Embodied Conversational AI Agents in a Multi-modal Multi-agent Competitive Dialogue

Rahul R. Divekar[1] , **Xiangyang Mou**[1] , **Lisha Chen**[1] ,
Maira Gatti de Bayser[2] , **Melina Alberio Guerra**[2] and **Hui Su**[1,2]
[1]Rensselaer Polytechnic Institute
[2]IBM Research
{divekr, moux4, chenl21}@rpi.edu, {mgdebayser, melinag}@br.ibm.com, huisuibmres@us.ibm.com

Abstract

In a setting where two AI agents embodied as animated humanoid avatars are engaged in a conversation with one human and each other, we see two challenges. One, determination by the AI agents about which one of them is being addressed. Two, determination by the AI agents if they may/could/should speak at the end of a turn. In this work we bring these two challenges together and explore the participation of AI agents in multi-party conversations. Particularly, we show two embodied AI shopkeeper agents who sell similar items aiming to get the business of a user by competing with each other on the price. In this scenario, we solve the first challenge by using headpose (estimated by deep learning techniques) to determine who the user is talking to. For the second challenge we use deontic logic to model rules of a negotiation conversation.

1 Introduction and Motivation

We have developed a conversational setting in which two AI agents playing the role of shopkeepers who sell similar items want to get the business of a user. AI agents do this by competing on the price. They are referred to A1 and A2 hereon. This scenario is developed to teach Chinese as a foreign language and culture through conversational role play with embodied AI. Enthusiastic readers are encouraged to see [Allen *et al.*, 2019] [Divekar *et al.*, 2018c] and [Divekar *et al.*, 2018a]'s work to read more about the context of the project.

Part of learning the new culture is learning to negotiate with street-market vendors which is uncommon for our users. To build AI agents that can participate in such a conversation, we explore how turn taking would work in a situation where multiple agents and humans are engaged in negotiations.

In any conversational setting, the first challenge towards determining whether an AI agent should respond is to determine if it is being addressed. It has been shown that the common practice of using a wake-word while talking to AI is not preferable in long conversations by [Divekar *et al.*, 2019]. Research in multi-modal addressee detection by [Ravuri and Stolcke, 2015], [Tsai *et al.*, 2015], [Sheikhi and Odobez, 2015], [Akhtiamov *et al.*, 2017], [Le Minh *et al.*, 2018] and [Norouzian *et al.*, 2019] has inspired us. It is a premise of the interactive aspect of our demo that it is common for people to look at the AI agent that they are speaking to especially when the AI agent is embodied as an animated avatar. As in [Divekar *et al.*, 2019]'s work, our system uses headpose as a primary determiner of addressee. It coupled with visual feedback from the agent to make the interaction smoother. Their system uses a facial landmark based approach for headpose detection. Our environment's lack of lighting and unusual camera position throw additional challenges. Here traditional facial landmark based estimation techniques fall short. Hence we use a deep learning approach to tackle this shortcoming. Details of the challenge and solution are described in Section 2.3. The addressee is determined by calculating the time overlap between the user's headpose intersection with the embodiments of the AI and the user's speech.

It is usually straightforward that once an addressee is clearly determined, the addressee must respond. However, addressee detection alone cannot trigger the non-addressed AI agents to participate in the conversation thereby making the agents reactive to users input rather than proactive. In a competitive setting, it is essential for the agents to be proactive in pitching their sale. Yet, they must not reply to every turn to the extent of being annoying. They must also not just talk with each other and leave the user out of the conversation. Therefore, they require a more complex set of rules that govern the conversation in order to determine the answer to the second challenge, i.e. when should the AI agent respond. [Andrist *et al.*, 2016] and [Khouzaimi *et al.*, 2016] have motivated the problem of turn-taking in AI. For our conversational setting, we explore the potential of using deontic logic to model rules of turn taking as previously shown by [de Bayser *et al.*, 2018b] and [de Bayser *et al.*, 2018a]. They have modeled social rules of multi agents but in collaboration conversations. We use their tool to model rules we wrote for competitive agents as shown in Section 2.2.

The interdependence of addressee detection and rules of turn-taking, specifically, in our said scenario is clear by the following example —

Situation 1: Addressee and thus speaker is determined by headpose.

User: (looking at A1) How much for water?

A1: $5

A2: No response (Erroneous: Rules of competition are not understood)

Situation 2: Speaker is determined by turn taking rules

A1: I can do $5

User: (looking at A2) Can you do better?

A1: Yes I can do $4 (Erroneous: User meant to talk to A2. System did not consider headpose to ascertain addressee)

We thus integrate headpose based addressee detection and deontic rules for a negotiation conversation to create a more intelligent interaction. We show a proof of concept in which two agents can successfully compete with each other and have a conversation with the user.

2 Technologies Involved

2.1 Dialogue and Integration

User's voice input is transcribed by Automatic Speech Recognition (ASR) and tagged with an addressee based on which agent was looked at more by the user while speaking. Then, each AI agent generates output text based on the intent detected from the utterance and the state of the dialogue tree following [Divekar *et al.*, 2018b]'s architecture. Whether the output text gets broadcasted/spoken will be decided by Ravel (tool to model social rules) described in Section 2.2.

It can be seen from the two scenarios in Section 1 that the two technologies (addressee detection and turn taking rules) can provide conflicting results. One way to solve such conflicts is to convert the output from the headpose-based addressee module to text that signifies addressee (e.g. @A1). Then, instead of separately using headpose and Ravel to determine whether a turn should be allowed or not and then breaking the tie, use this headpose result as an input to Ravel. Ravel allows us to apply rules about what should happen in case an addressee is detected.

2.2 Norm Specification Using Deontic Logic

Ravel maintains a Finite State Machine (FSM) representation of the conversation. Rules can be applied on the state transitions. Every incoming utterance (human and machine) is classified into an intent and gets tagged with it. Ravel decides whether the intent/utterance has a valid transition from the current state i.e. decides whether the agent that generated the utterance is *obligated*, *allowed* or *prohibited* to respond with that intent. If the agent is *obligated* or *allowed*, the system *broadcasts* the message to all participants by using JSON messages for AI agents and voice output for the user. Each agent receives the broadcasted output as input and generates a response which follows the same loop. If the agent is *prohibited* then its response is *blocked*.

We crafted the following rules. Their application can be seen in Table 1.

R1: User is always *allowed* to reply.

R2: AI Agents are *prohibited* from responding to themselves.

R3: If direct addressee detected, the direct addressee has the *obligation* to respond. Other AI agents are *prohibited*

R4: AI Agents are *allowed* to respond to a price pitch

To further show the applicability of rules, we made our agents agnostic to message sender. Thus they try to out-bid themselves (Turn 5 in Table 1). Our defined social rules block this utterance and add intelligence to the interaction.

Sender	Utterance	Status	Rule
User	@A1 Do you have water?	Broadcast	R1
A1	I will give it for $5	Broadcast	R3
A2	I will give it for $4	Block	R3
A1	I can do a better price	Block	R2
A2	I can do a better price	Broadcast	R4

Table 1: Application of Rules to Dialogue Turns

2.3 Head Orientation Estimation Using Deep Learning Techniques

The headpose estimation system takes image input from cameras to detect and track a face, detect facial landmarks and estimate headpose based on those landmarks. Using cameras enables non-intrusive markerless interactions. In our environment[1], the camera is constrained to be on the ground in a low-light condition (used to accentuate displays) and the users stand more than 3 meters from the camera, causing a low resolution face. Further, the position of the face w.r.t. the camera causes large pitch pose which affects the accuracy of even the state-of-the-art landmark detection algorithms trained on benchmark dataset [Bulat and Tzimiropoulos, 2017].

We therefore combine a generative model [Zhu *et al.*, 2019] and a probabilistic deep model [Chen and Ji, 2018]. Specifically, frontal faces captured in the environment are annotated, then large pose faces along with their landmark annotations are generated to fine-tune the probabilistic model [Chen and Ji, 2018] for facial landmark detection.

To calculate headpose, we assume a weak perspective projection model, where we have a 3D mean face shape $\bar{\mathbf{y}}_{3d}$, a 3D rotation matrix R, translation vector T and a camera intrinsic matrix W obtained from camera calibration. Given the detected 2D landmark points $\mathbf{y}_{2d}$, we estimate headpose by minimizing the weighted projection error, i.e. $R^*, T^* = \arg\min_{R,T} \|\mathbf{y}_{2d} - \frac{1}{\lambda} W[R,T]\bar{\mathbf{y}}_{3d}\|_C^2$ (in homogeneous coordinate), C consists of the inverse of the determinant of the predicted covariance for facial landmarks. Headpose is obtained from the rotation matrix R^*. The estimated headpose and translation T w.r.t. the camera coordinate is then transformed to the room coordinate using the camera extrinsic matrix. The probabilistic model quantifies uncertainty to avoid over-confident erroneous predictions, i.e. we reject predictions with corresponding uncertainty above threshold.

3 Conclusion and Future Work

We show the integration of headpose-based addressee detection and turn-taking rules in a negotiation conversation between two AI agents and one human. With this demo, we can give culture/language learners an opportunity to practice negotiation skills. We plan to conduct experiments to evaluate its effectiveness. This demo will be used to further understand the rules of a conversation through various approaches e.g. machine learning and, explore ways to empower the agents with stronger negotiation strategies in multi-agent settings.

[1] Demonstration Video - https://youtu.be/z6CJJ3ig8Hs

References

[Akhtiamov *et al.*, 2017] Oleg Akhtiamov, Maxim Sidorov, Alexey A Karpov, and Wolfgang Minker. Speech and text analysis for multimodal addressee detection in human-human-computer interaction. In *INTERSPEECH*, pages 2521–2525, 2017.

[Allen *et al.*, 2019] David Allen, Rahul R Divekar, Jaimie Drozdal, and Lilit Balagyozyan. The Rensselaer Mandarin Project—a Cognitive and Immersive Language Learning Environment. In *Thirty-third AAAI Conference on Artificial Intelligence*, 2019.

[Andrist *et al.*, 2016] Sean Andrist, Dan Bohus, Bilge Mutlu, and David Schlangen. Turn-taking and coordination in human-machine interaction. *AI Magazine*, 37(4):5–6, 2016.

[Bulat and Tzimiropoulos, 2017] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[Chen and Ji, 2018] Lisha Chen and Qiang Ji. Kernel density network for quantifying uncertainty in face alignment. In *3rd Bayesian Deep Learning Workshop of Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[de Bayser *et al.*, 2018a] Maira Gatti de Bayser, Melina Alberio Guerra, Paulo Cavalin, and Claudio Pinhanez. Specifying and implementing multi-party conversation rules with finite-state-automata. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[de Bayser *et al.*, 2018b] Maira Gatti de Bayser, Claudio Pinhanez, Heloisa Candello, Marisa Affonso, Mauro Pichiliani Vasconcelos, Melina Alberio Guerra, Paulo Cavalin, and Renan Souza. Ravel: A MAS orchestration platform for human-chatbots conversations. 2018.

[Divekar *et al.*, 2018a] Rahul R Divekar, Jaimie Drozdal, Yalun Zhou, Ziyi Song, David Allen, Robert Rouhani, Rui Zhao, Shuyue Zheng, Lilit Balagyozyan, and Hui Su. Interaction challenges in AI equipped environments built to teach foreign languages through dialogue and task-completion. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018*, pages 597–609. ACM, 2018.

[Divekar *et al.*, 2018b] Rahul R Divekar, Matthew Peveler, Robert Rouhani, Rui Zhao, Jeffrey O Kephart, David Allen, Kang Wang, Qiang Ji, and Hui Su. CIRA: An Architecture for Building Configurable Immersive Smart-Rooms. In *Proceedings of SAI Intelligent Systems Conference*, pages 76–95. Springer, 2018.

[Divekar *et al.*, 2018c] Rahul R Divekar, Yalun Zhou, David Allen, Jaimie Drozdal, and Hui Su. Building human-scale intelligent immersive spaces for foreign language learning. *iLRN 2018 Montana*, page 94, 2018.

[Divekar *et al.*, 2019] Rahul R Divekar, Jeffrey O Kephart, Xiangyang Mou, Lisha Chen, and Hui Su. You talkin' to me? - a practical attention-aware embodied agent. In *Human-Computer Interaction – INTERACT 2019*, 2019.

[Khouzaimi *et al.*, 2016] Hatim Khouzaimi, Romain Laroche, and Fabrice Lefèvre. Reinforcement learning for turn-taking management in incremental spoken dialogue systems. In *IJCAI*, pages 2831–2837, 2016.

[Le Minh *et al.*, 2018] Thao Le Minh, Nobuyuki Shimizu, Takashi Miyazaki, and Koichi Shinoda. Deep Learning Based Multi-modal Addressee Recognition in Visual Scenes with Utterances. *IJCAI 2018*, pages 1546–1553, 2018.

[Norouzian *et al.*, 2019] Atta Norouzian, Bogdan Mazoure, Dermot Connolly, and Daniel Willett. Exploring attention mechanism for acoustic-based classification of speech utterances into system-directed and non-system-directed. *arXiv preprint arXiv:1902.00570*, 2019.

[Ravuri and Stolcke, 2015] Suman Ravuri and Andreas Stolcke. Recurrent neural network and LSTM models for lexical utterance classification. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[Sheikhi and Odobez, 2015] Samira Sheikhi and Jean Marc Odobez. Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions. *Pattern Recognition Letters*, 66:81–90, 2015.

[Tsai *et al.*, 2015] TJ Tsai, Andreas Stolcke, and Malcolm Slaney. A study of multimodal addressee detection in human-human-computer interaction. *IEEE Transactions on Multimedia*, 17(9):1550–1561, 2015.

[Zhu *et al.*, 2019] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3D total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):78–92, 2019.