Abstract
Competitions that directly pit software agents against one another have proven to be an effective and entertaining way to advance the state of the art in a multitude of AI domains. Less frequently, human-agent competitions have been held to gauge the relative competence of humans vs. agents, or agents vs. agents as measured indirectly by their performance against humans. We are developing a platform that supports a new type of AI competition that involves both agent-agent and human-agent interactions situated in an immersive environment. In this competition, human buyers haggle (in English) with two life-size AI agents that attempt to sell them various goods. We describe several research challenges that arise in this context, present the platform architecture and accompanying technologies, and report on early experiments with simple agents that establish feasibility and suggest that human participants enjoy the experience.

CCS Concepts
• Human-centered computing → Natural language interfaces; Gestural input; Usability testing; Interaction design; Computing methodologies → Multi-agent systems;

Introduction
For at least the past two decades, competitions that directly pit two software agents (or two teams of software agents)
against one another have been an effective and entertaining way to advance the state of the art in a multitude of AI domains, such as robotic soccer and agent-based electronic commerce [36]. To a lesser degree, competitions have been held to assess the relative competence of humans and agents [11] or to evaluate which agents perform best against humans [28].

There are natural scenarios in which interactions among humans and multiple agents are of interest. An example we draw inspiration from is an educational scenario in which students practice Mandarin Chinese language and culture through spoken role-play with embodied AI agents in an immersive environment. Initial studies of AI-assisted language education [3, 13, 17] had shown that immersion has a beneficial impact. In the Mandarin education scenario, the agents play various roles, including shopkeepers who compete with each other for the student’s business [15]. As we tried to develop competent negotiation strategies that would engage the students, we realized that, with suitable extensions, our platform could be used as a basis for a new type of AI competition that blends aspects of agent-agent and human-agent interactions, and brings those interactions to life by situating them in an immersive environment (Fig. 1). The interactions also have the potential of letting users compare complex products/services in new ways.\(^1\)

With the support of members of the ANAC (Automated Negotiating Agents Competition) board (web.tuat.ac.jp/~katfuji/ANAC2019), we are preparing an international AI competition in which buyers haggle (in English) with two life-size AI agents that attempt to sell them various goods. After describing several research challenges that arise in this context and a brief literature review, we present the platform architecture and accompanying technologies. Finally, we report on early experiments with simple agents that establish feasibility and suggest that human participants find the experience enjoyable.

---

\(^1\) E.g. users could interact with agents that represent different products (e.g. competitor car brands) or perspectives on the same product (e.g. style consultant, engineer).

---

**Author Keywords:**
Agent competition; Immersive environment; Multimodal dialogue; Multiparty dialogue; Negotiation; Dialogue Systems; Mixed Reality

---

**Literature Review**

Two essential challenges that arise in the context of non-dyadic interactions among humans and agents include how an agent can know a) when it is being addressed and b) when it may speak.

Several authors, including [33], [38], [35], [1], [25], [5] and [31]) have sought means of determining the addressee without resorting to a wake word by means of various multimodal cues such as intonation, pitch, head-gaze, vocal energy, etc. to determine the addressee in human-kiosk, human-robot, human-human, and human-human-agent conversations. [23] [35], [30], [20] and many others have addressed this problem in the Human Robot Interaction
field using approaches such as identifying visual focus of attention or moving the robot’s head to signify turns. Recent work [14] shows that a simple approach based on head pose coupled with semiotics of inferred user attention by the avatar may suffice. Encouraged by this result, we adopt a similar approach but employ a different head pose recognition algorithm [10] that copes better with the low light and high pose angle that typifies our environment (Fig. 1).

Multi-party, multi-modal conversations have been studied from various perspectives (e.g. linguistics, algorithms) and embodiments (kiosk based, etc.) [4], [24], [19], [7], [37], [2], [32]; but rarely in a multi-agent competitive, immersive setting with natural interactions. We find the recent work of Gatti de Bayser et al. [18], [12] to be most appropriate for our use case. Their approach based on deontic logic explicitly models turn-taking in conversations involving humans and multiple AI agents. It can enforce rules on the structure of the conversation without requiring individual agents to understand or implement the model themselves.

AI competitions that pit software agents against one another have a long history that includes annual events with multiple competitive leagues, such as the RoboCup robotic soccer competition (https://www.robocup.org/), the Trading Agent Competition (https://strategicreasoning.org/trading-agent-competition/), and the Automated Negotiating Agents Competition (web.tuat.ac.jp/~katfuji/ANAC2019) [26, 22, 29]. Among the smaller set of human-agent competitions in existence is the ANAC human-agent league [27], wherein humans negotiate with a single agent using a desktop-based interface that allows them to input structured text supplemented with emojis to indicate emotion. As far as we are aware, ours is the first competition that features direct simultaneous negotiations between a human and multiple competing agents. Our competition is further distinguished by occurring in an immersive environment that incorporates multi-modal interactions involving speech understanding, speech synthesis, and addressee detection based upon head pose estimation – all of which combine to support a more natural form of interaction between humans and software agents thereby attempting to provide a sense of realism, visual and social presence (defined as sense of being with another intelligent entity [6]).

**Negotiating in an Immersive Environment**

The purpose of the immersive environment is to provide audio-visual immersion and presence, i.e. a feeling of being in a different place. Fig. 1 illustrates the 360-degree panoramic screen that is used to provide visual immersion; audio immersion is provided via spatial audio techniques [9, 8] that enable one to control the apparent location of sound sources. While similar physical immersive systems exist [21, 34], they tend to be used for other situational contexts such as combat simulation.

The screen depicts a virtual street scene in Shanghai (implemented in Unity Game Engine) inhabited by two street vendor avatars. Wearing a lapel microphone, the user looks at the avatar with whom they wish to speak. A central system transcribes the speech and infers the addressee using head pose information, and forwards this information to all of the agents. The central system maintains decorum and fairness by using a Finite State Automaton to enforce certain predetermined turn-taking rules, including ones that specify when interjections are permitted.

During each of several rounds in the competition, a human buyer starts by stating which goods they are interested in purchasing (e.g. eggs, milk, sugar, flour, chocolate) in an effort to acquire ingredients to complete a task (e.g. bake cakes). Agents may choose to respond to such requests.
with offers consisting of a bundle of ingredients and a price. Such offers are rendered as synthesized speech in such a way as to appear to emanate from that agent. The human may respond with a counter-proposal, and so on until an agreement is reached. While humans may primarily direct their attention to one agent, all agents are aware of all negotiation messages exchanged in the system, and (under conditions regulated by a Moderator) they may interject with convincing counter-arguments/offers, possibly causing the human to re-direct their attention. An accompanying video\textsuperscript{3} illustrates such interactions. Agents are evaluated quantitatively according to their total financial gain, while humans are evaluated according to a utility function that depends on the number and quality of cakes that can be assembled from their ingredients. Agents may also be judged qualitatively according to their perceived degree of engagement.

**Architecture and Technical Details**

Extended from [16] which was limited to single agent, wake-word based interactions, Fig. 2 shows the overall architecture that enables the interactions described above. Each module in the diagram can reside on a different machine and can communicate using publish-subscribe software (RabbitMQ) and RESTful API.

In step 1, the Mic and Cameras detect raw input. The mic connects to a commercial cloud-based Automatic Speech Recognition (ASR) service. The machine connected to the camera processes images to detect head pose according to a method described by [10]. In step 2, the text utterance and head pose coordinates are sent to the Attention Manager, which infers the addressee as the agent towards which the human was looking primarily over the course of the utterance (Agent 1 in our example). In step 3, the Attention Manager forwards the utterance $U$, the speaker $S$, and the inferred addressee $A$ to the Moderator, which updates the global conversation state and if the utterance is allowed, forwards it to the Agent Executors, which generate dialogue and visual actions of the avatars that embody the agents.

The architecture permits participant developers to submit independently-written Agent Executors that understand the messages generated by the system, compute actions (offers, acceptances, etc.) according to their strategies (for instance, bidding strategy), and speech utterances that represent those actions. The only other requirement is that the agents provide a list of possible responses ahead of time to enable the intent classifier of the Moderator to be trained.

Both Agent Executors receive the utterance, speaker and addressee, and they may choose to generate a proposed utterance. The architecture places no restrictions on the utterance or on the means by which it is produced. In our current implementation of the agents, they use an intent-entity based dialogue engine along with local conversation state variables to select a dialogue node. They use a naive negotiation strategy in which the default bidding agent's behavior simply decrease the last heard bid by a predefined amount until a predefined lower limit is reached. Agent Executors may propose any response they want to speak and pass it to the moderator in step 4. For example, we shall use the example of Table 1: after the user said “I want to buy tea” while looking mostly at Agent 1, the Agents 1 and 2 might propose “Yes, for $5” and “Yes, for $6”, simultaneously and pass it to the Moderator.

The Moderator is a centralized controller that regulates the interaction, and protects against either of the agents unfairly (or annoyingly) hijacking the interaction. It has its own separate intent classification engine. This helps maintain a global notion of intents that may differ from that of the agents. It consults the classification engine in steps 4.1 and 4.2 to get the intent of each proposed utterance. This

\textsuperscript{3}Interaction Demo Video: https://youtu.be/KBV9z9IAD0
Table 1: Sample Dialogue - Multi-party turn taking. T: Turn; S: Sender; H: Head pose

<table>
<thead>
<tr>
<th>T</th>
<th>S</th>
<th>H</th>
<th>Utterance</th>
<th>Status</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>U</td>
<td>A1</td>
<td>I want to buy tea</td>
<td>Broadcast</td>
<td>R1</td>
</tr>
<tr>
<td>t1.1</td>
<td>A1</td>
<td>Yes, for $5</td>
<td></td>
<td>Broadcast</td>
<td>R3</td>
</tr>
<tr>
<td>t1.1</td>
<td>A2</td>
<td>Yes, for $6</td>
<td></td>
<td>Block</td>
<td>R3</td>
</tr>
<tr>
<td>t1.2</td>
<td>A2</td>
<td>I can give it to you for cheaper</td>
<td></td>
<td>Broadcast</td>
<td>R4</td>
</tr>
</tbody>
</table>

AI Agents are allowed to respond to a price pitch. Thus, in Table 1, Agent 1’s pitch (t1.1, A1) passes through while that of Agent 2 is blocked (t1.1, A2). The blocking helps from agents speaking over one another. A2 will be allowed in the next turn to counter-reply A1’s accepted utterance. The Moderator, in step 5 (Fig. 1), passes the allowed utterance on to the appropriate avatar, which is rendered using the Unity Engine and a commercial text-to-speech engine in conjunction with the spatial audio system [8]. The Moderator also sends the accepted utterance to both the Agent Executors. Note that this time the attention manager can be skipped as the addressee is the entire room. Here, the agents upon hearing each others’ (or their own) bids may continue the cycle afresh by proposing a counter-pitch (seen in t1.2, A2). The counter pitches are received by the Moderator, at most one is allowed, and so the cycle continues until an agreement is reached. The architecture is
imaginably scalable to more participants and conversational contexts\(^4\).

**Discussion from various perspectives**

**Administrator’s Perspective**
An administrator’s role is split into two parts: pre-competition preparation and in-competition facilitation. Prior to the start of the competition, she requests and collects all potential phrase variations that the agents may utter and submits them to the *Moderator* and *Ravel*. During the competition itself, she uses a web-based-backend UI to generate rounds, indicate start/end of competition to parties, generate utility functions that provide incentives for agents and human competitors and, validate the final offers\(^5\).

**Participant Developer Perspective**
At the beginning of a round, the agents receive fresh utility functions from the Admin, which serve as incentives that drive their negotiation behavior. We expect that the competition will provide wide scope for research on both the strategic and psychological aspects of negotiation with humans; for example the agents may try to gain an advantage by expressing their bids in an engaging or attractive way, or by dising other agents. As a first test of the platform, two external groups successfully created agents in a pilot test. Detailed documentation and sample code will be made publicly available well in advance of the competition.

**Human Negotiator Perspective**
Two in-house agents were employed with the main purpose of role-playing haggling with users to learn a foreign language. We used this as a pilot to test whether the proposed competition sufficiently engages human participants. 13 college students (6 female, 7 male) participated. Prior to the study, they were told how to direct an utterance towards an agent using head pose. They were *not* told that other agents might interject even when not addressed. We evaluated the interaction using a post-experience questionnaire.

To judge the overall experience, we asked users to rate on a Likert scale of 1-5 whether they agreed that the interaction was usable and (in a separate question) likable. The responses to usability had a mean of 4.08 ± 0.86, while the responses to likability had a mean of 4.38 ± 0.65. A one-sample test revealed that with p=0.003 and p=0.0002 we could say that the true median for usability and likability was greater than 3 (neutral). We also asked them to rate the appropriateness of agent’s turn taking on a Likert scale. The mean score for responses to it was 4.36 ± 0.51. We did a One-sample Sign-test on the data and with confidence of p=0.0048 we could say that the true median was greater than 3 (neutral). Overall, we find that the interaction closely matched the users’ natural expectations and thus the design appears to suffice for the competition.

**Conclusion and Future Work**
We have provided an overview of the architecture and technology underlying a new AI competition that we hope to hold as part of ANAC at an upcoming AI conference: IJ-CAI 2020. Our preliminary evaluations with simple agents and a small number of human participants indicate that the platform supports independently-programmed agents and that humans find the experience engaging. Such conversational interactions may augur a new way for businesses to advocate their wares and users to compare attributes of complex goods or services. We hope and believe that this multi-agent, multi-modal negotiation platform and the competition it supports will — as RoboCup and TAC (Trading Agent Competition) have done before it — spur new and interesting research in the realm of multi-lateral negotiation algorithms, multi-modal dialogue, and the human dynamics of such interactions.
REFERENCES


