SMILEE: Symmetric Multi-modal Interactions with Language-gesture Enabled (AI) Embodiment

Sujeong Kim, David Salter, Luke DeLuccia, Kilho Son, Mohamed R. Amer and Amir Tamrakar*
SRI International, 201 Washington Rd, Princeton, NJ08540
https://sites.google.com/view/smilee

Abstract

We demonstrate an intelligent conversational agent system designed for advancing human-machine collaborative tasks. The agent is able to interpret a user's communicative intent from both their verbal utterances and non-verbal behaviors, such as gestures. The agent is also itself able to communicate both with natural language and gestures, through its embodiment as an avatar thus facilitating natural symmetric multi-modal interactions. We demonstrate two intelligent agents with specialized skills in the Blocks World as use-cases of our system.

1 Introduction

Recent advances in speech recognition and natural language processing techniques have resulted in increasing use of intelligent assistants, such as Google Assistant, Siri, and Alexa, in our daily lives, replacing keyboard or touch interfaces. However, the interactions with these assistants are still limited to just the verbal modality.

In this paper, we present an intelligent conversational agent system (SMILEE) designed to advance the state of the art in human-machine interaction. The main idea underlying this system is the observation that **non-verbal behavior** (primarily gestures) encodes information that both complements and supplements speech in human-to-human communication. Our studies in the Blocks-World (BW) have shown that gestures are frequently used with speech taking on both complementary roles (reinforcing the meaning) and supplementary roles (adding information to what was verbalized), and contribute towards facilitating communication (Kim et al., 2018). Thus we assert that gestures need to be taken into account when deducing the meaning of complex ideas exchanged during communication and this has to be done in a joint-inferencing process along with the natural language understanding process.

In the same vein, we also assert that communication between humans and computers should be symmetric and multi-modal in both directions for it to be truly natural. Thus, in order to facilitate the communication of a machines complex ideas to the human, the machine's utterances also need to be embellished with appropriate non-verbal behaviors. This argues for using a computer-generated avatar for embodying the machine. Only then will humans be able to naturally communicate with machines, as they do with other humans. Not only will this ease the communication but it also allow the communication to be more accurate.

In the following sections, we summarize various components of SMILEE and demonstrate two preliminary use-cases we have built.

2 Related Work

(Foster et al., 2009) explores different strategies for generating instructions given to the users by the robot for collaborative toy assembly type of tasks. (Perera et al., 2017) propose a system to learn new concepts (e.g., unseen structure) in BW environments from the descriptions of its properties. (She et al., 2014) propose a method to teach a robotic arm new actions for placing blocks using dialogues. There have also been efforts to understand language in the BW domain. Human-generated instructions in virtual blocks world are collected by (Bisk et al., 2016). Using this dataset, a neural-network model for 3D spatial operations has proposed (Krishnaswamy et al., 2018). In terms of non-verbal behaviors, (Salter et al., 2015) present a dataset containing two people building structures with basic shapes while only communicating with non-verbal queues.The key difference

*amir.tamrakar@sri.com

Proceedings of NAACL-HLT 2018: Demonstrations, pages 86–90
New Orleans, Louisiana, June 2 - 4, 2018.

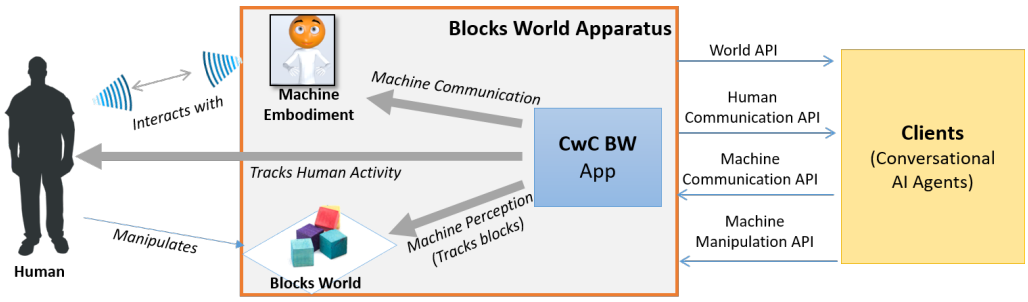

Figure 1: The Functional Architecture of the BW Apparatus. See text for details.

of our work from these work is that we use speech and gestures together for understanding human intents.

3 The Blocks World (BW) Apparatus

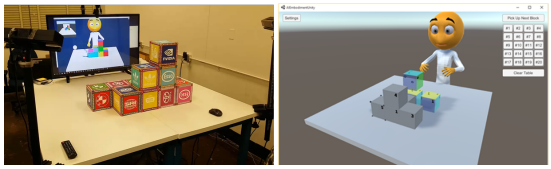

Figure 2: **Blocks World Apparatus** Physical setup (left) and a virtual setup (right).

The Blocks World Apparatus (Figs.1 & 2) is a computer vision based platform to provide I/O abilities for AI agents (Son et al., 2015). It was designed for the purpose of having intelligent conversations about blocks on a table, i.e., the Blocks World. The platform acts as the eyes and ears for the AI agent, tracking not only the blocks on the table (Son et al., 2016) but also multi-modal behaviors of the human interacting with it, both verbal and non-verbal (Siddique et al., 2015). It also provides an embodiment of the machine in the form of a simple humanoid avatar for the users to interact with. The Apparatus provides a number of APIs for the conversational AI agents to interface with the system.

The apparatus is also available as a virtual system (Fig. 2). The virtual version was developed using Unity (Unity, 2017) as an option for users who are not able to setup the full physical system. It allows the users to use the same exact APIs to interact with the apparatus. The virtual system is also deployed as a web app for remote users. This system is publicly available for use by the research community (Salter et al., 2017)[1].

[1] https://sites.google.com/view/playwithsmilee/home

4 SMILEE

SMILEE is the realization of our goal to allow natural verbal and non-verbal communication between a human and a conversational AI agent. Although the current system has been built for having conversations in the Blocks World, it is not restricted to this use case. We have used The Rochester Interactive Planning System (TRIPS) architecture (Ferguson and Allen, 1998) for realizing our system. The TRIPS framework approaches conversations as collaborative problem solving tasks. It provides a parser and an interpretation manager to normalize many forms of utterances into logical forms (LF) which is then formed into problem solving acts represented in AKRL. The TRIPS agent architecture utilizes a loose coupling of a number of independent agents communicating with each other via standardized KQML messages. In this section, we describe various agents that compose the SMILEE system (Fig. 3).

4.1 Scene Perception Agent

The scene perception agent processes the block tracking information from the BW apparatus to generate perceptual interpretations of the overall structure defined by the collection of positions and orientations of the tracked blocks. The perceptual interpretations include geometric relations between blocks, physical constraints, and perceptual grouping of blocks into larger substructures (such as row, stack, columns, etc). This grouping can also be done based on an attribute of the blocks, such as color, to generate other sub-grouping proposals to aid in the communication. The set of these relations are used to describe the scene.

4.2 Deictic Gesture Interpretation Agent

The gesture interpretation agent interprets the raw human communication signals from the BW apparatus. Current implementation of our gesture inter-

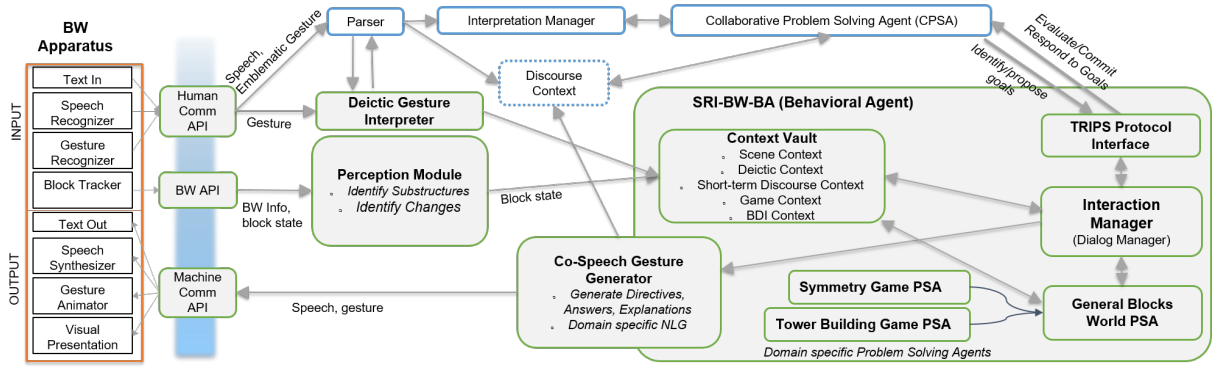

Figure 3: The SMILEE conversational AI Agent and its interactions with BW Apparatus and TRIPS modules. The modules highlighted with green are our own agents while the modules in blue are preexisting modules in the TRIPS system.

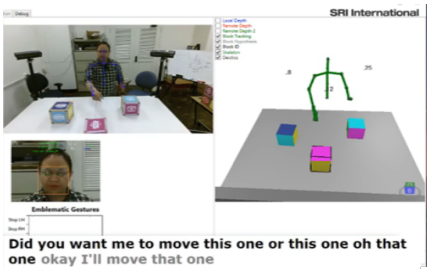

Figure 4: Real-time deictic gesture recognition. (Right) The block pointed by the user is highlighted with bold edges. (Bottom) The words from the live speech recognition. (Left-Top) Raw video input (Left-middle) facial behavior recognition.

pretation module converts the tokenized emblematic gestures directly into verbal utterances and allows them to be parsed in the normal manner. For deictic gesture interpretation, we use a real-time ASR engine for the verbal input and temporally align it with the information from the deictic gesture recognition stream to perform reference resolution (Fig.4). The pronoun-resolved utterance string goes through the parser to the rest of the interpretation modules. Our system is capable of tracking both the pointing and touching types of deictic gestures.

4.3 Co-speech Gesture Generation Agent

This agent is responsible for generating communication behaviors for the AI embodiment, in terms of deciding 'what to say' and 'what to do'. We approach this problem in two separate steps (Kim and Tamrakar, 2017). First, we convert the targeted instruction (machine computed plan step) into a fully verbal utterance and optimize it using appropriate contexts. Next, we explore the possibility of replacing parts of this utterance with gestures or simply supplementing them for better communication.

Below, we show various utterances generated by the agent for the same instruction "put a block at a position (x, y, z)", along with the cost (complexity) of the instruction. The spatial and task-related contexts significantly lowers the complexity.

$cost = 0$, Please put a block on top of it
$cost = 0.5$, Please put a new block on top of the block you just placed
$cost = 0.5$, Please put a new block behind the stack you just made
$cost = 1$, Please put a new block on top of the block on the top
$cost = 2$, Please put a new block on top of the block that is at the back and on the left

Deictic gestures also increase the chance of correct and efficient communication, by simplifying the dialogue or by confining the region of interest. Note that even when a gesture increases the length of the utterance, the communication can still be more effective. For example, "put a block on top of the stack there" can make it easier for the user to locate the stack than just saying "put a block on top of the stack".

4.4 SRI-BW Behavioral Agent (BA)

This behavioral agent (BA) is the main agent that enables the continuous interactions between the user and the machine. When the agent understands the user's intent without ambiguity, it performs an appropriate action. Otherwise, the agent asks clarifying questions to resolve missing or ambiguous information. For example:

USR: Put a block on the table.
SYS: Where on the table?

USR: Right there. [points to a location].
SYS: OK. I am putting a block at $(0, 0, 0)$
USR: Put another block next to it.
SYS: Can you be more specific? in front of/behind it? or to the right or left of it? or on top of it?
USR: Put it to the left of it.
SYS: But there is already a block there. Do you mean to the right of it?
USR: Yes.
SYS: Ok. I am putting a block at $(1, 0, 0)$.

The behavior agent consists of the following modules:

TRIPS protocol Interface handles the TRIPS choreography. It routes messages to the appropriate Problem Solving Agents.

Interaction Manager manages the state machines for dialogs/interactions.

Context Vault keeps track of all the states and contexts for the Problem Solving Agents (PSAs).

The BW PSAs are the domain specific problem solving agents for the Blocks World in the TRIPS architecture. They contain problem-specific general knowledge including appropriate representations and specific capabilities needed to carry out the actions requested by the user. They also keep track of the state of the problem.

We demonstrate use-cases in Section 5.

5 Use cases

We demonstrate two use-cases of specialized PSAs, namely, the general blocks-world PSA and the symmetry game PSA. Please see the videos on the website for more details.

5.1 The General Blocks-World PSA

The general blocks-world agent has knowledge about general tasks and properties of the BW environment. The properties of the world consist of physical properties like measurements, counts of the blocks with particular attributes and/or in particular locations, and spatial relationships between the sub-structures, etc. The tasks are related to manipulating blocks to build a structure, such as putting or moving blocks to certain positions. Either the user or the machine can take initiative to ask the other to perform a task-related actions.

5.2 The Symmetry Game PSA

The symmetry game is a collaborative turn-based game in the Blocks World. The user and the computer collaborates to create a symmetric structure out of the blocks. Both the user and the agent have total creative control on the building process but they can also choose to collaborate towards a common goal. The structure should be symmetric at the end of each turn.

The symmetry game agent understands the concept of symmetry within the blocks world and performs tasks like evaluating symmetry of an arbitrary structure. The agent utilizes a planner to formulate moves (plans) to construct symmetric structures, predicting potential solutions several steps ahead of the current step. The symmetry game agent also utilizes a state machine to keep track of the game state based on the rules of the game.

One of the highlights of the symmetry game agent is that it is capable of explaining itself – explaining its plans and the reasonings for its actions.

6 Conclusion and Future Work

We have demonstrated a system for symmetric natural communication with a computer which can interact with its users with verbal and non-verbal communication allowing it to have more robust conversation. We demonstrated two use cases in the BW domain. These use cases are like skills that can be rapidly expanded upon.

We are working on expanding our capabilities for interpreting additional non-verbal cues such as iconic and pantomimed gestures. We are also working on generating natural behaviors for the avatar such as dynamically reacting to the user's actions.

We are also working on performing studies to evaluate the system in terms of the task performance and user experience based on the criteria described in (Kozierok et al., 2018). The SMILEE web page (see Section 3 for detail) is open to public and collects feedback from the users.

Acknowledgments

This work was funded by the Defense Advanced Research Projects Agency (DARPA) under agreement number W911NF-15-C-9244. The views, opinions, and/or conclusions contained in this paper are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied of the DARPA or the DoD. We would like to thank Lucian Galescu of IHMC for his help with TRIPS system. We also thank MITRE for providing the feedback about the system and evaluations.

References

Y. Bisk, D. Marcu, and W. Wong. 2016. Towards a dataset for human computer communication via grounded language acquisition. In *Proc. AAAI Workshop on Symbolic Cognitive Systems*.

George Ferguson and James F. Allen. 1998. Trips: An integrated intelligent problem-solving assistant. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98, pages 567–572, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Mary Ellen Foster, Manuel Giuliani, Amy Isard, Colin Matheson, Jon Oberlander, and Alois Knoll. 2009. Evaluating description and reference strategies in a cooperative human-robot dialogue system. IJCAI.

Sujeong Kim, David Salter, Timur Almaev, Tim Meo, and Amir Tamrakar. 2018. Dataset of human-human interactions while engaging in a tower-building task in the blocks world. Technical report, SRI International.

Sujeong Kim and Amir Tamrakar. 2017. Co-speech gesture generation for communication in the blocks world. Technical report, SRI International.

Robyn Kozierok, Lynette Hirschman, John Aberdeen, Cheryl Clar, Christopher Garay, Bradley Goodman, Tonia Korves, and Matthew Peterson. 2018. Darpa communicating with computers: Program goals and hallmarks.

Nikhil Krishnaswamy, Pradyumna Narayana, Isaac Wang, Kyeongmin Rim, Rahul Bangar, Dhruva Patil, Gururaj Mulay, J. Ross Beveridge, Jaime Ruiz, Bruce A. Draper, and James Pustejovsky. 2018. Learning interpretable spatial operations in a rich 3d blocks world. Association for the Advancement of Artificial Intelligence (AAAI).

Ian E. Perera, James F. Allen, Lucian Galescu, Choh Man Teng, Mark H. Burstein, Scott E. Friedman, David D. McDonald, and Jeffrey M. Rye. 2017. Natural language dialogue for building and learning models and structures. In *AAAI*.

D. A. Salter, A. Tamrakar, B. Siddiquie, M. R. Amer, A. Divakaran, B. Lande, and D. Mehri. 2015. The tower game dataset: A multimodal dataset for analyzing social interaction predicates. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 656–662.

David Salter, Sujeong Kim, and Amir Tamrakar. 2017. A virtual blocks world apparatus for the darpa cwc program. Technical report, SRI International.

Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Yue Chai, and Ning Xi. 2014. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *SIGDIAL Conference*.

Behjat Siddique, David Salter, Gregory Ho, Amir Tamrakar, and Ajay Divakaran. 2015. A blocks world apparatus for the darpa cwc program. Technical report, SRI International.

Kilho Son, David Salter, Tim Sheilds, Tim Meo, Jihua Huang, Sujeong Kim, and Amir Tamrakar. 2015. A blocks world apparatus for the darpa cwc program. Technical report, SRI International.

Kilho Son, David Salter, and Amir Tamrakar. 2016. Accurate block structure understanding in real-time with time-of-flight cameras. Technical report, SRI International.

Unity. 2017. Unity engine (5.6.3).