Research Statement

Zhou Yu

Communication is an intricate dance, an ensemble of coordinated individual actions [3]. Imagine a future where machines interact with us like humans, waking us up in the morning, navigating us to work, or discussing our daily schedules in a coordinated and natural manner. Current interactive systems being developed by Apple, Google, Microsoft, and Amazon attempt to reach this goal by combining a large set of single-task systems. But products like Siri, Google Now, Cortana and Echo still follow pre-specified agendas that cannot transition between tasks smoothly and track and adapt to different users naturally. My research draws on recent developments in speech and natural language processing, human-computer interaction, and machine learning to work towards the goal of developing **situated intelligent interactive systems**. These systems can coordinate with users to achieve effective and natural interactions. I made a number of contributions in approaching this goal.

- I proposed the *situation intelligence framework* that uses a statistical policy learned via reinforcement learning methods that considers both system situation and interaction history in selecting system actions to enable machines to conduct coordinated, effective and natural long-term interactions.
- I have successfully applied the proposed framework to various tasks, such as social conversation, job interview training and movie promotion. My team's proposal on engaging social conversation systems was selected to receive $100,000 from Amazon Inc. to compete in the **Amazon Alexa Prize Challenge** [1].
- I implemented end-to-end systems and conducted user studies to iteratively improve the framework. Two dialog system frameworks were developed for computer science education and research (used in tutorials and courses), and deployed for real-world use.

The situation intelligence framework that I have developed will be the basis of my future research towards building the next generation interactive systems that are more efficient to develop, more trustworthy in practice, and more generalizable across domains.

1 Situation Intelligence Framework

I propose a ***situation intelligence*** framework to develop coordinated, effective and natural interactive systems. The framework has three key components: situation awareness, actions to coordinate the situation, and a statistical planning policy to select among actions to achieve coordinated long-term interaction.

Situation Awareness for Coordination. Empowering systems with *situation awareness* is the first step towards coordinated interaction. Situation awareness refers to the understanding of a system's situational contexts, such as location, time and users. My work focused on user modeling, as users are the system's "dance partners" and also final evaluators. I modeled a rich repertoire of user situations for systems with different goals. These include mental states (e.g. attention) in instructional tasks [22], interpersonal relationships (e.g. friendship) in group collaboration [15]; and backgrounds (e.g. cultural backgrounds) in social conversations [16].

Conversation Strategies for Understanding and Adaptation. I then designed *conversation strategies* to coordinate with detected system situation. Conversation strategies or discourse strategies refer to system actions that foster effective and natural interaction, such

[1] https://developer.amazon.com/alexaprize

as coordinating user state, assisting natural language understanding, creating adaptive user experience, etc. Drawing inspiration from conversation theories, I synthesized these strategies through natural language processing and knowledge base information.

Statistical Policies for Long-term Interaction Planning. The system also needs a policy or a plan to choose among these strategies. An interactive system's choice of action hinges on history, therefore history tracking is critical in policy design. Previous work attempted to incorporate history by using reinforcement learning methods to track a system's pre-specified states [12]. My work took this further by removing pre-specified states and learning an interaction policy that considers both interaction history and system situation to select among designed conversation strategies instead.

Situation intelligence framework creates situated intelligent interactive systems that are coordinated, adaptive, effective and natural in long-term interactions.

2 Applications in Conversation Systems

Users may leave conversations at any time; thus, tracking *user engagement* is essential. User engagement is defined as "a user's interest to continue the conversation" [9]. I propose an instantiation of the situation intelligence framework, the ***engagement-coordination*** framework to create coordinated long-term engaging conversation. I applied the framework to various types of conversations.

2.1 Social Conversation

Social conversation refers to everyday conversation without specific goals. Currently, personal assistants, such as Apple's Siri are in high demand. However, none of them can handle social conversations. The demand for such functionally is huge, as 20% of user utterances spoken to Microsoft's Cortana are social chats [6]. In addition, social conversation systems are not simply cool toys, but are also tools for social good, such as providing elderly people with social companionship. In this section, I will describe the concrete algorithms that provide the systems with engagement awareness, conversation strategies and statistical policies; and how these abilities work together to achieve engagement-coordinated social conversations.

Engagement Awareness. First, to track user engagement, I built a statistical model that leverages automatically quantified human behaviors from multiple modalities, such as speech volume in the audio modality and smiles in the vision modality. In a user study, I found Chinese college students disengaged more often than their American counterparts when the system took a long time to respond, perhaps because of a requirement for more immediate feedback due to the influence of their collectivist culture [16]. Based on this finding, I modeled user cultural backgrounds in a multi-lingual conversational system design. For instance, I modeled the system to produce hedging phrases (e.g. "let me see") to fill in the long silence for Chinese users. A following user study confirmed our findings as Chinese users rated the culturally adapted version more engaging [13]

Conversation Strategies. I also encode conversation theories in conversation strategies to foster user coordination, understanding and adaptation. In particular, I designed three types of strategies. *Active participation strategies* engage users by actively contributing to the conversation, such as asking more information on the current topic [10]. *Grounding strategies* assist open-domain natural language understanding. Grounding strategies were automatically synthesized via leveraging knowledge-base (e.g. Google Knowledge Graph) information and natural language processing algorithms, such as named-entity detection and statistical language generation. For instance, if an ambiguous named-entity "Clinton" is mentioned, the system would provide users with details of two popular "Clintons" from the knowledge base to collaboratively resolve the ambiguity with users. *Personalized strategies* adapt to users by leveraging automatically extracted information from individual user's conversation history. An

example personalized strategy is to suggest talking more about movies knowing the user was previously engaged in this topic.

Statistical Policies. Equipping systems with strategies is not enough to achieve long-term optimal outcomes. Therefore I introduced a statistical policy trained with reinforcement learning methods. It combined both conversational history and user engagement in the state design; and proposed conversation strategies and system responses (generated using an ensemble of sequence-to-sequence neural models [7] and information retrieval models [17]) in the action design. The reward function included both global metrics, such as conversation depth and variety, and local constraints derived from conversation theories, such as penalizing repeated usage of the same strategy. Thus the statistical policy is both tractable and globally optimal [22].

The long-term engagement-coordinated system elicited longer and more engaging social conversations compared to a rule based system [13]. The recognition of the system is broad, including wining a $100,000 research grant from Amazon Inc. for the Amazon Alexa Prize competition; and being selected as one of the six systems to participate in the shared tasks for data collection and labeling in two workshops collocated with conferences: LREC 2016 [4] and IVA 2016 [5].

2.2 Generalization to Task-Oriented Conversation

The engagement-coordination framework developed in social conversations are also generalizable to task-oriented conversation.

Job Interview Training. Job interview training is crucial for job hunting. However, effective and targeted training is expensive and cannot be easily accessed. Therefore, I designed a job interview training system to enrich user's interview experience and improve second-language learners' communication abilities. I designed the system following the engagement-coordination framework and added an extra type of conversation strategy: *positive feedback strategies* to coordinate user engagement, based on education literature. In a user study, users were found more engaging and provided more self-disclosure with the engagement-coordinated system. Also, statistics suggested that users significantly improved their communication ability after interacting with the system repeatedly [18]. The system was also deployed for classroom usage in countries with a high demand for English language learning, such as China, Japan and Brazil.

Movie Promotion. Another usage of social conversation is to act as a conversation smoother for complex tasks and at the same time to gauge user information that works in favor of the task goal. I designed a system that interleaves social content with task content, in particular, to promote a movie based on individual user's preferences. I also adopted the engagement-coordination framework in the design and expanded the reinforcement learning policy to provide a smooth transition between social and task contents. The system with the engagement-coordination was rated more engaging and elicited longer conversations [13]. These type of systems could be adapted to other tasks to achieve various goals, such as exercise reminding, health monitoring, and language learning.

3 Interactive System Implementation

I also built end-to-end systems and conducted user studies to iteratively refine proposed algorithms for different conversation systems. To support system development, I designed and implemented two conversation system frameworks: *TickTock* and *HALEF*. Both frameworks are open-source and compatible with various devices, including mobile phones. They also support multi-threading and cloud deployment.

TickTock contributed to research and computer science education. Both the social conversation system and the movie promotion system were based on it. TickTock is also used in the *CMU Dialog System Labs* course since 2015 and CMU LTI Capstone projects.

HALEF is another dialog system framework, I co-designed with ETS researchers [19]. The job

interview trainer was based on HALEF. It was also introduced in tutorials in various conferences and workshops, such as the SIGDIAL workshop in 2015 [2] and Cluecon conference in 2016 [3]; and used for various other tasks, such as presentation training, customer service and interruption study, and has collected tens of thousands of interactions with human users worldwide.

4 Future Directions

My research so far has leveraged recent developments across AI to create situated intelligent interactive systems that are adaptive, effective and natural. In the future, I plan to make these systems more efficient to build, more trustworthy in practice, more generalizable across domains and applicable to various areas, such as education, health care and entertainment.

Interactive Machine Learning. Modeling human-machine interaction is difficult because user actions depend on interaction history, and a small change could lead to an entirely different action sequence. Therefore, I propose to use *interactive machine learning*, which iteratively leverages human feedback to improve model performance at a faster pace [11], to make interaction modeling efficient. The performance of interactive learning hinges on good user interface design to enable people to provide more effective and accurate feedback, and powerful online learning methods to incorporate human feedback effectively.

Human-Machine Hybrid Systems. Trust is essential in high risk areas, such as health care and education. Therefore, I propose to design hybrid systems that combine automatic systems and crowd-powered systems [8] to make interactive systems more accountable. Hybrid system research hinges on how to combine automatic and crowd-powered systems. These systems can automate simple user requests and leverage human knowledge to solve complex requests in order to achieve a balance between costs and benefits. Therefore, the key research problem is to build better computation models to select system actions. The interaction history of these systems also provides more training data for further automation.

Domain Generalization. Interactive systems require domain knowledge to achieve various goals, thus they are hard to generalize across domains. To address this challenge, I propose to first use information extraction and knowledge inference methods to automatically learn domain knowledge structures from data. I will then use statistical generation and planning methods to encode the learned knowledge structures in system action and policy. The recent work on unsupervised dialog semantic slot induction is the first step towards leveraging knowledge bases to learn domain specified structures from unlabeled conversations [2]. Additionally, the recent development of structural sequence-to-sequence networks, such as long-short-term-memory (LSTM) [19] with attention models [1], have provided building blocks for developing advanced algorithms for automatic domain knowledge extraction and encoding in conversations.

Applications. Intelligent interactive system research is useful in various applications, such as health care, education and entertainment, and can be extended to robot design. I built virtual agents to facilitate clinical depression and PTSD assessments during my internship with Prof. Louis-Philippe Morency [20], and I intend to extend my research to collaborative therapies for dementia, aphasia and autism. I designed job interview training systems, in collaboration with ETS researchers, for educational purposes [20], and I plan to apply such system to learning at scale, such as assisting group discussions in Massive Online Open Courses (MOOCs). I also studied movie promotion systems for entertainment purposes [21] and wish to contribute to building automatic virtual agents in games. In addition to these virtual systems, I have also worked on physically embodied robots. For example, I worked on a direction-giving robot that coordinates user attention for effective communication, in a joint project with Eric Horvitz and Dan Bohus at Microsoft Research [14]. I want to extend this design to other collaborative robots, such as assistive robots for physical therapy and home robots for nursing.

[2] https://sites.google.com/site/yrrsdsmmxv/home
[3] https://cluecon.com/

References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *ArXiv preprint arXiv:1409.0473*, 2014.

[2] Y.-N. Chen, W. Y. Wang, and A. I. Rudnicky, "Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing," in *ASRU, 2013 IEEE*, IEEE.

[3] H. H. Clark, "Using language.," 1999.

[4] L. F. D'Haro, B. A. Shawar, and Z. Yu, "1st re-wochat share task report," in *RE-WOCHAT: Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents, LREC 2016*.

[5] L. D'Haro, B. A. Shawar, and Z. Yu, "2rd re-wochat share task report," in *RE-WOCHAT: 2nd Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents, IVA 2016*, p. 39.

[6] J. Jiang, A. Hassan Awadallah, R. Jones, U. Ozertem, I. Zitouni, R. Gurunath Kulkarni, and O. Z. Khan, "Automatic online evaluation of intelligent assistants," in *WWW*, ACM, 2015, pp. 506–516.

[7] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in neural information processing systems*, 2015, pp. 3294–3302.

[8] W. S. Lasecki, R. Wesley, J. Nichols, A. Kulkarni, J. F. Allen, and J. P. Bigham, "Chorus, a crowd-powered conversational assistant," in *UIST*, ACM, 2013.

[9] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi, "A model of attention and interest using gaze behavior," in *IVA , Kos, Greece*, 2005.

[10] D. Wendler, "Improve your social skills," *CreateSpace Independent Publishing Platform*, 2014.

[11] J. D. Williams, E. Kamal, H. A. Mokhtar Ashour, J. Miller, and G. Zweig, "Fast and easy language understanding for dialog systems with microsoft language understanding intelligent service (luis)," in *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, p. 159.

[12] J. D. Williams and S. Young, "Partially observable markov decision processes for spoken dialog systems," *Computer Speech & Language*, vol. 21, no. 2, pp. 393–422, 2007.

[13] Z. Yu, "Situated intelligent interactive systems," PhD thesis, Carnegie Mellon University, 2016.

[14] Z. Yu, D. Bohus, and E. Horvitz, "Incremental coordination: Attention-centric speech production in a physically situated conversational agent," in *SIGDIAL*, 2015, p. 402.

[15] Z. Yu, D. Gerritsen, A. Ogan, A. W. Black, and J. Cassell, "Automatic prediction of friendship via multi-model dyadic features," in *SIGDIAL*, 2013, pp. 51–60.

[16] Z. Yu, X. He, A. Black, and R. Alex, "User engagement modeling in virtual agents under different cultural contexts," in *International Conference on Intelligent Virtual Agents*, 2016.

[17] Z. Yu, A. Papangelis, and A. Rudnicky, "Ticktock: A non-goal-oriented multimodal dialog system with engagement awareness," in *AAAI Spring Symposium Series*, 2015.

[18] Z. Yu, V. Ramanarayanan, P. Lange, and D. Suendermann-Oeft, "An open-source multimodal dialog system with real-time engagement tracking for job interview training applications," in *ICASSP submitted*, 2016.

[19] Z. Yu, V. Ramanarayanan, R. Mundkowsky, P. Lange, A. Ivanov, A. W. Black, and D. Suendermann-Oeft, "Multimodal halef: An open-source modular web-based multimodal dialog framework," in *IWSDS*, 2016.

[20] Z. Yu, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell, "Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs," in *Proceedings of SEMDIAL*, 2013, pp. 160–169.

[21] Z. Yu, Z. Xu, A. Black, and A. Rudnicky, "Film promotion chatbot: Interleave social chats in tasks," in *IEEE SLT*, 2016.

[22] ——, "Strategy and policy learning for non-task-oriented conversational systems," in *SIGDIAL*, 2016.