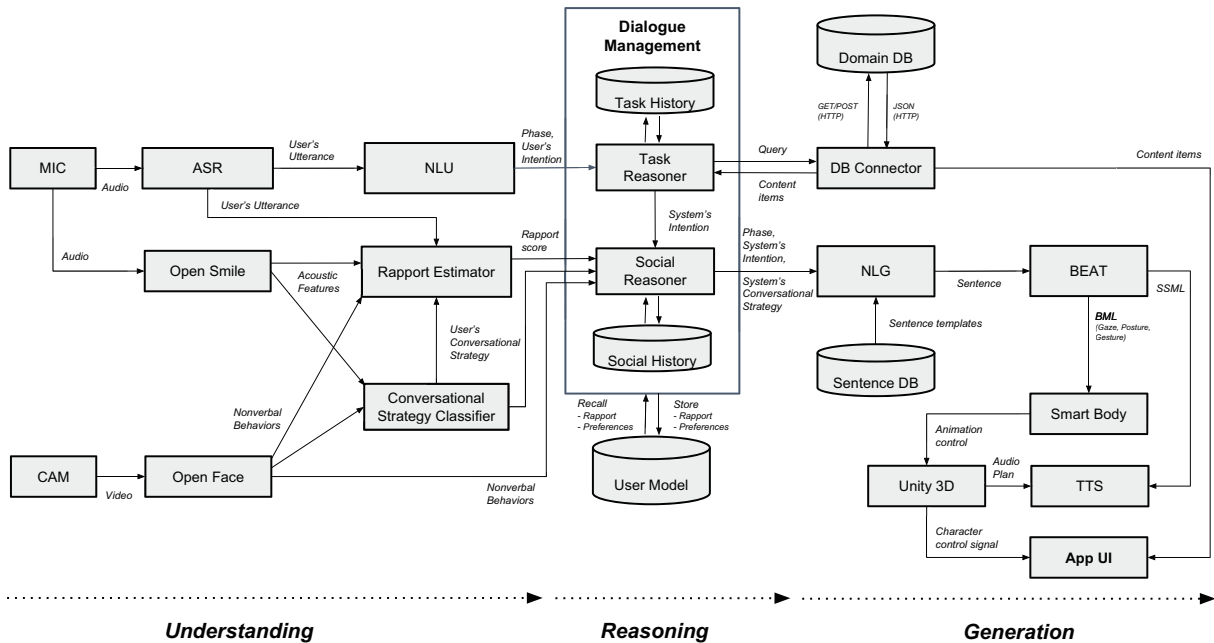
Figure 1: SARA Architecture

these also serve as input to the rapport estimator (smiles, for example, have been shown in the corpus we trained the estimator on to have a strong impact on rapport (Zhao et al., 2016b)).

2.2 Conversational Strategy Classifier

We implemented a multimodal conversational strategy classifier to automatically recognize particular styles and strategies of talking that contribute to building, maintaining or sometimes destroying a budding relationship. These include: self-disclosure (SD), elicit self-disclosure (QE), reference to shared experience (RSD), praise (PR), and violation of social norms (VSN). By analyzing rich contextual features drawn from verbal, visual and vocal modalities of the speaker and interlocutor in both the current and previous turns, we can successfully recognize these dialogue phenomena in user input with an accuracy of over 80% and with a kappa of over 60% (Zhao et al., 2016a).

2.3 Rapport Estimator

We also implemented an automatic multimodal rapport estimator, based on the framework of temporal association rule learning (Guillame-Bert and Crowley, 2012), to perform a fine-grained investigation into how sequences of interlocutor behaviors lead to (are followed by) increases and decreases in interpersonal rapport. The behaviors analyzed include visual behaviors such as eye gaze and smiles and verbal conversational strategies

such as SD, RSE, VSN, PR and BC. The rapport forecasting model involves two-step fusion of learned temporal associated rules: in the first step, the goal is to learn the weighted contribution (vote) of each temporal association rule in predicting the presence/absence of a certain rapport state (via seven random-forest classifiers); in the second step, the goal is to learn the weight corresponding to each of the binary classifiers for the rapport states, in order to predict the absolute continuous value of rapport (via linear regression) model (Zhao et al., 2016b). Ground truth comes from annotations of rapport in videos of peer tutoring sessions divided into 30 second slices which are then randomized (see (Zhao et al., 2014) for details).

2.4 Dialogue Management

The dialogue manager is composed of a task reasoner that focuses on obtaining information to fulfill the user's goals, and a social reasoner that chooses ways of talking that are intended to build rapport in the service of better achieving the user's goals. A task and social history, and a user model, also play a role in dialogue management, but will not be further discussed here.

2.4.1 Task Reasoner

The Task Reasoner is predicated on the system maintaining initiative to the extent possible. It is implemented as a finite state machine whose transitions are determined by different kinds of trig-

gering events or conditions such as: user's intents (extracted by the NLU), past and current state of the dialogue (stored by the task history) and other contextual information (e.g., how many sessions the agent has recommended so far). Task Reasoner's output can be either a query to the domain database or a system intent that will serve as input to the Social Reasoner and hence the NLG modules. In order to handle those cases where the user takes the initiative, the module allows a specific set of user intents to cause the system to transition from its current state to a state which can appropriately handle the user's request. The task Reasoner use a statistical discriminative state tracking approach to update the dialogue state and deal with error handling, sub-dialog,s and grounding acknowledgements, similar to the implementation of the Alex framework (Jurčíček et al., 2014).

2.4.2 Social Reasoner

The Social Reasoner is designed as a network of interacting nodes where decision-making emerges from the dynamics of competence and collaboration relationships among those nodes. That is, it is implemented as a Behavior Network as originally proposed by (Maes, 1989) and extended by (Romero, 2011). Such a network is ideal here as it can efficiently make both short-term decisions (real-time or reactive reasoning) and long-term decisions (deliberative reasoning and planning). The network's structure relies on observations extracted from data-driven models (in this case the collected data referenced above). Each node (behavior) corresponds to a specific conversational strategy (e.g., SD, PR, QE, etc.) and links between nodes denote either inhibitory or excitatory relationships which are labeled as precondition and post-condition premises. As preconditions, each node defines a set of possible system intents (generated by the Task Reasoner, e.g., "self_introduction", "start_goal_elicitation", etc.), rapport levels (high, medium or low), user conversational strategies (SD, VSN, PR, etc.), visuals (e.g., smile, head nod, eye gaze, etc.), and system's conversational strategy history (e.g., system has performed VSN three times in a row). Postconditions are the expected user's state (e.g., rapport score increases, user smiles, etc.) after performing the current conversational strategy, and what conversational strategy should be performed next. For instance, when a conversation starts (i.e., during the greeting phase) the most likely sequence of nodes could be: [ASN, SD, PR, SD ...VSN ...] i.e., initially the system establishes a cordial and respectful communication with user (ASN), then it uses SD as an icebreaking strategy, followed by PR to encourage the user to also perform SD. After some interaction, if the rapport level is high, a VSN is performed. The Social Reasoner is adaptive enough to respond to unexpected user's actions by tailoring a reactive plan that emerges *implicitly* from the forward and backward spreading activation dynamics and as result of tuning the network's parameters which determine reasoner's functionality (more oriented to goals vs. current state, or more adaptive vs. biased to ongoing plans, or more thoughtful vs. faster.).

2.5 NLG and Behavior Generation

On the basis of the output of the dialogue manager (which includes the current conversational phase, system intent, and desired conversational strategy) sentence and behavior plans are generated. The Natural Language Generator (NLG) selects syntactic templates associated with the selected conversational strategy from the sentence database and then fills them in with content from database queries performed by the task reasoner. The generated sentence plan is sent to BEAT, a non-verbal behavior generator (Cassell et al., 2004), which tailors a behavior plan (including relevant hand gestures, eye gaze, head nods, etc.) and outputs the plan as BML (Behavior Markup Language), which is a part of the Virtual Human Toolkit (Hartholt et al., 2013). This plan is then sent to SmartBody, which renders the required non-verbal behaviours.

3 Dialogue examples

SARA was demoed at the World Economic Forum in Tianjin China in June 2016 where it served as the front end to the event app. Over 100 participants interacted with SARA to get advice on sessions to attend and people to meet. The system operated with a Wizard of Oz GUI serving as backup in the case of recognition (speech recognition and natural language understanding), task reasoning errors, and network disruptions. Table 1 shows an extract from an actual interaction with the system, annotated with the outputs of the different modules as the system works to meet social and task goals.

Table 1: Extract from an actual interaction

Turn	Speaker	Utterance	Intent	Strategy
1	System	I'm here to be your personal assistant and help you get the most out of the conference, but to be honest, I'm a little nervous. This is my first time here, is it yours as well?	introduce, do_attendance_elicitation	SD, QE
2	User	Actually, this is my first time here as well.	first_time_attending	SD
3	System	I'm finding it overwhelming. I bet you are too.	feedback_attendance_elicitation_first_time	SD, QE
4	User	Ya, there's just so much going on. But i'm getting a hang of it.	positive_confirmation	SD
5	System	Great. Now, If you wouldn't mind sharing your goals for attending the world economic forum this year, I can try to help you achieve them.	do_goal_elicitation	ASN
6	User	Well, I'm here to meet people who are passionate about the use of technology in education, and to know more about the latest developments in the field.	recommend_people, recommend_sessions	None
7	System	Great. Good goals. Can you tell me a little bit about your work so I understand your interests better?	feedback_goals, do_interest_elicitation	ASN
8	User	I am the founder of a non profit educational startup which helps deliver quality education to the remote areas.	like {startup, education}	None
9	System	Very Interesting. Let's start by finding some interesting people for you to meet that are relevant to your interests. This is my favorite part. Let me look this up. One moment.	feedback_interest_elicitation, do_person_recommendation	ASN, SD
10	User	Sure	positive_confirmation	None
11	System	I think you would enjoy meeting this attendee. On the screen are some more details. Well? What do you think?	outcome_person_recommendation, end_person_recommendation	ASN, VSN

4   Conclusion and Future Work

We have described the design and first implementation of an end-to-end socially-aware embodied intelligent personal assistant. The next step is to evaluate the validity of our approach by using the data collected at the World Economic Forum to assess whether rapport does increase over the conversation. Subsequent implementations will, among other goals, improve the ability of the system to collect data about the user and employ it in subsequent conversations, as well as the generativity of the NLG module, and social appropriateness of nonverbal behaviors generated by BEAT. We hope that data collected at SIGDIAL will help us to work towards these goals.

References

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*, pages 163–185. Springer.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM.

Mathieu Guillame-Bert and James L. Crowley. 2012. Learning temporal association rules on symbolic time sequences. pages 159–174.

A Hartholt, D Traum, SC Marsella, A Shapiro, G Stratou, A Leuski, LP Morency, and J Gratch. 2013. All together now: Introducing the virtual human toolkit. In *Int. Conf. on Intelligent Virtual Humans*.

Filip Jurčíček, Ondřej Dušek, Ondřej Plátek, and Lukáš Žilka. 2014. Alex: A statistical dialogue systems framework. In *Text, Speech and Dialogue: 17th International Conference, TSD*, pages 587–594.

Pattie Maes. 1989. How to do the right thing. *Connection Science*, 1(3):291–323.

Alexandros Papangelis, Ran Zhao, and Justine Cassell. 2014. Towards a computational architecture of dyadic rapport management for virtual agents. In *Intelligent Virtual Agents*, pages 320–324.

Oscar J. Romero. 2011. An evolutionary behavioral model for decision making. *Adaptive Behavior*, 19(6):451–475.

Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *Intelligent Virtual Agents*, pages 514–527.

Ran Zhao, Tanmay Sinha, Alan Black, and Justine Cassell. 2016a. Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *17th Annual SIGdial Meeting on Discourse and Dialogue*.

Ran Zhao, Tanmay Sinha, Alan Black, and Justine Cassell. 2016b. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *16th International Conference on Intelligent Virtual Agents*.