11-866 Artificial Social Intelligence **Spring 2023**

Week 9: Conversational AI

Instructors: Louis-Philippe Morency, Paul Liang *Synopsis Leads: Zhengyang Qi, Ji Min Mun*

Edited by Paul Liang *Scribes: Santiago Benoit, Ryan Liu*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/asi-course/spring2023/

Summary: Artificial social intelligence (ASI) encompasses the research effort towards advancing AI systems that can perceive and interpret human social information (e.g., communication patterns, pragmatics, intent) and engage in seamless human-AI interactions.

In week 9's discussion session, the class focused on discussing use of language . The following was a list of provided research probes:

1. As a first step in reviewing research related to artificial social agents, we start with language-based interactions (aka non-embodied interactions, conversational agents). From our previous weeks discussing skills and abilities related to social intelligence and competence, what subset of these skills and abilities are primarily driven by language-based conversations? Is language central to all of them?
2. A big enabler of recent progress in this area of (language-based) conversational AI is related to large-scale language models (e.g., GPT, OPT, T5, TNLG) and their extensions to conversation modeling (e.g., DialogGPT, Blenderbot). What aspects of social intelligence are particularly well suited to be learned by these large-scale models? What social intelligence skills and abilities may not be well modeled by these large-scale models?
3. In parallel to this recent progress in large-scale models, researchers are still continuing to study more theoretically-inspired research in conversational modeling. What aspects of social intelligence are particularly well suited for these approaches, not based primarily on large-scale models? What social intelligence skills and abilities may not be well modeled by these approaches?
4. We start seeing hybrid approaches that integrate both theory-inspired approaches with large-scale models. How would you propose to integrate these two lines of research? Should large-scale models be the foundation of these integrated approaches? Or should large-scale models be seen just as a module of a larger system? Can you think of other integrative/hybrid approaches?
5. Evaluation of these conversational AI systems is still a big issue. Focusing on the social intelligence parts of the problem, how can we properly evaluate conversational AI systems? How should automatic metrics be used as part of the evaluation and creation of conversational AI systems? Similarly, how should we integrate human feedback?

As background, students read the following papers:

1. Socio-conversational systems: Three challenges at the crossroads of fields [Clavel et al., 2022]
2. Commonsense Reasoning for Conversational AI: A Survey of the State of the Art [Richardson and Heck, 2023]
3. Empathy and Prosociality in Social Agents [Paiva et al., 2021]
4. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs [Sap et al., 2022]
5. ValueNet: A New Dataset for Human Value Driven Dialogue System [Qiu et al., 2022]
6. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset [Rashkin et al., 2018]

We summarize several main takeaway messages from group discussions below:

1 Role of Language (Probe 1)

Language based interactions are generally more informative and text-based data is easy to find, which has led to the success of language-model based conversational AIs. However, in terms of social intelligence and the human feedback loop, language may not be the sole sufficient modality for a large-scale model to poses a high level of social intelligence. Non-verbal cues, such as body language, facial expression, and eye gaze, are sometimes more informative than language in terms of revealing the emotional status of human. Following from this, a social conversational AI should receive feedback from not only language but also the multi-modal information during a conversation with human and use human gestures as an additional set of rewards, or use the implicit responses from human to improve the original reward functions.

1.1 Language for Social Skills

Despite not being sufficient, some skills require language. Language is key to collaboration, engagement, ability to manage relationships, ability to memorize past conversations, social memory, and social learning, especially declarative knowledge. Language is still the best way, with more information with more efficiency in storage and communication, to clarify non-verbal or the latent variables in social intelligence.

2 Large Language Models (Probe 2)

2.1 Limitations from Human Feedback

Reinforcement Learning from Human Feedback (RLHF) has proved to be an effective way to train conversational AIs based on large language models, such as ChatGPT. Although such methodology has shown promising results on many current conversational models, human feedback has also its limitations, which may result in limitations in some perspectives of large language models [Sap et al., 2022].

The current human feedback (HF), usually in the form of thumbs up and down, tends to train LLMs majorly on how to avoid errors. Although HF can learn certain elements of human interactions, it does not explicitly contribute to the model's ability to reason about the correctness, i.e., ChatGPT having very confident incorrect answers. Moreover, in comparison to human's learning, LLMs can not be securely sure to remember certain social rules immediately and tend to forget what they were instructed. This may result in low social competence to unforeseen social situations. As a result, the improvement made by our current HF systems could taper to some point where input to HF does not proportionally improve social intelligence. As the summed-up effect of the current HF systems, LLMs possess a certain level of social intelligence but still needs improvement on social reasoning and adaptability, which is important social competence and skills.

3 Theory-inspired Approach (Probe 3)

In parallel to large language models, theoretically inspired approaches offer another set of advantages in conversational AIs. One of the biggest shortcomings of large language models is the uncertainty of responses. Human learns explicit rules quickly and is usually error prune, which is not an ability guaranteed for language models. Structural knowledge is also more difficult to teach to language models than human. Theoretically, the emergence of social intelligence for human is through a gradual process driven by individual needs, such as collaborating on a task, sharing feelings, and communicate thoughts, etc, while language models tend to get social intelligence through extensive exposure to human language.

The above observations led to a theory-inspired approach to use agent-based simulation to study social intelligence by how it satisfies evolution needs. An example would be placing a group of agents in an simulated environment that requires collaboration and group-level coordination to get rewards. Such agents may possess baseline-level understanding of their environment, means of communications, and primitive reasoning abilities, but unnecessarily social intelligence or social knowledge. The end result will be observing how the needs for communication may improve level of social intelligence in agents.

4 Hybrid Approaches (Probe 4)

To address the aforementioned problems of uncertainty in responses and difficulty in learning social rules, another approach is to mix language and multi-modal models to structural controls, such as behavioral trees. In certain cases where people would prefer more deterministic actions, such as game NPC, a better form of AI may be one that perceive social cues very well and performs a limited set of pre-designed actions according to the cues. The decision-making process to choose with one actions or the other may be simple and deterministic as long as it exhibits social intelligence.

Another approach is to draw inspirations from a neuroscience and cognitive science perspective. Specifically, we may connect neural models as modules to emulate how human brain functions and evolve. For example, we may use perceptive models as information receivers and language models as memory and connect them together. In a broader sense, more processes can be simulated, such as how procedural memory and declarative memory differentiates, how human learn by imitating others, and how memory is updated.

5 Evaluation (Probe 5)

How then should we evaluate these skills that require language? More specifically, how should we measure engagement, empathy, etc. in the agents or between humans?

One way in which these evaluations should be considered is through goal achievement [Rashkin et al., 2018]. Dialogues between humans occur to achieve certain goals as we have read in previous weeks; therefore, it is important to consider the different goals and consistency in identities and theory of mind of the conversational agents.

However, it is difficult to measure specific aspects of conversations in concrete and systematic ways. Looking at human evaluations such as specific testing methods used for social intelligence might be one way of trying to evaluate methods more systematically in an offline way.

References

Chloé Clavel, Matthieu Labeau, and Justine Cassell. Socio-conversational systems: Three challenges at the crossroads of fields. *Frontiers in Robotics and AI*, 9, 2022.

Ana Paiva, Filipa Correia, Raquel Oliveira, Fernando Santos, and Patrícia Arriaga. Empathy and prosociality in social agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, pages 385–432. 2021.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. Valuenet: A new dataset for human value driven dialogue system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11183–11191, 2022.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018.

Christopher Richardson and Larry Heck. Commonsense reasoning for conversational ai: A survey of the state of the art. *arXiv preprint arXiv:2302.07926*, 2023.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*, 2022.