Agent-Based Dynamic Collaboration Support in a Smart Office Space

Yansen Wang, R. Charles Murray, Haogang Bao, and Carolyn P. Rosé
Language Technologies Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA, 15213
yansenwa,rcmurray,haogangb,cp3a@andrew.cmu.edu

Abstract

For the past 15 years, in computer-supported collaborative learning applications, conversational agents have been used to structure group interactions in online chat-based environments. A series of experimental studies has provided an empirical foundation for the design of chat-based conversational agents that significantly improve learning over no-support control conditions and static-support control conditions. In this demo, we expand upon this foundation, bringing conversational agents to structure group interaction into physical spaces, with the specific goal of facilitating collaboration and learning in workplace scenarios.

1 Introduction

AI-Enhanced human learning is a broad area of research with a history at least 50 years long (Aleven and Kay, 2016), with Carbonell's SCHOLAR system being among the earliest systems (Carbonell, 1970). However, while great strides to introduce technologies to enhance both individual and collaborative learning have been made in relatively structured environments such as the lab and the classroom over the decades of research since that time, less progress has been made in more unstructured environments such as the workplace, where the stakes are far higher and social and political pressures play a more substantial role. This demo presents an apparatus for support of collaboration and learning in workplace scenarios using a Virtual Human facilitator interacting face-to-face through speech and gesture.

Large scale quantitative research, including experimental studies and carefully controlled quasi-experimental corpus studies, are the basis for learning generalizable principles (i.e., causal models) that underlie data-driven design of effective AI-enabled systems that support human learning. In recent decades, process data such as click logs, discourse data, biometric sensors, and images are used to understand the process of human learning more deeply (Lang et al., 2017). Models trained over this process data are also used to enable real-time monitoring and support of learning processes even as groups learn through multi-party discussion (Adamson et al., 2014; Rosé and Ferschke, 2016). Thus, the ability to draw causal inferences to motivate effective interventions and the ability to trigger personalized, just-in-time support go hand-in-hand towards development of AI-enhanced learning experiences.

Much of the prior research on workplace learning is qualitative work, which focuses on deep understanding of individual contexts rather than producing generalizable principles through intervention studies. Thus, there is a dearth of empirical research that can rigorously motivate design of effective AI-enabled interventions to support workplace learning, and the data from such research is unable to support model-enabled real-time sensing technology that would facilitate just-in-time support for learning in the workplace. In response, we have constructed a "Smart Office Space" in which to run lab studies with simulated work conditions in order to discover causal mechanisms that can form the foundation for design.

2 Smart Office Space

2.1 Technical Description

As a resource for exploring how to introduce analytics and just-in-time support for collaborative learning during work, we have assembled a "Smart Office Space" which has been instrumented for behavioral sensing (See Figures 1 and 2). It is designed as a foundation for simulating workplace conditions for collaborative and individual desk work. Figure 1 displays the layout of the room while Figure 2 describes the architecture of the

Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 19 - 22
July 1-3, 2020 ©2020 Association for Computational Linguistics

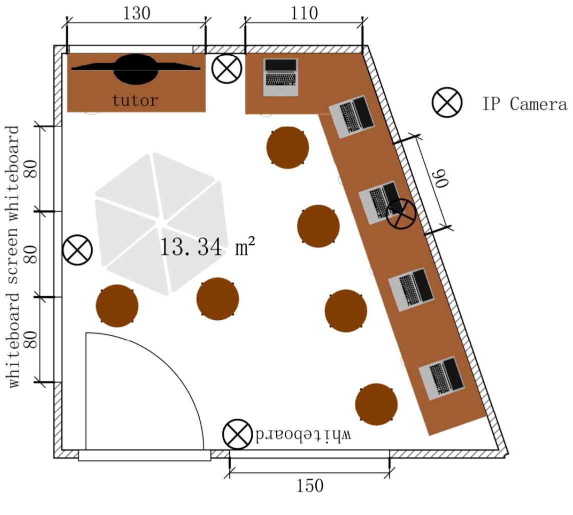

Figure 1: Room layout.

software infrastructure for monitoring and support.

The foundation for the dialogue-based support offered within the Smart Office Space is the Bazaar toolkit (Adamson et al., 2014), which has been used extensively as support for online collaborative learning groups. In this past work, Bazaar agents use what is referred to as an academically productive talk(APT)-based approach, which uses reasoning-focused prompts that encourage participants to articulate and elaborate their own lines of reasoning, and to challenge and extend the reasoning of their teammates in a group discussion. In order for students to learn and contribute to group discussions, it is important for students to articulate their reasoning and build on each other's reasoning. This allows them to identify gaps in their knowledge and to observe how others think differently and might possess knowledge that they are missing. In this way, they have the opportunity to construct knowledge together as a group. The Bazaar toolkit has extensive authoring capabilities that enable a wide range of activities to be authored for virtually any topic area. Dozens of studies of group learning have been conducted with an online, text-based version of Bazaar. Here we place it within the Smart Office Space to communicate, not with text, but with speech and gesture within a physical space.

The room has been instrumented with a variety of sensors including four Lorex 4K cameras with microphones, a Kinect camera with a microphone array, an Intel RealSense depth-sensing camera, and an AWS DeepLens camera. Key software components include the Microsoft Platform for Situated Intelligence (PSI) (Bohus et al., 2017) for coordination across datastreams, CMU Sphinx (Lamere et al., 2003) and the Azure Speech Recognizer for speech recognition, the USC Institute for Creative Technologies Virtual Human Toolkit (VHT) to present an embodied conversational agent (Hartholt et al., 2013), OpenFace for face recognition (Amos et al., 2016), OpenPose for sensing body movement and positioning (Cao et al., 2017), and Bazaar for sensing collaboration-relevant events (such as ideas that have not yet been elaborated or that no one has responded to or built on yet) and triggering support for collaboration in response (such as prompts that direct participants to consider and respond to the contribution of another participant) (Adamson et al., 2014).

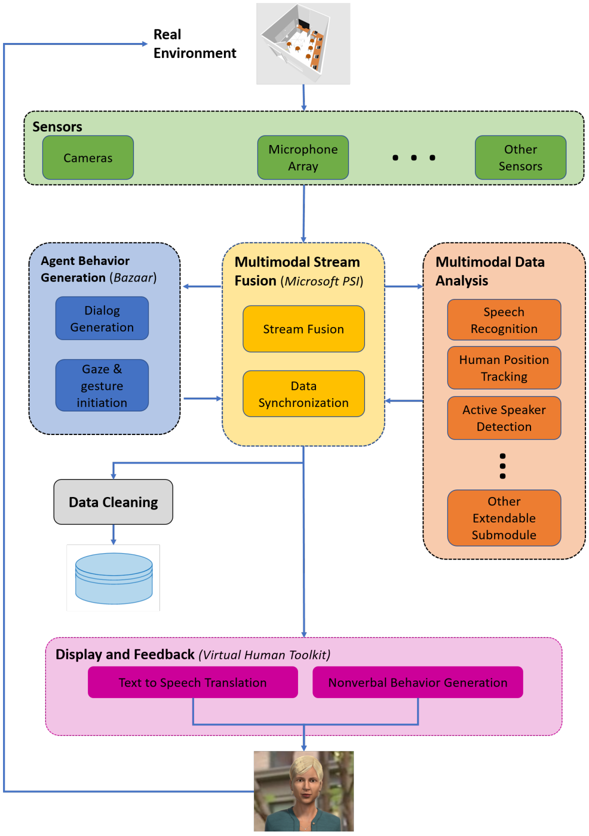
Real Environment
Sensors
Cameras
Microphone Array
Other Sensors
Agent Behavior Generation (Bazaar)
Dialog Generation
Gaze & gesture initiation
Multimodal Stream Fusion (Microsoft PSI)
Stream Fusion
Data Synchronization
Multimodal Data Analysis
Speech Recognition
Human Position Tracking
Active Speaker Detection
Other Extendable Submodule
Data Cleaning
Display and Feedback (Virtual Human Toolkit)
Text to Speech Translation
Nonverbal Behavior Generation

Figure 2: The software infrastructure.

2.1.1 Information Flow

Information flow for operating the Smart Office Space is displayed in Figure 2. As we develop the Smart Office Space, we are first focusing on using the audio and video data provided by the Lorex cameras and the Kinect microphone array to communicate with users via VHT. The data captured by the cameras and the microphone array are sent in separate streams to PSI. PSI passes the streams to audio and video recognizers. As the recognizers detect events, they pass event messages back to PSI: Visual information is translated to semantic text describing body position and facial expressions; location information is translated to polar coordinates; and speech is translated to text. Recognition of the various audio and visual events may occur at different speeds, so PSI may receive event messages out of order. PSI therefore synchronizes event messages by their originating time, then passes on the translated events as appropriate: User location changes are passed directly to VHT to update agent gaze direction with low latency; speech translations along with user locations and visual events like a raised hand are passed to Bazaar; and some events are discarded, such as recognition that the agent itself is speaking. Messages between PSI and Bazaar use an internally developed multimodal message format that associates a user identifier with any combination of the following easily-expandable list of user attributes: location, speech text, body position, facial expression, and any detected emotion. Bazaar uses the information it receives from PSI to decide when and how to respond to events in the room, passing response messages through PSI, which coordinates verbal and nonverbal communication, to VHT for communication with users as a virtual agent.

2.1.2 Multimodal Stream Fusion

To handle multiple data streams, PSI (Bohus et al., 2017) provides a runtime environment for parallel, coordinated computation across data streams along with a set of tools for visualization, data processing and machine learning. We run PSI on Windows 10. PSI associates timestamps with the data it receives from video and audio streams and includes these timestamps as it passes the data on to the appropriate video and audio recognizers. When the recognizers detect events, they include the originating timestamps with the event messages that they return to PSI. PSI uses these timestamps to synchronize messages received on different streams, enabling it to identify both simultaneous events and the correct order of event sequences. PSI's messages to Bazaar combine synchronous audio and video events. For instance, PSI might combine video recognition that a user is speaking, audio and visual recognition of the user's location, and audio recognition of the user's words in a single message to Bazaar. In addition, PSI logs all data that it receives for playback, analysis, and offline machine learning.

2.1.3 Multimodal Data Analysis

We incorporate multiple video and audio recognizers to process the video and audio streams received through PSI. Video recognizers run on a Linux GPU server for faster processing of neural network models. We use OpenFace (Amos et al., 2016) to find and recognize people facing any of the four cameras, including recognizing whether two face inputs are the same person. To detect body position and key body points, we use OpenPose(Cao et al., 2017). OpenPose forwards its body points for the nose and neck to a location detector which maps these to lines in real space to triangulate users' locations. For location verification, we are currently using a Kinect microphone array and we plan to try adding inputs from an Intel RealSense depth-sensing camera. For audio speech-to-text recognition, we are currently testing two packages integrated with PSI: CMU Sphinx (Lamere et al., 2003) and Microsoft Azure Speech to Text [1].

2.1.4 Agent Behavior Generation

Bazaar receives event updates from PSI and uses this information to decide exactly how and when to respond to events. For instance, when a user enters the room, PSI sends Bazaar a message specifying a newly created internal user identifier along with the user's location within the room, specified in terms of polar coordinates. Bazaar saves this information as the beginning of its user model. As additional information is acquired about the user – including spoken words as text, body position, facial expression, and apparent emotion – PSI sends event updates and Bazaar updates its user model accordingly. Bazaar's responses can be tailored to the context. For example, if Bazaar wants to respond to an assertion by prompting the user to explain her reasoning, it can identify the user by associating the location of the speech source with the user's saved location, call up the user's name, and respond to the user by name while gazing in her direction.

2.1.5 Communication to Users

To communicate to users both verbally and nonverbally, Bazaar sends messages through PSI to VHT (Hartholt et al., 2013). VHT's display to the user can be designed to represent a 3-dimensional setting with one or more actors that communicate to

[1] https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/

Figure 3: Virtual tutor

users using speech, facial expressions, gaze directions, body position, and gestures. Speech is sent to VHT as text while non-verbal behavior is specified using the Behavior Markup Language (BML) realization library, “Smartbody”(Feng et al., 2012). At this stage, we communicate nonverbally using facial expressions, gaze direction, and simple arm and hand gestures. Facial expressions are specified in terms of lips, brows, and eyes, while gaze direction is realized through coordinated rotation of the shoulders, neck, head and eyes. We use these nonverbal cues to present some common non-verbal expressions – neutral, listening, confused, angry, happy, and amazed – and to gaze at individual users. For instance, if user Ron has offered an idea and Joan has not contributed to the ongoing group discussion in a while, the VHT may turn towards Joan and say, “Joan, can you build on the idea that Ron has offered?” Using the Smart Office Space, we are working towards collecting multiple datasets in collaboration with industry partners who help inform the characteristics of workplace scenarios for our studies including support for maintaining social distancing during intensive collaborative learning.

3 Demo Session

The video presentation of the demo for the online demo session will display scenarios in which groups of individuals work together on a task, with the VHT providing guidance for task structuring and collaborative work processes. What makes the demo unique among other applications of in person multi-party dialogue is the use of the virtual human as a group learning facilitator, enabled through the Bazaar architecture.

Acknowledgments

This work was funded in part by NSF grants IIS 1822831 and 1917955 and funding from Microsoft.

References

David Adamson, Gregory Dyke, Hyeju Jang, and Carolyn Penstein Rosé. 2014. Towards an agile approach to adapting dynamic collaboration support to student needs. *International Journal of Artificial Intelligence in Education*, 24(1):92–124.

Vincent Aleven and Judy Kay. 2016. *International Journal of AI in Education, 25th Anniversary Edition, volume 26(1-2)*. Springer.

Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science.

Dan Bohus, Sean Andrist, and Mihai Jalobeanu. 2017. Rapid development of multimodal interactive systems: A demonstration of platform for situated intelligence. In *ICMI 2017 Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 493–494.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299.

Jaime R Carbonell. 1970. Ai in cai: An artificial-intelligence approach to computer-assisted instruction. *IEEE transactions on man-machine systems*, 11(4):190–202.

Andrew Feng, Yazhou Huang, Yuyu Xu, and Ari Shapiro. 2012. Automating the transfer of a generic set of behaviors onto a virtual character. In *International Conference on Motion in Games*, pages 134–145. Springer.

Arno Hartholt, David Traum, Stacy C Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. 2013. All together now. In *International Workshop on Intelligent Virtual Agents*, pages 368–381. Springer.

Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. 2003. The cmu sphinx-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, volume 1, pages 2–5.

Charles Lang, George Siemens, Alyssa Wise, and Dragan Gasevic. 2017. *Handbook of learning analytics*. SOLAR, Society for Learning Analytics and Research.

Carolyn Penstein Rosé and Oliver Ferschke. 2016. Technology support for discussion based learning: From computer supported collaborative learning to the future of massive open online courses. *International Journal of Artificial Intelligence in Education*, 26(2):660–678.