YRRSDS 2024

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems

Proceedings of the Workshop

September 16 - 17, 2024
Kyoto, Japan

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-162-9

Sponsor

The Association for Natural Language Processing

In Collaboration With

Preface

We are thrilled to present the opening remarks for the 20th Young Researchers Roundtable on Spoken Dialogue Systems (YRRSDS) 2024, a workshop dedicated to PhD candidates, PostDocs, and emerging researchers in the field of Spoken Dialogue Systems. YRRSDS 2024 was held in conjunction with the Special Interest Group on Discourse and Dialogue (SIGDIAL) 2024. The workshop took place on September 16-17, 2024, at Kyoto University in Kyoto, Japan. This year's YRRSDS was conducted in an in-person format.

Young researchers submitted a 2-page position paper detailing their current research topics, interests, and the key points they hoped to discuss during the workshop's roundtable sessions. Each submission was carefully reviewed by two senior researchers from our Advisory Committee. We extend our deep gratitude to the Advisory Committee members for their exceptional and insightful reviews. Their contributions have been invaluable in offering critical feedback to the workshop participants at this pivotal stage in their careers.

Participants accepted into the program were required to deliver a brief oral presentation based on their submissions. This year, YRRSDS accepted all 32 submissions received. The roundtable discussions covered topics such as LLMs, multimodality, explainability, evaluation, trustworthiness, ethics, safety, interdisciplinarity, human cognition, and the future of SDSs. Alongside the oral sessions and roundtables, the program featured two outstanding keynote presentations. We would like to express our gratitude and acknowledge our keynote speakers: Koichiro Yoshino (Associate Professor, Tokyo Institute of Technology) and Yoichi Matsuyama (Associate Research Professor, Waseda University and CEO of Equmenopolis, Inc.) for their inspiring talks.

We extend our gratitude to the organizers for making sure the conference ran seamlessly and was enjoyed by all attendees. Finally, we sincerely appreciate the support provided by our sponsor, The Association for Natural Language Processing (ANLP).

Organizing Committee, YRRSDS 2024

Organizing Committee

Organizers:

Koji Inoue, *Kyoto University*
Yahui Fu, *Kyoto University*
Agnes Axelsson, *Delft University of Technology*
Atsumoto Ohashi, *Nagoya University*
Brielen Madureira, *University of Potsdam*
Yuki Zenimoto, *Nagoya University*
Biswesh Mohapatra, *INRIA*
Armand Stricker, *Université Paris-Saclay*
Sopan Khosla, *Amazon Web Services AI Lab*

Advisory Committee:

Timo Baumann
Ryuichiro Higashinaka
Mikio Nakano
Marilyn Walker
Srinivas Bangalore
Ronald Cumbal
Luis Fernando D'Haro
Nina Dethlefs
Mikey Elmers
Tatsuya Kawahara
James Kennedy
Kazunori Komatani
Udo Kruschwitz
Marek Kubis
Divesh Lala
Pierre Lison
Alexandros Papangelis
Giuseppe Riccardi
Pawel Skorzewski
David Traum
Stefan Ultes
Nigel Ward
Hendrik Buschmeier
Kallirroi Georgila
Julia Hirschberg
Michimasa Inaba

Table of Contents

Conference Program

Monday September 16

13:00–13:10 Opening

13:15–14:00 Keynote 1: Koichiro Yoshino

14:20–15:20 Position Talks (Oral) 1

Conversational XAI and Explanation Dialogues
Nils Feldhus

Enhancing Emotion Recognition in Spoken Dialogue Systems through Multimodal Integration and Personalization
Takumasa Kaneko

Towards Personalisation of User Support Systems
Tomoya Higuchi

Social agents for positively influencing human psychological states
Muhammad Yeza Baihaqi

Personalized Topic Transition for Dialogue System
Kai Yoshida

Elucidation of psychotherapy and development of new treatment methods using AI
Shio Maeda

Assessing Interactional Competence with Multimodal Dialog Systems
Mao Saeki

Faithfulness of Natural Language Generation
Patricia Schmidtova

Knowledge-Grounded Dialogue Systems for Generating Interesting and Engaging Responses
Hiroki Onozeki

Towards a Dialogue System that Can Take Interlocutors' Values into Account
Yuki Zenimoto

15:45–16:15 Roundtable 1

Topic 1: Multimodality
Chair: Yuki Zenimoto, Koji Inoue

Topic 2: Data and Techniques
Chair: Yahui Fu, Armand Stricker

Topic 3: Evaluation
Chair: Brielen Madureira, Atsumoto Ohashi

16:45–17:45 Position Talks (Oral) 2

Multimodal Spoken Dialogue System with Biosignals
Shun Katada

Timing Sensitive Turn-Taking in Spoken Dialogue Systems Based on User Satisfaction
Sadahiro Yoshikawa

Towards Robust and Multilingual Task-Oriented Dialogue Systems
Atsumoto Ohashi

Toward Faithful Dialogs: Evaluating and Improving the Faithfulness of Dialog Systems
Sicong Huang

Cognitive model of listener response generation and its application to dialogue systems
Taiga Mori

Topological Deep Learning for Term Extraction
Benjamin Matthias Ruppik

Dialogue Management with Graph-structured Knowledge
Nicholas Thomas Walker

Towards a co-creation dialogue system
Xulin Zhou

Enhancing Decision-Making with AI Assistance
Yoshiki Tanaka

Ontology Construction for Task-oriented Dialogue
Renato Vukovic

18:00–19:00 Casual Dinner

Tuesday September 17

09:30–10:15 Keynote 2: Yoichi Matsuyama

10:30–11:45 Position Talks (Oral) 3

Generalized Visual-Language Grounding with Complex Language Context
Bhathiya Hemanthage

Towards a Real-Time Multimodal Emotion Estimation Model for Dialogue Systems
Jingjing Jiang

Exploring Explainability and Interpretability in Generative AI
Shiyuan Huang

Innovative Approaches to Enhancing Safety and Ethical AI Interactions in Digital Environments
Zachary Yang

Leveraging Linguistic Structural Information for Improving the Model's Semantic Understanding Ability
Sangmyeong Lee

Multi-User Dialogue Systems and Controllable Language Generation
Nicolas Wagner

Enhancing Role-Playing Capabilities in Persona Dialogue Systems through Corpus Construction and Evaluation Methods
Ryuichi Uehara

Character Expression and User Adaptation for Spoken Dialogue Systems
Kenta Yamamoto

Interactive Explanations Through Dialogue Systems
Isabel Feustel

Towards Emotion-aware Task-oriented Dialogue Systems in the Era of Large Language Models
Shutong Feng

Utilizing Large Language Models for Customized Dialogue Data Augmentation and Psychological Counseling
Zhiyang Qi

Toward More Human-like SDSs: Advancing Emotional and Social Engagement in Embodied Conversational Agents
Zi Haur Pang

13:30–14:00 Roundtable 2

Topic 4: Explainability and Trustworthy
Chair: Atsumoto Ohashi, Yuki Zenimoto

Topic 5: Taking Inspiration from Human Cognition
Chair: Brielen Madureira, Armand Stricker

Topic 6: Interdisciplinarity
Chair: Koji Inoue, Yahui Fu

14:30–15:00 Roundtable 3

Topic 7: Present and Future of SDSs
Chair: Atsumoto Ohashi, Yahui Fu

Topic 8: Possibilities and limits of LLMs
Chair: Armand Stricker, Yuki Zenimoto

Topic 9: Ethics and Safety
Chair: Koji Inoue, Brielen Madureira

15:30–15:45 Wrap up

15:45–16:00 Photo Session

16:00–18:00 Social Activity

Keynotes

Keynote 1: Multimodal Dialogue System Research and Careers, the Past 10 Years, the Future 10 Years

Koichiro Yoshino (Associate Professor, Tokyo Institute of Technology, Japan)

Abstract:

Over the past decade, advances in deep learning have made it easier for dialogue systems to handle various modalities in the real world, and research on multimodal dialogue systems has advanced enormously. Voice input devices such as Amazon Alexa and Google Home have entered our daily lives in the past ten years. Agent robots that handle various modalities will be realized as real-world services in another 10 years. What should young researchers consider in such an environment as they develop their careers? I participated as a PhD student at YRRSDS2014 10 years ago. Based on my experiences in academia over the past 10 years, I would like to discuss what career path you should follow in the next 10 years.

Biography: Koichiro Yoshino is an associate professor at Tokyo Institute of Technology, and is cross-appointed with RIKEN as a team leader. He received his B.A. from Keio University in 2009, and M.E. and Ph.D. in informatics from Kyoto University in 2014. He worked at Kyoto University and NAIST. From 2019 to 2020, he was a visiting research of Heinrich-Heine-Universität Düsseldorf, Germany. He is working on areas of spoken and natural language processing, especially robot dialogue systems. Dr. Koichiro Yoshino received several honors, including the best paper award of IWSDS2020, IWSDS2024, and the best paper award of the 1st NLP4ConvAI workshop. He is a member of IEEE Speech and Language Processing Technical Committee (SLTC), a member of Dialogue System Technology Challenge (DSTC) Steering Committee, an action editor of ACL rolling review (ARR), a board member of SIGdial and a board member of association for The Association for Natural Language Processing.

Keynote 2: From Dialogue System Research to Social Innovation

Yoichi Matsuyama (Co-Founder and CEO, Equmenopolis, Inc.)

Abstract: In the rapidly evolving field of dialogue systems, researchers hold a unique position, equipped to drive advancements in dialogue processing technologies while addressing emerging social needs. This talk will trace the journey from academic research to founding a university spin-out startup, showing how insights from cutting-edge technologies and user experiences can tackle real-world challenges and generate a social impact. Drawing from my experience as a dialogue systems researcher turned entrepreneur, I'll discuss the role of dialogue technologies in shaping the future of human-computer interaction and their broader implications for social innovation, encouraging young researchers to think creatively beyond academic boundaries.

Biography: Yoichi Matsuyama is the Co-Founder and CEO of Equmenopolis, Inc. and an Associate Research Professor at Waseda University in Tokyo. The mission of Equmenopolis is "Towards a Human-AI Co-Evolving Society," where we dispatch conversational AI agents to schools and workplaces to improve creativity and productivity. He specializes in developing computational models of human conversation, integrating AI, linguistics, social science, and human-agent interaction. Before his current role, he was a Postdoctoral Fellow at the ArticuLab, School of Computer Science, Carnegie Mellon University. His work has garnered attention from major media outlets, including MIT Technology Review, The Washington Post, CNBC, BBC, CNET, Popular Science, Nikkei, and NHK. He holds a B.A. in cognitive psychology and media studies, as well as an M.E. and Ph.D. in computer science from Waseda University, earned in 2005, 2008, and 2015, respectively.

Organizers' Notes of the Roundtable Discussions

Roundtable 1: Multimodality (Chair: Yuki Zenimoto, Koji Inoue)

Goal: Discuss the diverse aspects of multimodality in SDSs, including the utilization of visual information, environmental context, gestures, emotions, and personalization. Explore how these aspects can be effectively combined and what effect can be achieved.

Summary: In this discussion, we shared our individual works, focusing on multimodal dialogue systems as well as the challenges of sensing and annotation. We began by exploring the concept of ideal multimodal communication, emphasizing the need for dialogue to be smooth and duplex. In this context, we highlighted the significance of multimodal processing. We then addressed the challenges of annotating subjective phenomena like emotion labels. Additionally, we examined various measurable multimodal behaviors such as gestures, respiration, eye-gaze, and heart rate. Lastly, we delved into the interface of multimodal dialogue systems, comparing robots with CG agents/avatars and referencing the uncanny valley theory.

Roundtable 2: Data and Techniques (Chair: Yahui Fu, Armand Stricker)

Goal: Discuss the challenges and innovative approaches related to data creation, collection, and learning techniques for advanced SDSs, such as adaptation for unseen data, controllability, effective use of LLMs, and efficiency.

Summary: In this discussion, we first examined the challenges of prompting robustness and parameter-efficient fine-tuning techniques like prefix-tuning and LoRA, highlighting that language models are sensitive to prompt variations and benefit from training on diverse prompts to adapt to unseen data. Then, we compared synthetic data generation with human data labeling: synthetic data offers scalability but risks model degeneration and lacks diversity, while human-labeled data is richer but costly and prone to subjective interpretations and low inter-annotator agreement, especially in emotion recognition tasks. Lastly, we discussed large language models' effectiveness in processing speech data, noting they handle ASR noise well in tasks like emotion recognition but face difficulties in slot filling and with non-English accents. These insights emphasize the need for innovative data creation and learning techniques to improve adaptability, controllability, and efficiency in advanced SDSs.

Roundtable 3: Evaluation (Chair: Brielen Madureira, Atsumoto Ohashi)

Goal: Critically examine the current evaluation practices for SDSs and their limitations. Explore innovative automated evaluation metrics and methodologies for various domains, such as in non-task-oriented dialogues.

Summary: In this discussion, we focused on the challenges of evaluating dialogue systems, particularly LLMs, and emphasized the limitations of existing evaluation metrics such as BLEU. These metrics are especially problematic for open-domain dialogue systems, where human-like qualities are difficult to measure objectively. We debated the need for a more holistic approach that considers aspects like coherence and common sense. Benchmarking was another key topic, with concerns raised about models becoming over-specialized and "gaming" the system to perform well on specific tests rather than improving general performance. The balance between human and automatic evaluations was discussed. Participants concluded by stressing the importance of real-world testing and aligning evaluations with user needs, rather than purely focusing on making systems human-like.

Roundtable 4: Explainability and Trustworthy (Chair: Atsumoto Ohashi, Yuki Zenimoto)

Goal: Discuss the importance and methods of making SDSs more explainable and trustworthy, including the development of conversational explainable AI (XAI), the evaluation of reliability, the controllability of language generation, and dealing with closed proprietary models.

Summary: In this discussion, we discussed various aspects of explainability in dialogue systems, focusing on the challenges of making ML models interpretable and understandable for both scientists and general users. We identified a gap in tools that enable interactive explainability for general users and discussed the ethical implications of trusting explanations generated by models. We also touched on methods for evaluating the quality of explanations, with simulatability being one approach where users predict the model's output based on its explanation. Additionally, we raised the challenges of working with large black-box models (e.g., those accessed via APIs), where researchers lack insight into their inner workings.

Roundtable 5: Taking Inspiration from Human Cognition (Chair: Brielen Madureira, Armand Stricker)

Goal: Explore how insights from human cognitive processes, language acquisition, and social interaction can inform the development of more advanced SDSs, including the integration of physiological signals and the development of collaborative and creative systems.

Summary: The session centered on examining how insights from human cognitive processes can inform the development of more advanced spoken dialogue systems (SDSs). By integrating elements such as theory of mind, physiological signals, multimodal perception, and visual cues, dialogue systems can become more adaptive, capable of keeping information up to date, and better at integrating new facts in real-time. A major focus was on how these systems can emulate human-like understanding and illocutionary intent, as well as the implications of such advancements for user expectations, where more human-like behaviors tend to amplify the impact of errors when they occur. Additionally, the discussion explored how long-term interactions with SDSs could be improved by employing strategies like smoother turn-taking, meta-learning, and curriculum learning, ensuring that systems adapt to communication pace over time.

Roundtable 6: Interdisciplinarity (Chair: Koji Inoue, Yahui Fu)

Goal: Discuss how we can foster collaboration between the fields of SDSs and other disciplines, such as linguistics, psychology, robotics, and social sciences. Explore the benefits and ways to incorporate insights from other fields into practical SGSs development.

Summary: This discussion emphasized the importance of interdisciplinary collaboration, particularly integrating insights from psychology, linguistics, robotics, and social sciences. Participants highlighted the need for collaboration to create more human-like and efficient systems. Understanding users' emotions and adapting to text-based and spoken communication styles was noted as crucial, with experiments requiring careful design and sufficient participant numbers, potentially more than 100 in diverse real-world settings. While large language models (LLMs) provide a cost-effective way to test systems, they cannot replace the need for real human input. Psychological insights can improve LLM performance, but human evaluations are essential for quality. In robotics, transferring knowledge between systems like CommU and ERICA presents challenges. Ultimately, interdisciplinary collaboration and real human interaction are key to advancing SDSs.

Roundtable 7: Present and Future of SDSs (Chair: Atsumoto Ohashi, Yahui Fu)

Goal: Critically think about the current directions in the SDSs field and the reasons and necessity for doing so. Discuss future research directions, including to what extent human-like SDSs are desirable and the ideal relationship between humans and SDSs.

Summary: In this discussion, we shared our individual motivations for SDSs, such as the need for AI systems that fulfill communication needs and offer companionship. One key point of debate was the necessity of a physical body in AI companions, where some argued that emotional bonds could be formed through voice alone, while others maintained that physical interaction was essential in certain contexts, such as companionship or therapeutic relationships. There was also a discussion about whether modular or integrated approaches like LLMs would be more effective for future AI systems. Some highlighted the advantages of modular systems (e.g., better control and faithfulness), while others pointed out that multimodal models could simplify interaction. Finally, the discussion concluded with thoughts on the future of AI, with some expressing optimism about integrating advanced models with robots for even more sophisticated interactions.

Roundtable 8: Possibilities and limits of LLMs (Chair: Armand Stricker, Yuki Zenimoto)

Goal: Discuss the capabilities and limitations of LLMs in the context of SDSs. Explore how to effectively incorporate LLMs into SDSs, such as architecture design, controllability, and handling of multimodal dialogues.

Summary: In this discussion, we discussed the capabilities and limitations of Large Language Models (LLMs) in spoken dialogue systems (SDSs). We began by questioning the necessity of incorporating LLMs into SDSs and shared both successful and challenging experiences. A key topic was the constraints of autoregressive models, which generate responses token-by-token, potentially limiting their planning capabilities and output fidelity. We debated whether vision-language models or LLMs trained on multimodal data suffice for capturing complex meanings or if symbolic representations are necessary. We also explored the need to look beyond model responses to understand internal behaviors. The discussion then turned to the architecture of multimodal dialogue systems, weighing end-to-end against modular designs, and considering the limitations of prompt optimization in ensuring controllable and adaptable systems. Participants suggested combining prompt and internal modifications, such as chain-of-thought decoding and control modules, to refine the generation process. Ultimately, we emphasized the importance of balancing general model capabilities with task-specific requirements for optimal performance in SDS applications.

Roundtable 9: Ethics and Safety (Chair: Koji Inoue, Brielen Madureira)

Goal: Raise awareness about the ethical considerations and potential risks associated with the development and deployment of SDSs, such as when working with powerful but opaque models and creating human-like SDSs. Address concerns related to privacy, data rights, toxicity, and misinformation.

Summary: This discussion emphasized the importance of addressing ethical concerns in the development and use of spoken dialogue systems. It calls for ethical oversight during human trials and data collection, especially in sensitive areas like emotions or mental health. Ethical responsibility should not be entirely shifted to external bodies, with individual accountability being crucial. The lack of transparency in newer technologies like large language models (LLMs) poses challenges in explaining errors. There is a need to manage harmful behaviors, such as toxicity and inappropriate personalization, with reversible actions and consideration of cultural diversity. It also warned about the risks of over-reliance on commercial LLMs and advocated for transparency and open models for development and debugging. Governments may need to regulate the field, and developers should work to raise awareness of these limitations and responsibilities.

Nils Feldhus

German Research Center for Artificial Intelligence (DFKI)
Alt-Moabit 91c
10559 Berlin
Berlin, Germany

nils.feldhus@dfki.de
https://nfelnlp.github.io/

1 Research interests

My main research interest is **human-centric explainability**, i.e., making language models more interpretable by building applications that lower the barrier of entry to explanations. I am enthusiastic about **interactive systems** that pique the interest of more people beyond just the experts to learn about the **inner workings of language models**. My hypothesis is that users of language model applications and dialogue systems are more satisfied and trusting if they can look behind the curtain and get easy access to explanations of their behavior.

1.1 Dialogue-based explainability

Human-centered XAI is concerned with incorporating insights from Human-Computer Interaction (HCI) into the field of XAI (Miller, 2019; Ehsan and Riedl, 2020; Weld and Bansal, 2019). Many XAI systems have interactive components, elaborate user interfaces and are evaluated with user studies (Chromik and Butz, 2021; Bertrand et al., 2023). Only recently, however, there has been a push towards conceptualizing dialogue-based XAI systems. Lakkaraju et al. (2022) proposed four modules which are necessary for explanatory conversational systems: Natural language understanding (NLU), explanation algorithm, response generation, and a graphical user interface. Representative systems like TalkToModel (Slack et al., 2023), ConvXAI (Shen et al., 2023), InterroLang (Feldhus et al., 2023), and LLMCheckup (Wang et al., 2024) all implement these four modules.

However, the current conversational XAI systems exhibit a lack of understanding the user and responding to them. This is because they do not consider context and often resemble question answering setups (request and provide explanations). They lack a dedicated dialogue management, as traits of information-seeking (Stepin et al., 2024), mixed-initiative (or proactive) dialogues (Deng et al., 2023), argumentation dialogues (Bex and Walton, 2016) and teacher-student (or tutorial) dialogues (Wachsmuth and Alshomary, 2022; Lee et al., 2023; Liu et al., 2024b) are necessary for a natural explanatory dialogue.

Current research in computational argumentation (Bex and Walton, 2016; Madumal et al., 2019) provides valuable insights into explanatory dialogue interactions, yet it remains relatively abstract and does not cover the full range of explanation moves. Similarly, while didactics literature (Wachsmuth and Alshomary, 2022; Hennessy et al., 2016) defines many moves, it lacks a comprehensive dialogue strategy.

I am currently working on a concept for an explanatory dialogue management which is able to take context into account and easily adapt to user needs. I conduct user studies to examine if LLM-generated explanations are able to take dialogue context into account and, at the same time, beat conventional template-based answers in terms of likeability and perceived faithfulness.

LLMs are getting increasingly better at synthesizing natural language explanations (Wiegreffe et al., 2022) and offer the possibility to hold conversations in various styles, e.g. concise vs. elaborate explanations (Liu et al., 2024a). On top of that, they have been shown to perform dialogue state tracking exceptionally well (Heck et al., 2023). However, LLMs also introduce issues with ground truth, which recent work has started to analyze with test suites (Atanasova et al., 2023) and user studies (Si et al., 2024). I intend to answer the question of whether the faithfulness as perceived by the user matches the actual faithfulness as measured by explanation evaluation and LLM factuality evaluation methods.

1.2 Explanations in tutoring systems

Explanations can also be framed as instructions, e.g. in didactics, where a teacher instructs a student on a concept or topic (Wachsmuth and Alshomary, 2022). Didactics research often debates which teaching strategies lead to the best learning outcome (Roelle et al., 2015). I am investigating if language models can reliably detect if a teacher follows good practices as defined by teaching strategies Feldhus et al. (2024). It turns out that this requires a very thorough definition of acts and high expertise of annotators to achieve a sufficient agreement and trustworthy evaluation results.

A language model that can extract explanation and

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 1–4
September 16–17, 2024.

teaching moves would be helpful for didacticians to self-check and scale up assessments. This is why I am also looking into evaluation measures for generated text, specifically those for measuring how close teachers stick to lesson planning (Feldhus et al., 2024) and accessibility such as readability (Hsu et al., 2024).

Several works have pointed out the difficult of using neural language models for the purpose of tutoring (Macina et al., 2023; Wang and Demszky, 2023). A final goal would be personalized tutoring chatbots that are aware of the user's personality and can adapt their explanatory processes to the expertise and mental model of the user (Fernau et al., 2022).

2 Spoken dialogue system (SDS) research

I believe that SDS research play a vital role in many domains, such as medicine (clinical decision support) and journalism (fact checking). Assistants have a growing presence in our everyday lives and they need to be trustworthy and accountable. Faithful explanations that are grounded in the data, architecture and documentation of the models need to accompany dialogue systems for that reason.

In the coming years, SDS research needs a higher focus on user studies and human evaluation rather than architectures, scaling and exuberant claims of emergent capabilities or agency. With a focus on evaluation and the collection of valuable resources for the growing range of downstream tasks and with the purpose of filling pressing gaps in a multilingual landscape, we can mitigate the actual and present risks for society from uncontrolled systems that already extrude falsehoods and augment harmful biases.

3 Suggested topics for discussion

- How should we design effective explanatory dialogue and conversational XAI systems?
  - Under which circumstances can they depend on LLMs?
  - What findings from other disciplines such as didactics and argumentation should we take into account when building such systems?
- How can the quality of explanation dialogues be evaluated?
- Are LLMs reliable and trustworthy tutoring systems?

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Toronto, Canada, pages 283–294. https://doi.org/10.18653/v1/2023.acl-short.25.

Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2023. On selective, mutable and dialogic xai: A review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, CHI '23. https://doi.org/10.1145/3544548.3581314.

Floris Bex and Douglas Walton. 2016. Combining explanation and argumentation in dialogue. *Argument & Computation* 7(1):55–68. https://doi.org/10.3233/AAC-160001.

Michael Chromik and Andreas Butz. 2021. Human-XAI interaction: a review and design principles for explanation user interfaces. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18*. Springer, pages 619–640. https://doi.org/10.1007/978-3-030-85616-8_36.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 10602–10621. https://doi.org/10.18653/v1/2023.findings-emnlp.711.

Upol Ehsan and Mark O. Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*. Springer International Publishing, Cham, pages 449–466. https://doi.org/10.1007/978-3-030-60117-1_33.

Nils Feldhus, Aliki Anagnostopoulou, Qianli Wang, Milad Alshomary, Henning Wachsmuth, Daniel Sonntag, and Sebastian Möller. 2024. Towards modeling and evaluating instructional explanations in teacher-student dialogues. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*. Association for Computing Machinery, New York, NY, USA, GoodIT '24, page 225–230. https://doi.org/10.1145/3677525.3678665.

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023. InterroLang: Exploring NLP models and datasets

through dialogue-based explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 5399–5421. https://doi.org/10.18653/v1/2023.findings-emnlp.359.

Daniel Fernau, Stefan Hillmann, Nils Feldhus, Tim Polzehl, and Sebastian Möller. 2022. Towards personality-aware chatbots. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Edinburgh, UK, pages 135–145. https://doi.org/10.18653/v1/2022.sigdial-1.15.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Toronto, Canada, pages 936–950. https://doi.org/10.18653/v1/2023.acl-short.81.

Sara Hennessy, Sylvia Rojas-Drummond, Rupert Higham, Ana María Márquez, Fiona Maine, Rosa María Ríos, Rocío García-Carrión, Omar Torreblanca, and María José Barrera. 2016. Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, Culture and Social Interaction* 9:16–44. https://doi.org/https://doi.org/10.1016/j.lcsi.2015.12.001.

Yi-Sheng Hsu, Nils Feldhus, and Sherzod Hakimov. 2024. Free-text rationale generation under readability level control. *arXiv* abs/2407.01384. https://arxiv.org/abs/2407.01384.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. *HCAI @ NeurIPS 2022* https://arxiv.org/abs/2202.01875.

Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive step-by-step explanatory dialogues to answer children's why and how questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, CHI '23. https://doi.org/10.1145/3544548.3581369.

Yinhong Liu, Yimai Fang, David Vandyke, and Nigel Collier. 2024a. TOAD: Task-oriented automatic dialogs with diverse response styles. *To appear in ACL 2024 Findings* https://arxiv.org/abs/2402.10137.

Zhengyuan Liu, Stella Xin Yin, Carolyn Lee, and Nancy F. Chen. 2024b. Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions. *arXiv* abs/2404.03429. https://arxiv.org/abs/2404.03429.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Opportunities and challenges in neural dialog tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, pages 2357–2372. https://aclanthology.org/2023.eacl-main.173.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '19, page 1033–1041. https://dl.acm.org/doi/abs/10.5555/3306127.3331801.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38. https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007.

Julian Roelle, Claudia Müller, Detlev Roelle, and Kirsten Berthold. 2015. Learning from instructional explanations: Effects of prompts based on the active-constructive-interactive framework. *PLOS ONE* 10(4):e0124115. https://doi.org/10.1371/journal.pone.0124115.

Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023. ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-AI scientific writing. In *Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, New York, NY, USA, CSCW '23 Companion, page 384–387. https://doi.org/10.1145/3584931.3607492.

Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. 2024. Large language models help humans verify truthfulness – except when they are convincingly wrong. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, Mexico City, Mexico, pages 1459–1474. https://aclanthology.org/2024.naacl-long.81.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learn-

ing models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence* https://doi.org/10.1038/s42256-023-00692-8.

Ilia Stepin, Katarzyna Budzynska, Alejandro Catalá, Martin Pereira-Fariña, and Jose Maria Alonso-Moral. 2024. Information-seeking dialogue for explainable artificial intelligence: Modelling and analytics. *Argument & Computation* https://content.iospress.com/articles/argument-and-computation/aac220011.

Henning Wachsmuth and Milad Alshomary. 2022. "mama always had a way of explaining things so I could understand": A dialogue corpus for learning to construct explanations. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pages 344–354. https://aclanthology.org/2022.coling-1.27.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Josef Genabith, Leonhard Hennig, and Sebastian Möller. 2024. LLMCheckup: Conversational examination of large language models via interpretability tools and self-explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Mexico City, Mexico, pages 89–104. https://aclanthology.org/2024.hcinlp-1.9.

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, pages 626–667. https://doi.org/10.18653/v1/2023.bea-1.53.

Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62(6):70–79. https://doi.org/10.1145/3282486.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, pages 632–658. https://doi.org/10.18653/v1/2022.naacl-main.47.

Biographical sketch

Nils Feldhus is a final-year PhD student at DFKI under the supervision of Prof. Dr.-Ing. Sebastian Möller at TU Berlin. His research focus is making language model explanations accessible to more target groups such as domain experts and NLP beginners. He presented his work at various conferences, including EMNLP (2021 & 2023), SIGDIAL (2022), ACL (2023), and IJCAI (2022). He holds degrees in computational linguistics (BA) from Heidelberg University and cognitive systems (MSc) from Potsdam University. He is an area chair for ACL Rolling Review since February 2024 for the Interpretability and Analysis of NLP Models track. In his free time, he enjoys music production, cycling, board games, and nature photography.

Takumasa Kaneko

University of Electro-Communications
Chofu, Tokyo
Japan

k2440004@gl.cc.uec.ac.jp

1 Research interests

Emotion recognition is vital for improving the quality of human–human and human–machine inter- actions. Especially in spoken dialogue systems (SDSs), responses can be generated by predicting users' emotions and considering their intentions. The response content can change depending on the user's emotion for similar utterances, allowing for natural communication. However, emotion recognition remains a challenging task, and responses based on incorrect emotion predictions can significantly impair the user experience.

One of my research interests is multimodal emotion recognition. Studies have proposed several emotion recognition models using multiple modalities to improve recognition accuracy (Sun et al., 2023; Wu et al., 2023; Yang et al., 2023). Multimodal emotion recognition models perform better than unimodal ones.

Another area that caught my attention is **personalization** in the emotion recognition task. Specifically, methods for personalization without fine-tuning have been recently proposed (Tran et al., 2023). Personalization without fine-tuning can greatly improve the utility and user experience of dialogue systems.

1.1 Speech Emotion Recognition

Speech emotion recognition models have improved year by year through various efforts. Zou et al. (2022) proposed an emotion recognition model that uses three types of acoustic information as input: raw waveform data, Mel-Frequency Cepstrum Coefficient (MFCC), and spectrogram. This model extracts features from the three acoustic data types and fuses the extracted features with a coattention mechanism. Kim et al. (2022) developed a model that combines the focus attention mechanism and the calibration attention mechanism. This proposed attention mechanism allows us to focus more on the important regions in the feature space of speech data. Pan et al. (2024) proposed a model that uses contrastive learning and gender information. Using information other than speech for prediction, such as gender information, is important for improving accuracy. Hence, many proposed models use text and video in addition to speech.

Although several multimodal models have been proposed that use video, text, and speech as input (Sun et al., 2023; Wu et al., 2023; Yang et al., 2023), their use is currently limited to video analysis and other applications that do not consider real-time performance because of the high computational complexity of handling video. In an SDS, it is difficult to use all the user's video images during speech because the speed of emotion recognition is important.

Furthermore, a multimodal dataset is more expensive to create than a unimodal dataset. Therefore, multimodal emotion recognition datasets are not available, but many cases have dealt with emotion recognition datasets with facial expressions and speech. Against this background, I aim to construct a multimodal emotion recognition model with facial expressions and speech using emotion recognition datasets for each modality. I construct a multimodal emotion recognition dataset by pairing similar labels from each dataset and trained a multimodal model. I compare the difference in performance between the multimodal model trained on the constructed dataset and the model trained on the unimodal dataset to confirm the effectiveness of the method.

1.2 Personalization

Typical speech emotion recognition tasks aim to predict emotion labels such as happiness, sadness, anger, and neutral. However, these labels often fall short of capturing the complexity of human emotions. An alternative approach is to use emotion attributes as suggested by core affect theory (Russell, 2003). Emotion attributes are represented as continuous scores in dimensions such as arousal (calm versus active), valence (unpleasant versus pleasant), and dominance (weak versus strong) for more nuanced expressions of emotions.

The prediction of valence is known to heavily depend on the speech characteristics of individual speakers, making it more challenging than predicting the other two attributes (Sridhar et al., 2018). However, adopting personalization techniques can address the variability in expression among different speakers. This approach allows for accurate predictions by considering each speaker's unique features.

Sridhar and Busso (2022); Tran et al. (2023) showed that prediction accuracy can be improved by embedding

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 5–7

speaker characteristics from training data and identifying speakers with similar characteristics in test data. This method necessitates that the training data include a diverse array of speakers. If the training data does not contain a sufficient number of speakers, adding new trainable speaker data may be necessary. In such cases, existing methods require retraining not only the speaker embedding module but also the speech encoder weights after each data addition.

To address this, I introduce the concept of continuous prompt tuning, in which speaker prompts are added to the inputs of each speech encoder layer. In this approach, the weights of the speech encoder are frozen, and only the speaker prompts are updated to learn the speaker's characteristics. This allows for the addition of new speaker data without retraining the weights of the speech encoder.

2 Spoken dialogue system (SDS) research

Recently, the development of large language models has made it possible to perform tasks that had been constrained by technical limitations and costs and has allowed us to achieve high performance in demanding tasks. For example, in natural language processing, tasks such as document summarization, translation, and question–answering systems, which were considered challenging, can now be executed with high accuracy. This advancement has allowed for various practical applications such as information retrieval and customer support.

However, many unresolved challenges remain. One significant difficulty is the integration and consistent processing of multiple modality information (e.g., text, images, audio). Effective methods to model the interactions between these modalities are not yet fully established.

Additionally, there is a demand for the rapid processing of such diverse information. Current models consume a vast computational resources, making real-time response challenging. In speech dialogue systems, understanding the user's intent quickly and accurately is paramount; any latency can degrade the user experience. Hence, improving computational efficiency remains a critical research area.

Moreover, current models have limitations in accurately understanding human intent and emotions. While many language models are trained on large datasets and are adept at understanding general patterns and contexts, they continue to struggle with grasping subtle nuances and emotional changes in specific situations. For instance, they may misinterpret sarcasm or the use of polysemous words.

Future research will resolve these issues and further advance SDSs. If technology can be established to manage multiple modalities of information in an integrated manner and process it at high speed, a system that can understand the user's intentions more accurately will be realized. In addition, if emotions and intentions can be accurately captured, more natural and humanlike dialogue will be possible. As a result, SDSs are expected to find applications in a wide range of fields, including medicine, education, and entertainment.

3 Suggested topics for discussion

I suggest discussing the following:

- How can we improve the accuracy of emotion recognition in SDSs?
- What are the challenges in multimodal emotion recognition?
- Can personalization be applied to tasks other than speech emotion recognition tasks?
- How can reinforcement learning be used in dialogue systems?

References

Junghun Kim, Yoojin An, and Jihie Kim. 2022. Improving Speech Emotion Recognition Through Focus and Calibration Attention Mechanisms. In *Proc. Interspeech 2022*. pages 136–140. https://doi.org/10.21437/Interspeech.2022-299.

Yu Pan, Yanni Hu, Yuguang Yang, Wen Fei, Jixun Yao, Heng Lu, Lei Ma, and Jianjun Zhao. 2024. Gemo-clap: Gender-attribute-enhanced contrastive language-audio pretraining for accurate speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 10021–10025.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110(1):145.

Kusha Sridhar and Carlos Busso. 2022. Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech. *IEEE Transactions on Affective Computing* 13(4):1959–1972.

Kusha Sridhar, Srinivas Parthasarathy, and Carlos Busso. 2018. Role of regularization in the prediction of valence from speech. *Interspeech 2018* .

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. 2023. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 658–670.

Minh Tran, Yufeng Yin, and Mohammad Soleymani. 2023. Personalized adaptation with pre-trained speech encoders for continuous emotion recognition. *arXiv preprint arXiv:2309.02418* .

Shaoxiang Wu, Damai Dai, Ziwei Qin, Tianyu Liu, Binghuai Lin, Yunbo Cao, and Zhifang Sui. 2023. Denoising bottleneck with mutual information maximization for video multimodal fusion. *arXiv preprint arXiv:2305.14652* .

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 7617–7630.

Heqing Zou, Yuke Si, Chen Chen, Deepu Rajan, and Eng Siong Chng. 2022. Speech emotion recognition with co-attention based multi-level acoustic information. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 7367–7371.

Biographical sketch

Takumasa Kaneko is a PhD student in the Department of Informatics at the University of Electro- Communications. His master's research focused on developing a speech dialogue system that considers emotions.

Tomoya Higuchi

The University of Electro-Communications
Tokyo, Japan

h2430109@gl.cc.uec.ac.jp

1 Research interests

My research interests lie on the development of advanced **user support systems**, emphasizing the enhancement of user engagement and system effectiveness. The field of user support systems aims to help users accomplish complex tasks efficiently while ensuring a pleasant and intuitive interaction experience. I explore how to incorporate engaging and context-appropriate assistance into these systems to make the task completion process more effective and enjoyable for users.

A key area of my research is user support system personalization, which includes methods for adapting system behavior, interface elements, and assistance strategies based on user profiles, skill levels, and interaction histories. I am specifically interested in approaches that can achieve personalization without extensive manual configuration, allowing the support system to dynamically adjust to each user's evolving needs and preferences. To achieve this in a news commentary dialog system, I propose multiple question candidates with varying levels of difficulty to the user and, based on the selected questions, estimate and adapt the user's level of understanding of the news article.

1.1 Building a conversational question answering system

Conversational question generation involves producing multiturn questions related to a document, aiming to fulfill the user's information needs through conversation. Methods include generating questions based on dialog history, consisting of question–response pairs, and supporting sentences. Pan et al. (2019) developed a consistent question generation process using reinforcement learning. Do et al. (2023) proposed a two-stage framework for conversational question generation, determining what to ask and how to ask based on a semantic graph. These methods use datasets for conversational question answering, such as DoQA (Campos et al., 2020), QuAC (Choi et al., 2018), and CANARD (Elgohary et al., 2019), generating simple one-word-answer questions.

Qin et al. (2023) and Chernyavskiy et al. (2023) used large language models (LLMs) to generate fluent responses based on preselected knowledge. Following these methods, this study uses LLMs to generate questions that elicit explanatory answers in a free-form manner.

1.2 News chatbot

The media uses dialog content related to news articles for their clarity, but they are manually created by journalists. Manual creation is inefficient because it requires significant cost and time. To address this, Laban et al. (2020) proposed a method to automatically construct chatbots from news articles. This method presents question candidates to the user, but individual user characteristics are not considered in the creation of these candidates. Therefore, this study aims to provide desired question candidates by generating them based on the user's understanding.

1.3 Question generation considering user characteristics

If the user's social group characteristics differ, the questions are also expected to vary. Stewart and Mihalcea (2022) developed a method to generate questions reflecting user characteristics. This method trains a text generation model using social media data, considering social groups such as domain expertise. Additionally, An et al. (2021) designed a prototype conversation agent that generates speech based on what the user knows and does not know, verifying the effectiveness of incorporating the user's knowledge. Inspired by these methods, this research uses LLM to generate questions that consider the extent to which the user understands news articles. Specifically, the user is presented with questions with three levels of difficulty, and their understanding of the news article is assessed based on the difficulty level of the question they select.

1.4 Question generation with adjusted difficulty level

In educational contexts, the generation of questions with controlled difficulty is gaining momentum. Controlling the difficulty of questions in a question–answer learning system allows for learning to be tailored to individual users. Cheng et al. (2021) defined the difficulty of questions based on the number of inference steps required to answer them and proposed a method that gradually increases the difficulty through step-by-step rewrit-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 8–10
September 16–17, 2024.

ing. However, in this study, the difficulty of questions stems from factors such as the background knowledge required for comprehending news articles, rendering this definition unsuitable. Therefore, we generate questions with adjusted difficulty using an LLM through few-shot learning (Brown et al., 2020), following examples manually created in advance.

2 Spoken dialogue system (SDS) research

The advancement of LLMs has made it possible to build user support dialogue systems for a wide range of users. However, challenges remain in adapting these systems to individual users. Personalization in voice dialogue systems, which uses emotions and intentions derived from voice features, is particularly promising. By analyzing tone, speed, and accent, user profiling becomes more precise, potentially offering personalized support.

Using voice in user support dialogue systems also improves accessibility. Voice interfaces allow the system to be accessed by visually impaired users and those in situations where manual input is difficult, such as while driving. In my research on news article explanation interfaces, I reduced the user's burden by automatically generating question candidates. Voice can further alleviate the user's burden by eliminating the need to input long texts manually. Moreover, designing appropriate voice dialogues and endowing the system with a personality to strengthen emotional connections with users could enhance engagement.

However, several challenges need to be addressed to realize these benefits, including improving speech recognition accuracy, effectively utilizing audio features, and reducing hallucinations during task execution.

3 Suggested topics for discussion

I suggest discussing the following topics:

- What are the benefits and challenges of converting existing text-based user support dialog systems to voice-based systems?
- Can a multimodal LLM become an assistant spoken dialog system (SDS) that exceeds existing text-based LLM?
- Can the use of voice features in an SDS help personalize the system?

References

Sungeun An, Robert Moore, Eric Young Liu, and Guang-Jie Ren. 2021. Recipient design for conversational agents: Tailoring agent's utterance to user's knowledge. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. pages 1–5.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - accessing domain-specific FAQs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pages 7302–7314.

Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. pages 5968–5978.

Alexander Chernyavskiy, Max Bregeda, and Maria Nikiforova. 2023. PaperPersiChat: Scientific paper discussion chatbot using transformers and discourse flow management. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. pages 584–587. https://doi.org/10.18653/v1/2023.sigdial-1.54.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. pages 2174–2184.

Xuan Long Do, Bowei Zou, Shafiq Joty, Tran Tai, Liangming Pan, Nancy Chen, and Ai Ti Aw. 2023. Modeling what-to-ask and how-to-ask for answer-unaware conversational question generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 10785–10803.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. pages 5918–5924.

Philippe Laban, John Canny, and Marti A. Hearst. 2020. What's the latest? a question-driven news chatbot. In *Proceedings of the 58th Annual Meeting of the Association for Computational Lin-*

guistics: System Demonstrations. pages 380–387. https://doi.org/10.18653/v1/2020.acl-demos.43.

Boyuan Pan, Hao Li, Ziyu Yao, Deng Cai, and Huan Sun. 2019. Reinforced dynamic reasoning for conversational question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pages 2114–2124.

Lang Qin, Yao Zhang, Hongru Liang, Jun Wang, and Zhenglu Yang. 2023. Well begun is half done: Generator-agnostic knowledge pre-selection for knowledge-grounded dialogue. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. pages 4696–4709.

Ian Stewart and Rada Mihalcea. 2022. How well do you know your audience? toward socially-aware question generation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. pages 255–269. https://doi.org/10.18653/v1/2022.sigdial-1.27.

Biographical sketch

Tomoya Higuchi is a master's student at the Graduate School of Informatics and Engineering, University of Electro-Communications. He is interested in user support dialog systems and spoken task-oriented dialog systems. He has participated in several competitions on building dialog systems, including Dialog System Live Competition 6 and AIWolfDial2024jp. He is supervised by Assoc. Prof. Michimasa Inaba.

Muhammad Yeza Baihaqi

[1]Nara Institute of Science and Technology
[2]RIKEN
[1]8916-5 Takayama-cho, Ikoma, Nara
630-0192, Japan
[2]2-2-2 Hikaridai Seika-cho, Sorakugun,
Kyoto 619-0288, Japan

muhammad_yeza.baihaqi.lx2@naist.ac.jp
https://sites.google.com/view/mybaihaqi

1 Research interests

My research interest lies in the realm of **social interactive agents**, specifically in the development of **social agents for positively influencing human psychological states**. This interdisciplinary field merges elements of artificial intelligence, psychology, and human-computer interaction. My work integrates psychological theories with dialogue system technologies, including rule-based systems and large language models (LLMs). The core aim of my work is to leverage these systems to promote mental well-being and enhance user experiences in various contexts.

1.1 Social agent for psychological well-being

A significant focus of this research was designing dialogue systems that interact with users in ways that support their psychological well-being. In my master's thesis, I developed a social interactive robot aimed at reducing student anxiety during oral tests (Baihaqi, 2023). This study involved comparing different types of agent interactions that promote a positive psychological state by delivering certain dialogue such as small talk against a control group where the agent exhibited flat, robotic behavior. The triggers for delivering such dialogue were based on the similarity score of the student's answer to the answer key. In addition, I measured the students' anxiety levels during oral tests with a human examiner to provide a comprehensive comparison.

This study evaluated anxiety levels using subjective reports, behavioral observations, and physiological measures. Subjective reports were collected through structured questionnaires, behavioral observations were made by annotating recorded videos, and physiological states were assessed using self-designed measurement devices.

I have also undertaken several efforts in this field, including proposing a robot demonstration method to introduce social robotics to university students (Baihaqi and Xu, 2024), conducting a literature review on the practical applications of small robots as social robots and fuzzy techniques (Baihaqi and Xu, 2022a,c), exploring the emotion classification technique (Baihaqi et al., 2023), and designing customer service robot interactions in shopping malls using the seven stages of action (Baihaqi and Xu, 2022b).

1.2 Human-agent rapport

Currently, my doctoral research focuses on human-agent rapport, specifically exploring rapport-building dialogue strategies for multimodal dialogue agents (Baihaqi et al., 2024). Rapport refers to a warm and effortless connection marked by mutual comprehension, acknowledgment, and sympathetic harmony among individuals (Vanden-Bos, 2007). It ensures team members' sustained interest, involvement, and contentment, which eventually improves work results. Existing research highlights the vital role of rapport in enhancing task outcomes across various applications, such as healthcare (Johanson et al., 2020), tutoring (Sinha and Cassell, 2015), food services (Lee et al., 2012), and clinical interviews (Gratch et al., 2014).

Our research introduced a rapport-building dialogue strategy by integrating rapport-building utterances into the small talk with a virtual agent which was gathered from various successful existing studies of human-human rapport-building such as storytelling and praise expression. By integrating these curated utterances, our aim was to leverage the benefits of each utterance to diversify and enrich the agent's responses, ultimately enhancing the rapport between humans and agents.

The rapport-building dialogue strategy was embedded into the agent with two distinct strategies, free-form and predefined dialogue strategies. In the free-form strategy, the virtual agent gained the advantage of fostering a more natural and dynamic conversation, allowing users to express themselves authentically. This approach offered flexibility and adaptability, enhancing user engagement by responding to unique cues. However, drawbacks included potential inconsistency and missed opportunities for strategic rapport-building across many sessions. On the other hand, predefined elicitation ensured consistency and goal alignment but led to a more rigid and less personalized dialogue.

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 11–13
September 16–17, 2024.

The effectiveness of these strategies was assessed through questionnaires examining rapport scores and user experiences. Additionally, we are examining confounding factors such as total turn count and utterance length to understand their impact on rapport.

2 Spoken dialogue system (SDS) research

For the future of SDS research, I agree with the statements by Mattar et al. (2012). While the research on task-oriented SDS is well-developed, limiting SDS to task-oriented interactions is not sufficient and can negatively affect user experiences. Enhancing non-task-oriented dialogue systems is essential, as these systems engage users in trivial conversations, increasing engagement and satisfaction before crucial or main conversations. Remembering the importance of it, in the future, non-task-oriented dialogue system research is expected to become a trend.

In line with the growing importance of non-task-oriented dialogue, leveraging psychological theories will become a common strategy to achieve meaningful interactions. By incorporating these theories, SDS can exhibit more favorable behaviors, resulting in natural and relatable dialogue. However, implementing some theories will require a deep understanding of human utterances and non-verbal cues to discern implicit meanings, thereby increasing the demand for recognition techniques. For instance, recognizing when a human is not actively engaged in the conversation allows the agent to provide appropriate backchanneling behaviors. Instead of simply instructing the participant to pay attention, these behaviors may include employing psychological theories such as clarification, followed by reflective and active listening strategies. This approach may enhance the user experience, fostering connection between users and dialogue systems.

Last, unlike task-oriented dialogue that can be evaluated through computable metrics, non-task-oriented dialogue systems typically require human participants to evaluate performance. However, there is a growing discussion on utilizing LLMs to evaluate SDS. This method can possibly accelerate development through continuous feedback, reduce human evaluator biases, and save cost and time. It helps establish benchmarks and standards in SDS development. I believe this type of research will be a trend for the next five years.

3 Suggested topics for discussion

- **Implementing psychological theory to SDS:** Implementing psychological theory into SDS has primarily relied on rule-based methods and prompting LLMs. An alternative approach involves conducting human-human dialogue experiments and using the resulting dialogue corpus to train the language model. This enables the agent to adopt desired behaviors based on psychological principles. However, it requires a high cost and time to experiment and annotate. Is there a possible novel method to further integrate psychological theory with a language model?

- **Assessing human-agent rapport without human evaluation:** Currently, assessing rapport between humans and agents primarily relies on human evaluations through questionnaires. Some research has moved beyond questionnaires by observing specific behavioral patterns of participants. However, human evaluation is often questioned by reviewers for its subjectivity, generalization, and reliability. Is it possible to evaluate an SDS's ability to cultivate rapport using computable metrics, without relying on human evaluation?

- **Conducting SDS evaluation using other SDSs:** Evaluating SDS, especially for non-task-oriented SDS typically involves human interaction and feedback, which can be time-consuming and costly. Nowadays, there is a growing discussion about using other SDSs for evaluation purposes. As a result, there is also a growing debate about how closely the agent can replicate the human responses. Is it possible to build SDSs that can accurately represent human responses by incorporating diverse human personas?

References

Muhammad Yeza Baihaqi. 2023. *Human Behavioral, Subjective, and Physiological Assessments Under an Oral Test by a Humanoid Robot Examiner*. Master's thesis, National Taiwan University of Science and Technology.

Muhammad Yeza Baihaqi, Angel García Contreras, Seiya Kawano, and Koichiro Yoshino. 2024. Rapport-driven virtual agent: Rapport building dialogue strategy for improving user experience at first meeting. To appear in INTERSPEECH 2024.

Muhammad Yeza Baihaqi, Edmun Halawa, Riri Syah, Anniza Nurrahma, and Wilbert Wijaya. 2023. Emotion classification in indonesian language: A cnn approach with hyperband tuning. *Jurnal Buana Informatika* 14:137–146. https://doi.org/10.24002/jbi.v14i02.7558.

Muhammad Yeza Baihaqi and Sendren-Sheng Dong Xu. 2022a. The past and future of the fuzzy technique in social robots. In *Proceedings of the International Conference on Fuzzy Theory and Its Applications (iFuzzy 2022)*. page 1.

Muhammad Yeza Baihaqi and Sendren-Sheng Dong Xu. 2022b. Seven stages of action for the interaction design of the customer service of a robot in a shopping mall. In *Proceedings of the International Conference on Advanced Robotics and Intelligent Systems (ARIS 2022)*. pages 1–2.

Muhammad Yeza Baihaqi and Sendren-Sheng Dong Xu. 2022c. A survey on practical applications of small robots as social robots. In *Proceedings of the International Conference on System Science and Engineering 2022 (ICSSE 2022)*. page 1.

Muhammad Yeza Baihaqi and Sendren Sheng-Dong Xu. 2024. Impact of showing robot demonstration on introducing social robotics field to university students. *International Journal of Humanoid Robotics* 21(02):2350018–2350041. https://doi.org/10.1142/S0219843623500184.

Jonathan Gratch, Gale M. Lucas, Aisha Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: the impact of human-agent interaction in clinical interviews. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '14, page 85–92.

Deborah L. Johanson, Ho Seok Ahn, Craig J. Sutherland, Bianca Brown, Bruce A. MacDonald, Jong Yoon Lim, Byeong Kyu Ahn, and Elizabeth Broadbent. 2020. Smiling and use of first-name by a healthcare receptionist robot: Effects on user perceptions, attitudes, and behaviours. *Paladyn, Journal of Behavioral Robotics* 11(1):40–51. https://doi.org/doi:10.1515/pjbr-2020-0008.

Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Paul Rybski, John Antanitis, and Sarun Savetsila. 2012. Personalization in hri: A longitudinal field experiment. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, HRI '12, pages 319–326. https://doi.org/10.1145/2157689.2157804.

Nikita Mattar et al. 2012. Small talk is more than chitchat. In *KI 2012: Advances in Artificial Intelligence*. pages 119–130.

Tanmay Sinha and Justine Cassell. 2015. We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And InfLuence*. Association for Computing Machinery, New York, NY, USA, INTERPERSONAL '15, pages 13–20. https://doi.org/10.1145/2823513.2823516.

G. R. VandenBos. 2007. *APA dictionary of psychology*. American Psychological Association, Washington, D. C.

Biographical sketch

Muhammad Yeza Baihaqi earned his B.Eng. degree from President University in 2020, where he was recognized as the top graduate in Electrical Engineering. Subsequently, he completed his M.Sc. degree in the Graduate Institute of Automation and Control at the National Taiwan University of Science and Technology in 2023, receiving accolades as an Outstanding Student in the College of Engineering.

Currently, he is pursuing his Ph.D. in Information Sciences at the Nara Institute of Science and Technology. His research interests include social interactive agents and dialogue systems for psychological well-being.

Kai Yoshida

Nara Institute of Science and Technology / RIKEN GRP, Nara, Japan

yoshida.kai.yf1@is.nasit.jp
https://sites.google.com/view/kai-yos/

1 Research interests

In our research, we aim to achieve SDS capable of generating responses considering user preferences. While users have individual topic preferences, existing SDSs do not adequately consider such information. With the development of LLMs, SDSs are expected to be implemented in various tasks, including coexisting with humans in robotic applications. To become better partners with humans, systems are anticipated to memorize user preferences and utilize them in their response generation. Our future reserarch aim to realize SDSs that can remember and complement user information through dialogue, enabling personalized interactions.

1.1 Personalized Dialogue System

Persona dialogue is a dialogue task where systems generate responses by referring profile information called personas, aiming to induce certain social behaviors in LLMs. A persona refers to information such as desired personality traits and background that systems should exhibit, enabling SDSs to provide more natural and human-like conversations. The primary goal of persona dialogue is to enhance the naturalness, consistency of system personality, and character, thereby increasing user engagement. For instance, setting information like "Name: Alice, Age: 25, Occupation: Virtual Assistant, Hobbies: Reading, Traveling, Music" allows interactions with users based on this setup. When a user asks, "Hello, Alice. How's the weather today?" Alice might respond, "Hello! It's sunny today, and the temperature is warm. Perfect weather for a walk. Have you read any new books recently?" This setup enables SDSs to deliver human-like dialogues tailored to specific backgrounds.

An important aspect of this task is that it's impractical to pre-define all system profile information, leading to hallucinations where new profile information emerges in responses as dialogue turns increase. Failing to consider such hallucinated personas may result in inconsistencies in system character and lack of response coherence in subsequent dialogues.

Our past research [Yoshida et al. (2024b)] addressed this challenge by extracting and storing persona information from generated texts for retrieval-based response generation. However, effective methods have yet to be proposed due to challenges in experimental setups for long-term dialogues and factors like language model generation accuracy.

In future work, based on the knowledge gained so far, we aim to design long-term dialogue experiments and response generation systems to address the issue of persona hallucination in long-term dialogues.

1.2 Topic Transition on Dialogue

The research interests in Section 1 can be further divided into subsections, as found appropriate by the author. For SDSs and dialogue robots to establish rapport with users, it is anticipated that system-level personalization of users is necessary. Hence, our research focuses on topic transitions in dialogues. While users have preferences for specific topics, existing LLMs do not explicitly utilize these preferences for response generation. For instance, offering discussions on baseball to users who prefer it can enhance dialogue engagement. The goal is for systems to provide personalized dialogues based on such user preferences, fostering rapport between humans and systems.

As a preliminary step towards this goal, we have previously explored methods to naturally transition topics from current to desired topics by inducing word associations in LLMs [Yoshida et al. (2024a)]. This approach has suggested that it enables more natural and diverse transitions compared to transition methods using knowledge graphs.

However, we have not yet addressed the realization of personalized topic transitions using user information. Therefore, future efforts will focus on ini tiatives like biasing transition content based on user information.

1.3 Automatic Evaluation of Dialogue Topic Transition

While the dialogue performance of LLMs is advancing daily, current LLMs adopt passive dialogue strategies and lack the ability to lead conversations. Therefore, our research focuses on topic transitions in dialogues to enable SDSs to acquire the ability to lead conversations. Specifically, we are working on achieving personalized topic transitions for each user, which is crucial for LLMs to take the lead in dialogues.

One major challenge in this endeavor is the lack of automated evaluation metrics for assessing the naturalness of topic transitions. Existing studies often use

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 14–15
September 16–17, 2024.

benchmarks that measure the accuracy of topic transitions against correct labels in datasets, but reference-free evaluation metrics for topic transitions are still insufficient. In topic transitions, due to the characteristic that the next topic candidate is not uniquely determined, reference-free evaluation metrics are suitable for reference-based ones. Moreover, automated evaluation metrics for topic transitions are important for inferring natural transition targets.

Therefore, our future research will focus on developing reference-free automated evaluation metrics for topic transitions.

2 Spoken dialogue system (SDS) research

In the coming years, SDS research is expected to split into two major directions: practical applications and the study of dialogue mechanisms. Moreover, these areas are not independent of each other but are expected to mutually influence one another.

2.1 Interaction of dialogues

To generalize SDSs more broadly, it is necessary to make them more appealing to users. To achieve this, it is important to study the interactions between SDSs and users.

Before the advent of ChatGPT, dialogue research primarily focused on generating natural sentences or producing sentences according to specifications. However, with the advancement of LLMs, it has become possible to generate reasonably natural responses to given contexts. Consequently, the groundwork is being laid for examining human interaction when using LLMs as agents.

Given these developments, it is anticipated that future research will focus heavily on how generated sentences affect users and what dialogue strategies should be employed to influence users.

Therefore, advancing this research will likely require approaches that include fields such as psychology and linguistics. Hence, future researchers in the SDS field will need interdisciplinary knowledge that extends beyond engineering alone.

2.2 Social Adaptation

To further generalize and popularize SDS, it is essential to address problems set in real-world environments. For example, long-term dialogues spanning multiple sessions and turn-taking in multi-party conversations are expected to become important. Investigating the gap between SDS applications and experimental environments and establishing these as defined tasks is necessary. Additionally, collaboration with companies developing these services is crucial.

3 Suggested topics for discussion

The author would like to propose the following topics for discussion.

- What is the necessity of SDSs aimed specifically at dialogue rather than being just user interfaces? What do general users need from SDSs through conversation?
- The relationship between SDSs and users: Should SDSs act just as agents, or should they aim to become like friends or family?
- Privacy in conversational content. Nowadays, many SDS applications operate online via APIs, but is this preferable from a privacy perspective? If it is not preferable, how can this issue be resolved?

4 Acknowlegments

References

Kai Yoshida, Seiya Kawano, and Koichiro Yoshino. 2024a. *Analysis of the Topic Transition Graph Using Word Association based on Large Language Model*. Proceedings of the Annual Conference of JSAI2024, 4Xin2-94 (in Japanese).

Kai Yoshida, Koichiro Yoshino, Seitaro Shinagawa, Katsuhito Sudoh, and Satoshi Nakamura. 2024b. *Persona Selection and Response Generation in Persona-based Dialogue Systems*. Proceedings of the Thirtieth Annual Meeting of the Association for Natural Language Processing (in Japanese).

Biographical sketch

Kai Yoshida is a PhD student at NAIST in Japan. His research interests lie in personalized dialogues, open-domain dialogues, and in their interaction. As part of his master's thesis, he worked on persona dialogue system on long-term conversation settings. He enjoys eating ramen, reading comic book and discovering new ramen.

Maeda Shio

International Institute for Integrative Sleep Medicine University of Tsukuba
1-2-2 Kasuga
Tsukuba, Ibaraki 305-0821

maeda.shio.gn@u.tsukuba.ac.jp

1 Research Interests

My research aims to use **multimodal data in psychotherapy** to develop optimal analytical models for various information generated during therapy, to elucidate the process of psychotherapy, and to **create AI therapists** to develop new psychotherapies.

1.1 Past Research

The applicant has been conducting research in the field of psychology to elucidate the maintenance and prediction of symptoms of mental illness and other disorders and the transformation process in treatment from multimodal data such as facial expressions (Maeda and Yokotani, 2022; Maeda and Yokotani, 2023).

1.2 Current and planned future research

Current and future research that we are conducting aims to elucidate the process of psychotherapy using multimodal high-resolution data while utilizing the qualifications of licensed psychologists. As background for this study, psychotherapy is in high demand in society, but its effectiveness is only 60%. In recent decades, although enormous amounts of research funds and the time of many researchers worldwide have been devoted to improving the effectiveness of psychotherapy, the effectiveness of psychotherapy has come to a head (Leichsenring et al., 2022). To improve the effectiveness of psychotherapy, it is necessary to "dismantle" psychotherapy and verify the effectiveness of each component（Boschloo et al., 2019）, and multimodal machine learning is effective for this purpose. Therefore, we will establish an analytical model of psychotherapy using a registry(Psychotherapy Registry R-MAP : KAKENHI 22K20312) that stores high-resolution data during psychotherapy, which is owned by the laboratory to which the applicant belongs. We will also apply the obtained model to clinical data to clarify its usefulness in clinical practice.

Furthermore, she is working on the development of AI therapists, aiming to achieve natural communication with patients by utilizing voice interaction technology. With advancements in this voice interaction technology, AI therapists will be able to understand patients' emotions and intentions more accurately and provide personalized treatment anytime and anywhere. Figure 1 shows a demo screen of an actual implementation of an AI therapist, serving as a supportive bot for people experiencing parenting loneliness.

Figure 1 Demo screenshot of the snuggle bot for lonely childcare people.

2 Future of Spoken Dialog Research

In five years, voice dialogue research is expected to achieve further advancements, establishing technologies that enable more appropriate and natural interactions. These technologies will allow for a more accurate understanding of human speech and intentions, and the generation of appropriate responses. In ten years, these technologies are predicted to become widely adopted, with voice dialogue systems being utilized in various aspects of daily life, such as healthcare, education, and entertainment.

In the meantime, young researchers are expected to significantly contribute to improving the accuracy and practical application of voice dialogue systems. Specifically, they will aim to develop systems that can handle diverse languages, dialects, and accents. Additionally, research to improve the usability and safety of voice dialogue systems will be important. Furthermore, by working on the development of multimodal dialogue systems that integrate different modalities (e.g., voice,

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 16–17
September 16–17, 2024.

visual, gestures), they can provide a richer user experience.

To achieve the above goals, research on interface design and usability to create easy-to-use and safe dialogue systems will be necessary. Additionally, the development and evaluation of dialogue systems that combine modalities other than voice will be essential. Furthermore, to promote the field, research on the ethical issues and social impacts brought by the widespread use of dialogue systems and measures to address these issues will be indispensable. Through these studies, it will be possible to build the future of voice dialogue research and provide technologies beneficial to society.

3 Suggestions for discussion

- Accurate Detection of Emotions and Intentions and Optimization of Responses: Analytical techniques for understanding emotions and intentions and generating more human-like responses.
- Interface Design: Designing interfaces that affect how users interact with the system (visual and auditory design).
- Interpretation of Emotions and Other Paralinguistic Phenomena: Interpretation of Multimodal Information such as Emotions.

References

Boschloo, L., Bekhuis, E., Weitz, E. S., Reijnders, M., DeRubeis, R. J., Dimidjian, S., ... & Cuijpers, P. (2019). The symptom‐specific efficacy of antidepressant medication vs. cognitive behavioral therapy in the treatment of depression: Results from an individual patient data meta‐analysis. *World Psychiatry*, 18(2), 183-191.

Leichsenring, F., Steinert, C., Rabung, S., & Ioannidis, J. P. (2022). The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: an umbrella review and meta‐analytic evaluation of recent meta‐analyses. *World Psychiatry*, 21(1), 133-145.

Maeda, S., & Yokotani, K. (2022). Development of a detection model for negative facial expression in individuals with alexithymia tendencies. *Paper presented at the 48th Annual Meeting of the Japanese Association for Cognitive and Behavioral Therapies*, 242-243.

Maeda, S., Yokotani, K., Yokomitsu, K., Irie, T., & Kamata, M. (2023). Examination of emotional expression and synchrony of facial expressions between gamblers and therapists during psychological interviews. *Paper presented at the 49th Annual Meeting of the Japanese Association for Cognitive and Behavioral Therapies*, 474-475.

Biographical Sketch

The applicant obtained her Master's degree in Clinical Psychology from the Graduate School of Integrated Arts and Sciences, University of Tokushima, in March 2024. The applicant also earned the national qualification of Certified Public Psychologist in June of the same year. Recognized for her active research activities, innovative ideas, and analytical techniques during her graduate studies, Shio is currently working as the youngest researcher at the International Institute for Integrative Sleep Medicine (WPI-IIIS) at the University of Tsukuba, a research institution selected for the World Premier International Research Center Initiative (WPI) by the Ministry of Education, Culture, Sports, Science and Technology.

At WPI-IIIS, The applicant participates in various projects, including two clinical trials on new psychological therapies for insomnia utilizing technology (jRCT1030210575, jRCT1030210518) and research on the societal implementation of psychological therapies based on implementation science (UMIN000052911), under the Moonshot Research and Development Program titled "Unraveling and Manipulating the Two Types of Sleep: Sleep and Hibernation for New Generation Medical Development." This year, she has also joined as an analyst in the development of next-generation psychological therapies using human augmentation technologies (Grant-in-Aid for Scientific Research (B): 24K00492), contributing to research that is expected to gain international attention. Additionally, the applicant is collaborating with Dr. Francis X. Shen, a leading expert in the Ethical, Legal, and Social Issues (ELSI) of digital mental health, to conduct research on the ELSI of digital mental health. Together, they are planning and conducting patient surveys on AI agents to examine the facilitators and barriers to research data utilization, which is one aspect of Patient and Public Involvement (PPI) (jRCT1030220228).

Mao Saeki

Waseda University
Wasedamachi-27, Shinjuku-ku
Tokyo, Japan

saeki@pcl.cs.waseda.ac.jp
https://saeki-m.github.io/

1 Research interests

My research interests lie in **multimodal dialog systems**, especially in **turn-taking** and the understanding and generation of **non-verbal cues**. I am also interested in bringing dialog system research into **industry**, and making virtual agents practical in real world setting.

I have been working on the Intelligent Language Learning Assistant (InteLLA) system, a virtual agent designed to provide fully automated English proficiency assessments through oral conversations[1]. This project is driven by the practical need to address the lack of opportunities for second-language learners to assess and practice their conversation skills.

While recent advancements in large language models (LLMs) have enabled natural conversation, effective assessment requires several components that current LLMs cannot fully achieve. Based on interviewer guidelines for oral proficiency assessment (Liskin-Gasparro, 2003), the following functionalities have been identified as necessary for the agent to possess for an effective assessment:

1. Automated Proficiency Assessment
2. Confusion Detection
3. Multimodal Turn-Taking Prediction

My past research has focused on solving each of these problems, which will be explained in subsequent sections. Additionally, these research outcomes have been integrated into the InteLLA agent and evaluated for its effectiveness in end-to-end assessment.

1.1 Automated Proficiency Assessment

For scalable and reliable evaluations, an automated assessment model is essential. While handcrafted lexical and acoustic features have been extensively investigated for assessment, end-to-end approaches, particularly those utilizing visual features like facial expressions and eye gaze—critical components of interaction—have not been thoroughly explored. I proposed a multimodal oral proficiency assessment model incorporating lexical, prosodic, and visual cues (Saeki et al., 2021). The results demonstrated that end-to-end approaches using deep neural networks achieve a higher correlation with human scoring compared to those employing handcrafted features. Furthermore, the effectiveness of the modalities was found to be in the order of lexical, acoustic, and visual features.

Assessment is also important during the interview. For the user to demonstrate their full range of ability, they must be challenged with appropriate levels of questions, according to assessment during the interview. Saeki et al. (2022a) investigated the feasibility of incremental assessment of oral proficiency using an adaptive test format.

1.2 Confusion Detection

Language learners often face confusion, where they fail to understand what the system has said and may be unable to respond, leading to a conversational breakdown. Detecting such states and keeping the conversation moving forward by repeating or rephrasing system utterances is crucial. In Saeki et al. (2022b) we collected a dataset of user confusion using a psycholinguistic experimental approach and identified seven multimodal signs of confusion, some unique to online conversations. We trained a classification model of user confusion using these features. An ablation study showed that features related to self-talk and gaze direction were most predictive.

1.3 Multimodal Turn-Taking Prediction

Language learners often produce long silences while formulating their responses. Such pauses should not be interrupted, however, the system should promptly take its turn when the user has finished speaking to achieve a natural conversation flow. While the effectiveness of visual cues—such as gaze, mouth, and head movements—has been suggested, few studies have fully incorporated them into turn-taking models. We proposed a multimodal model for predicting the end-of-turn probability in spoken dialogue systems(Kurata et al., 2023). An ablation study on visual features showed that eye movements contributed more significantly than mouth and head movements. Additionally, an end-to-end visual feature extraction model utilizing 3D-CNN was employed to comprehensively capture these visual cues. Combining visual features with acoustic and verbal information, the AUC score for end-of-turn prediction improved from 0.896 to 0.920, demonstrating the effectiveness of these visual cues.

[1] https://youtu.be/RzCq5Z4cDBk?feature=shared

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 18–20
September 16–17, 2024.

1.4 InteLLA System Evaluation

The primary challenge in using dialogue systems for reliable language assessment of interactional skills lies in obtaining ratable speech samples that demonstrate the user's full range of abilities. We developed a multimodal dialogue system that employs adaptive sampling strategies and enables mixed-initiative interaction through extended interviews and role-play dialogues (Saeki et al., 2024). The interview is a system-led dialogue aimed at evaluating the user's overall proficiency. The system dynamically adjusts question difficulty based on real-time assessment to induce linguistic breakdowns, providing evidence of the user's upper proficiency limits. The role-play, on the other hand, is a mixed-initiative, collaborative conversation intended to assess interactional competence such as turn management skills.

Two experiments were conducted to evaluate our system in assessing oral proficiency. In the first experiment, involving an interview dataset of 152 speakers, our system demonstrated high accuracy in automatically assessing overall proficiency. However, linguistic breakdowns were less likely to occur among high-proficiency users, indicating room for improving the ratability of speech samples. In the second experiment, based on a role-play dataset of 75 speakers, the speech samples elicited by our system were as ratable for interactional competence as those elicited by experienced teachers, demonstrating our system's capability in conducting interactive conversations. Finally, we reported on the deployment of our system with over 10,000 students in two real-world testing scenarios.

1.5 Future Planned Work

Using the InteLLA system, a future direction I am planning is to automatically evaluate interactional competence the user is able to demonstrate. Interactional competence is an important metric in the context of language assessment; however, I believe it could also benefit Spoken Dialogue Systems (SDS). For example, interactional competence has been identified in the field of language assessment to include functions such as turn management strategy, which consists of timing, turn-allocation, overlap resolution, and preference organization. Measuring and closing the gap of interactional competence of turn management strategy between an SDS and a human interlocutor would mean the SDS is recognized more similarly to a human interlocutor, which is a measure of improvement for the SDS. Furthermore, if the relationship between the interlocutor's and user's turn management strategies is identified, we can effectively improve the system to achieve a more authentic spoken dialog experience.

2 Spoken dialogue system (SDS) research

In light of recent advancements in large language models (LLMs) and multimodal LLMs, I believe that this generation will witness the widespread usage of spoken dialogue systems (SDS) in everyday life. Currently, many SDS frameworks depend on multiple models, external APIs, and networks. An imperative field of research in the coming years will be the continuous monitoring and quality assurance of these complex systems. Automatic detection of bugs and conversation experience issues will be crucial for the widespread usage of SDS. Additionally, it will be essential to ensure that subsequent changes do not introduce new problems or degrade overall system performance.

Another exciting advancement for SDS would be the development of fully end-to-end models. For instance, models like GPT-4o exhibit impressive expressiveness; however, they are likely not yet capable of fully interactive conversations as they process user utterances in chunks. Such models would struggle with complex turn-taking phenomena like allowing users time to think, backchanneling, and handling overlapping responses without external modules. Research on incremental models that process and output audio in real-time will be essential to overcome these limitations and achieve more natural and fully duplex interactions.

3 Suggested topics for discussion

- Quality Assurance and Testing of SDS: With the rapid deployment of Spoken Dialogue Systems (SDS) in real-world applications, similar to other industrial software, automated testing is becoming crucial. How can we ensure that updates to the system do not degrade performance and genuinely improve the conversational experience? What methodologies or frameworks can be employed for objective and automated testing?

- Identifying and Implementing Improvements: Virtual agents in SDS have several aspects that can be enhanced, such as speech content, Text-to-Speech (TTS) quality, turn-taking mechanisms, and motion. How can we efficiently pinpoint bottlenecks in user experience to prioritize and implement improvements effectively?

- Leveraging Larger and Multimodal Models: The advent of large language models (LLMs) with multimodal capabilities suggests significant potential for handling spoken dialogue with natural speech. Is this the future direction for SDS development? What roles can university researchers and young professionals without huge computational resource play in the race for bigger models?

References

Fuma Kurata, Mao Saeki, Shinya Fujie, and Yoichi Matsuyama. 2023. Multimodal Turn-Taking Model Using Visual Cues for End-of-Utterance Prediction in Spoken Dialogue Systems. In *Proc. INTERSPEECH 2023*. pages 2658–2662. https://doi.org/10.21437/Interspeech.2023-578.

Judith E. Liskin-Gasparro. 2003. The ACTFL Proficiency Guidelines and the Oral Proficiency Interview: A brief history and analysis of their survival. *Foreign Language Annals* 36(4):483–490. https://doi.org/10.1111/j.1944-9720.2003.tb02137.x.

Mao Saeki, Weronika Demkow, Tetsunori Kobayashi, and Yoichi Matsuyama. 2022a. A woz study for an incremental proficiency scoring interview agent eliciting ratable samples. In *Conversational AI for Natural Human-Centric Interaction*. Springer Nature Singapore, Singapore, pages 193–201.

Mao Saeki, Masaki Eguchi, Hiroaki Takatsu, Shungo Suzuki, Shungo Suzuki, Fuma Kurata, Ryuki Matsuura, Kotaro Takizawa, Sadahiro Yoshikawa, and Yoichi Matsuyama. 2024. Intella: Intelligent language learning assistant for assessing language proficiency through interviews and roleplays. *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue* page to appear.

Mao Saeki, Yoichi Matsuyama, Satoshi Kobashikawa, Tetsuji Ogawa, and Tetsunori Kobayashi. 2021. Analysis of multimodal features for speaking proficiency scoring in an interview dialogue. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. pages 629–635. https://doi.org/10.1109/SLT48900.2021.9383590.

Mao Saeki, Kotoka Miyagi, Shinya Fujie, Shungo Suzuki, Tetsuji Ogawa, Tetsunori Kobayashi, and Yoichi Matsuyama. 2022b. Confusion detection for adaptive conversational strategies of an oral proficiency assessment interview agent. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2022-September:3988–3992. https://doi.org/10.21437/Interspeech.2022-10075.

Biographical sketch

Mao Saeki is a Ph.D. student in Computer Science at Waseda University. His research interests focus on multimodal conversational AI, particularly in the understanding and generation of non-verbal cues. He developed the InteLLA system, a virtual agent designed to automatically assess the English proficiency of language learners. Additionally, he is a founding member of Equmenopolis Inc., a company dedicated to integrating SDS research into societal applications.

Patrícia Schmidtová

Charles University
Institute of Formal and Applied Linguistics
Prague, Czech Republic

schmidtova@ufal.mff.cuni.cz
ufal.mff.cuni.cz/patricia-schmidtova

1 Research interests

My research interests lie in the area of **natural language generation** (NLG), more specifically, I focus on the **faithfulness** of NLG. Following Maynez et al. (2020), we define faithfulness as adherence to a given set of inputs (such as a result of a database lookup or a system action). These inputs can either be given by the user with the goal of a given transformation (e.g. data-to-text generation or summarization), or by a dialogue system to compose a reply to the user.

In contrast to numerous works that focus on factuality, i.e. the real-world truth value of a statement, (Azaria and Mitchell, 2023; Lin et al., 2022), I believe that faithfulness is a more useful quality in the realm of **dialog systems** since it measures whether the user received the information they asked for. Thus, my research is directly applicable to the task of **dialog response generation**.

My research is guided by two research questions:

1. How can we determine if a generated text is faithful to its source data?
2. Which factors affect the faithfulness of an LLM's output and how can we manipulate them to achieve better accuracy?

In the following sections, I will outline my progress and plans for how to evaluate faithfulness and thus answer the first research question (Sec 1.1), my plans to understand and improve the faithfulness of systems to seek answers to the second research question (Sec 1.2), and my previous work on treating script generation as a dialogue system task (Sec 1.3).

1.1 Evaluation of faithfulness

There are several challenges when designing a robust protocol to evaluate faithfulness. Most metrics rely on the presence of gold reference data (Papineni et al., 2002; Zhang et al., 2020; Kane et al., 2020) that is not always available. Furthermore, some problems have more than one correct solution, and comparing to an arbitrary reference might not always favor the best outputs.

Additionally, in our work examining **data contamination** (i.e. presence of testing data in the training data) (Balloccu et al., 2024), we found that many datasets with gold annotations, such as several variants of MultiWOZ (Budzianowski et al., 2018; Eric et al., 2020; Ye et al., 2022) and some datasets used for DSTC (Zhao et al., 2023) were leaked to closed-source language models by users. With works examining the presence of datasets in CommonCrawl (Li et al., 2024), we cannot even be entirely sure that open-weight models with secret training data, such as Mistral (Jiang et al., 2023) or Llama2 (Touvron et al., 2023) are free from data contamination. This casts a shadow of doubt on whether the models truly generalize well or whether a part of their success is due to data contamination.

Therefore, in my research, I will focus on reference-free evaluation methods that can be used on freshly mined data, such as the QUINTD dataset (Kasner and Dušek, 2024). We have seen some success using LLMs as evaluators for dialogue response generation (Plátek et al., 2023) and we are currently extending this work on new datasets and with comparison to crowd-workers of several proficiency levels determined based on a qualification screening test. Currently, there are also reference-free metrics based on **natural language inference** (NLI), however, they are not yet equipped to deal with structured data or with data of various lengths. We intend to address this issue in our future work.

We do not intend to replace human evaluation using these methods since insights gained by a well-performed human analysis are unparalleled. We rather see automatic evaluation as a proxy for situations where time and resources are limited, such as in a development cycle when trying to estimate the effect of a change. Additionally, human and automatic evaluation should complement each other to assess the strengths and weaknesses of a system comprehensively.

To simplify human (or LLM) annotation of LLM faithfulness errors, my colleagues and I developed a tool called factgenie[1] (Kasner et al., 2024), which will be presented at INLG as a demo the week after YRRSDS. Finally, we have prepared a comprehensive survey of how automatic evaluation is generally performed in NLG and extended a set of best practices (Schmidtová et al., 2024). This work will also be presented at INLG. One of our

[1] https://github.com/kasnerz/factgenie

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 21–24

main findings was that evaluation in NLG is currently very divided. The most prominently used metrics are based on N-gram overlap, such as BLEU (Papineni et al., 2002), which is unfortunate, since Reiter (2018) shows that they have little informational value in NLG.

1.2 Understanding and improving faithfulness

When trying to understand the faithfulness of LLMs to a given input, prompts are the easiest external factor to examine. Axelsson and Skantze (2023) observed that asking an LLM to stick to the provided facts indeed increases their faithfulness. In our research, we intend to explore how various circumstances, such as prompt length, grammatical correctness, or the presence of specific instructions, affect the faithfulness of a language model.

Moreover, we also intend to use **probing** to observe how the different prompts activate different parts of the network and thus elicit different results. We draw inspiration from work where probing was used to seek out and modify facts stored in LLMs' trained weights (Meng et al., 2022) or to classify if an LLM believes that a statement supplied by the user on the input is true (Azaria and Mitchell, 2023).

1.3 Previous work on theatre play script generation

The majority of my past work on theatre play generation was performed with a single language model predicting the next character utterance (Schmidtová et al., 2022). However, one of the downsides of this approach was the lack of consistency in the characters' personalities. As a small side project, we decided to treat this task as a conversation between three language models, each of them fine-tuned to represent a separate character (Schmidtová et al., 2022). To keep things simple, we classified characters in movie scripts into pessimists, optimists, and realists by observing the average sentiment of their utterances. We showed that by training each model separately, the consistency of characters was indeed improved.

2 Spoken dialogue system (SDS) research

The arrival of large language models trained using reinforcement learning from human feedback changed the way how the public perceives dialogue systems and what to expect from them. I believe there are two directions in research we should pay attention to in the next 5-10 years:

Multidisciplinary collaboration We might not even be aware of all the ways how the public uses dialogue systems and often only hear about the use cases where something went wrong, such as a lawyer citing non-existent cases (Merken, 2023). I believe it is important to connect with other fields, especially psychology, to have a better understanding of how SDSs impact the users so we can make more informed decisions about how we design and present the systems to make them safer.

Educating the public Last, but not least, we see many public figures make bold statements about how LLMs will make entire careers, such as programmers, obsolete. Generally, the boldest claims do not come from researchers, but rather from executives seeking to increase profits of the companies they run. For this reason, I believe that communicating research to the public has grown equally as important as the research itself. Scientists should be the figures that the public looks to with questions, yet they are often not very visible outside of academic grounds. As young researchers, we can start small, for example by giving talks to high school students or interested communities around us.

3 Suggested topics for discussion

These are the topics I would like to suggest for discussion:

- **Data contamination**: to what extent should we examine and worry about dialogue datasets being contained in CommonCrawl or the training sets of closed-source models?
- **Evaluation**: should we strive for a more general and unified set of evaluation practices or rather try to adapt the metrics used to the presented dialogue system?
- **Multidisciplinary collaboration**: Other fields, such as robotics or social sciences can be very beneficial to SDS and provide insights to make them better and safer. On the other hand, the structure and funding distribution of universities does not always favor such collaboration. How do others tackle this, if at all?

Acknowledgements

This research was co-funded by the European Union (ERC, NG-NLG, 101039303) and by Charles University projects GAUK 40222 and SVV 260 698.

References

Agnes Axelsson and Gabriel Skantze. 2023. Using large language models for zero-shot natural language generation from knowledge graphs. In Albert Gatt, Claire Gardent, Liam Cripwell, Anya Belz, Claudia Borg, Aykut Erdem, and Erkut Erdem, editors, *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*. Association for Computational Linguistics, Prague, Czech Republic, pages 39–54. https://aclanthology.org/2023.mmnlg-1.5.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 967–976. https://doi.org/10.18653/v1/2023.findings-emnlp.68.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, St. Julian's, Malta, pages 67–93. https://aclanthology.org/2024.eacl-long.5.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pages 5016–5026. https://doi.org/10.18653/v1/D18-1547.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 422–428. https://aclanthology.org/2020.lrec-1.53.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*. Association for Computational Linguistics, Online (Dublin, Ireland), pages 28–37. https://aclanthology.org/2020.evalnlgeval-1.4.

Zdeněk Kasner and Ondřej Dušek. 2024. Beyond reference-based metrics: Analyzing behaviors of open llms on data-to-text generation.

Zdeněk Kasner, Ondřej Plátek, Patrícia Schmidtová, Simone Balloccu, and Ondřej Dušek. 2024. factgenie: A framework for span-based evaluation of generated texts. https://arxiv.org/abs/2407.17863.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. An open source data contamination report for large language models. https://arxiv.org/abs/2310.17589.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, pages 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*. https://openreview.net/forum?id=-h6WAS6eE4.

Sara Merken. 2023. New york lawyers sanctioned for using fake chatgpt cases in legal brief. https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Ondřej Plátek, Vojtech Hudecek, Patricia Schmidtova, Mateusz Lango, and Ondrej Dusek. 2023. Three ways of using large language models to evaluate chat. In Yun-Nung Chen, Paul Crook, Michel Galley, Sarik Ghazarian, Chulaka Gunasekara, Raghav Gupta, Behnam Hedayatnia, Satwik Kottur, Seungwhan Moon, and Chen Zhang, editors, *Pro-*

ceedings of The Eleventh Dialog System Technology Challenge. Association for Computational Linguistics, Prague, Czech Republic, pages 113–122. https://aclanthology.org/2023.dstc-1.14.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics* 44(3):393–401. https://doi.org/10.1162/coli-a-00322.

Patrícia Schmidtová, Rudolf Rosa, David Košťák, Tomáš Studeník, Daniel Hrbek, Tomáš Musil, Josef Doležal, Ondřej Dušek, David Mareček, Klára Vosecká, Máša Nováková, Petr Žabka, Alisa Zakhtarenko, Dominik Jurko, Martina Kinská, Tom Kocmi, and Ondřej Bojar. 2022. *THEaiTRE: Generating Theatre Play Scripts using Artificial Intelligence*. Institute of Formal and Applied Linguistics, Prague, Czechia.

Patrícia Schmidtová, Dávid Javorský, Christián Mikláš, Tomáš Musil, Rudolf Rosa, and Ondřej Dušek. 2022. Dialoguescript: Using dialogue agents to produce a script. https://arxiv.org/abs/2206.08425.

Patrícia Schmidtová, Saad Mahamood, Simone Balloccu, Ondřej Dušek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondřej Plátek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. To appear at INLG 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy Li, Arash Ashrafzadeh, Daniel Hernández Garcia, Malihe Alikhani, David Vandyke, and Ondřej Dušek, editors, *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Edinburgh, UK, pages 351–360. https://doi.org/10.18653/v1/2022.sigdial-1.34.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr.

Chao Zhao, Spandana Gella, Seokhwan Kim, Di Jin, Devamanyu Hazarika, Alexandros Papangelis, Behnam Hedayatnia, Mahdi Namazifar, Yang Liu, and Dilek Hakkani-Tur. 2023. "what do others think?": Task-oriented conversational modeling with subjective knowledge.

Biographical sketch

Patrícia Schmidtová is a second-year PhD student at Charles University advised by Ondřej Dušek. Currently, she is researching how to comprehensively and reliably evaluate large language models, especially the faithfulness of their outputs to the data they are given. She plans to explore why LLMs respond to some prompts better than others by using interpretability techniques.

Before her PhD, she was a member of the THEaiTRE research team which succeeded in producing the world's first AI-scripted theater play. She also has six years of industry experience in NLP application development, mostly working on devising components for robotic process automation (RPA).

Hiroki Onozeki

The University of Electro-Communications
Chofu, Tokyo
Japan

o2330023@edu.cc.uec.ac.jp

1 Research interests

My research interests lie in the area of **building a dialogue system to generate interesting and entertaining responses**, with a particular focus on knowledge-grounded dialogue systems. Study of open-domain dialogue systems seeks to maximize user engagement by enhancing specific dialogue skills. To achieve this goal, much research has focused on the generation of empathetic responses, personality-based responses, and knowledge-grounded responses (Algherairy and Ahmed, 2024). In addition, interesting responses from the open-domain dialogue systems can increase user satisfaction and engagement due to their diversity and ability to attract the user's interest. Interesting responses are defined here as those that contain facts not generally well known but that provide surprise and engage the user. It has also been observed in task-oriented dialogue, user engagement can be increased by incorporating interesting responses into the dialogue. For example, Vicente et al. (2023) incorporated interesting responses into the spoken dialogue systems (SDSs) to support the user in performing complex tasks, making the experience pleasant and enjoyable for the user. However, even in the case of interesting responses, if the dialogue is incoherent, user engagement is likely to be significantly reduced. To create a dialogue system that is consistent and interesting in a dialogue context, I am working on using knowledge-grounded response generation methods to select interesting knowledge that is relevant to the dialogue context and to make responses that are based on that knowledge.

1.1 Introducing interesting knowledge into dialogue systems

Several studies have been conducted to investigate the use of interesting knowledge in dialogues to achieve engaging dialogues. Konrád et al. (2021) built a dialogue system that uses interesting knowledge obtained from crawling from Reddit with high similarity to the dialogue context in the dialogue for improving user engagement in open-domain dialogue. To incorporate the acquired knowledge in a conversational format, the method generated a follow-up question and connected it to the knowledge. However, in this approach, the timing of the insertion of interesting knowledge into the dialogue is determined by rules, and the content response does not take into account the dialogue context, resulting in unnatural dialogue. Vicente et al. (2023) proposed a method of introducing interesting knowledge into a spoken dialogue system to help users perform complex tasks using templates. In this approach, interesting knowledge gathered from web searches is introduced into the dialogue via a template, producing a dialogue that lacks naturalness and coherence for the dialogue context. To address these challenges, I am working on selecting appropriate knowledge, taking into account both the dialogue context and interestingness, and generating responses based on knowledge without using templates. This will enable the generation of natural and interesting responses that are consistent with the dialogue and will improve user satisfaction.

1.2 Knowledge-grounded dialogue systems

Knowledge-grounded dialogue systems are approaches that generate responses that are based on external knowledge relevant to the dialogue, and can generate diverse and informative responses. Knowledge-grounded dialogue systems basically consist of two modules: knowledge selection and response generation (Wang et al., 2023). The knowledge selection module selects knowledge for use in the next response from the candidate knowledge related to the dialogue, and the response generation module generates a response that is based on the content of the retrieved knowledge and the dialogue context. Kim et al. (2020) built a model of knowledge selection with continuous latent variables modeling past knowledge selection. Zhao et al. (2020) proposed an unsupervised approach to jointly optimize knowledge selection and response generation using a prior learning model. However, most existing methods mainly perform knowledge selection by considering only the dialogue context, resulting in responses that contain much general information and are uninteresting (Xu et al., 2023). To address this, Xu et al. (2023) modeled a shift in dialogue topics and built a model for selecting a variety of knowledge while remaining consistent with the dialogue context. Generating responses that are consistent and interesting in the dialogue context is considered necessary to build a dialogue system that is close to human and engaging. To this end, I am working on a model that es-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 25–27

timates the appropriate interestingness of the knowledge for use in a response to select knowledge that is based on this interestingness and the context of the dialogue. As responses that contain general content are generally preferred to interesting content at the beginning of a dialogue and topic switches, it is important to capture topic switches in a dialogue to estimate the appropriate interestingness of the knowledge used in a response.

1.3 Trade-off between fidelity and consistency

A dialogue system that produces responses that are engaging and interesting to the user must provide responses faithful to knowledge and consistent with the context of the dialogue. However, there is a recognized trade-off between generating responses that are consistent with the dialogue context and faithful to knowledge (Chawla et al., 2024). Rashkin et al. (2021) proposed a method of improving fidelity to knowledge by adding control tokens to the beginning of the model input. The results showed that improving fidelity to knowledge may sacrifice consistency in the dialogue context. Chawla et al. (2024) built an approach to generating responses that balance fidelity and consistency by planning the content of the responses for generation and then creating a response generation model. It is more important to generate responses that balance fidelity to knowledge and consistency with the dialogue context than to focus solely on generating responses that reduce the generation of hallucinations and are faithful to knowledge.

2 Spoken dialogue system (SDS) research

Due to the advent of large language models, text dialogue systems can now generate natural and fluent responses that are close to those of humans. Thus, it is expected that SDSs can be studied more actively and used in a wide range of aspects of society, such as restaurant reservations, product recommendations, and counseling. In particular, multimodal dialogue systems have attracted particular attention recently because they enable close communication with humans by using user information such as facial expressions and voice information. However, several challenges have to be resolved before SDSs can be used in many fields of society.

First, there are insufficient datasets to train SDSs. Audio data collection is much more costly and time-consuming than text data, resulting in a smaller number of datasets and smaller-size datasets. In particular, there is a lack of non-English audio datasets. This lack of datasets directly prevents SDSs from being used in a wide variety of situations. It is necessary to find a way to train each module using independent text, audio, and video data and combine them to build a model.

Further, SDSs need to reduce response time relative to text dialogue. Long response times are unnatural and provide the user a sense of distrust. The larger the model that generates the response, the more fluent it is, but the larger the computational resources required, the longer it takes to generate the response. Also, multimodal dialogue systems need to process voice and image information rather than text information, which takes time to respond. To increase user engagement, we must find ways to reduce response time.

Furthermore, in the use of SDSs in society, it is essential to reduce hallucinations. Responses containing incorrect information when making restaurant reservations or recommending products are a serious problem. The occurrence of hallucinations is thought to be the main reason that SDSs are not still widely used in society today. Methods that do not produce hallucinations with high reliability and methods that detect responses including hallucinations will be particularly important in the future.

3 Suggested topics for discussion

I suggest discussing the following topics:

- What methods can reduce the time to generate knowledge-grounded responses in SDSs?
- How should multimodal information of users such as facial expressions be used in knowledge-grounded dialogues?
- Are existing methods for assessing response diversity adequate? How can response diversity be appropriately and automatically assessed?

References

Atheer Algherairy and Moataz Ahmed. 2024. A review of dialogue systems: current trends and future directions. *Neural Computing and Applications* 36(12):6325–6351.

Kushal Chawla, Hannah Rashkin, Gaurav Singh Tomar, and David Reitter. 2024. Investigating content planning for navigating trade-offs in knowledge-grounded dialogue. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 2316–2335.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.

Jakub Konrád, Jan Pichl, Petr Marek, Petr Lorenc, Van Duy Ta, Ondrej Kobza, Lenka Hýlová, and Jan Sedivý. 2021. Alquist 4.0: Towards social intelligence using generative models and dialogue personalization. *CoRR* abs/2109.07968.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 704–718.

Frederico Vicente, Rafael Ferreira, David Semedo, and Joao Magalhaes. 2023. The wizard of curiosities: Enriching dialogues with fun facts. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, pages 149–155.

Ming Wang, Bo Ning, and Bin Zhao. 2023. A review of knowledge-grounded dialogue systems. In *2023 8th International Conference on Image, Vision and Computing (ICIVC)*. pages 819–824.

Lin Xu, Qixian Zhou, Jinlan Fu, and See-Kiong Ng. 2023. Cet2: Modelling topic transitions for coherent and engaging knowledge-grounded conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31:3527–3536.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 3377–3390.

Biographical sketch

Hiroki Onozeki is a master's student at the Graduate School of Informatics and Engineering, The University of Electro-Communications. He is interested in knowledge-grounded dialogue systems. He has participated in several competitions building dialogue systems, including Dialogue Robot Competition 2023, Dialogue System Live Competition 6, and AIWolfDial2024jp.

Yuki Zenimoto

Nagoya University
Naogoya, Aichi
Japan
zenimoto.yuki.u1@s.mail.nagoya-u.ac.jp
https://yukizenimoto.notion.site/

1 Research interests

I believe that for future dialogue systems to coexist with humans, it is crucial to consider the value of an interlocutor, such as their way of thinking and perceiving things. Currently, dialogue systems like ChatGPT are used by many users. However, the information of the interlocutor is only expressed in simplified sentences such as "I like cooking" in many dialogue systems (Lu et al., 2022; Zhang et al., 2018; Tsunomori and Higashinaka, 2024); the dialogue systems cannot consider the interlocutor's values. Therefore, I am dedicated to researching **dialogue systems for eliciting the interlocutor's values** and **methods for understanding the interlocutor's values from narratives**.

1.1 Dialogue system for eliciting the values of an interlocutor

It is essential to understand the values of an interlocutor to generate responses based on these values. Manual interviews or questionnaires can be considered as a method for collecting the values of the interlocutor. However, manual interviews could pose a considerable burden on the interviewer and it is difficult to explore the responses of the interlocutor in depth using questionnaires. Therefore, methods to collect the values of the interlocutor automatically and naturally during a chat are desirable.

In light of the above, I am conducting research on a question-guiding dialogue system that asks specific questions naturally during a chat to elicit the values of the interlocutor. The question-guiding dialogue system was constructed using a large language model (LLM) and the question-guiding corpus (QGC) constructed by Horiuchi and Higashinaka (2021, 2023). The QGC is a dialogue corpus between humans, where one speaker is instructed to ask questions such as "How old are you?" and "Do you have any specialties?" in a natural context to the other interlocutor.

I constructed the question-guiding dialogue system by fine-tuning an LLM using the QGC and using OpenAI's GPT-4. The GPT-4 prompts include dialogue data with natural guiding, which have been manually selected from the QGC. I am currently conducting evaluations of these question-guiding dialogue systems.

Eventually, I will evaluate the question-guiding performance for questions related to values, such as "Which is more important to you, life, or work?" or "What do you think about the circumstances of the protagonist in this movie?" In actual dialogues, the interlocutor may not always provide a clear answer to a question. Therefore, I am considering developing a method to evaluate the validity of the answers of the interlocutor and to ask follow-up questions when the response is not as expected. Moreover, in some situations, the interlocutor may not be willing to answer questions directly. Therefore, I would also like to explore the use of indirect questions to elicit information from users in a more sociable manner.

1.2 Understanding the narratives of the interlocutor

The values of the interlocutor are often expressed through narratives, such as stories about their past experiences or impressions of something (Schank, 1990). However, extracting the values of the interlocutor from a narrative is difficult because narratives generally involve complex events.

As a representation of a narrative from which to extract the values of the interlocutor, story intention graph (SIG), a structured format of narratives, will be useful (Elson, 2012). SIG focuses on important elements in the narrative, such as characters, actions, and intentions/motivations; it represents the entire narrative as a graph structure. However, research on automatic SIG generation methods is rare.

Hence, I am conducting research on methods for automatically generating SIGs using LLMs. I am currently developing a SIG generation system using OpenAI's GPT-4 and creating manually annotated SIGs for evaluation. Furthermore, SIGs can differ depending on the annotator (Lukin et al., 2016). Therefore, I plan to consider methods for evaluating the quality of SIGs.

2 Spoken dialogue system (SDS) research

For SDSs to be used daily in the future, they will engage in dialogue with the same user multiple times. In such cases, SDSs need the ability to remember dialogue histories and make responses considering those dialogue histories.

Recent LLMs can deal with very long inputs; however, generating consistent responses to all input information

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 28–29

is challenging. Furthermore, within publicly available dialogue datasets, few include long dialogue histories (Xu et al., 2022; Yamashita et al., 2023).

Future SDS research should focus on how to collect and use dialogue histories for long-term dialogues, as in Xu et al. (2022) and Tomashenko et al. (2020). In addition, human evaluation of long-term dialogues is more time-consuming and costly than a single-dialogue evaluation. Therefore, I believe that research on high-quality, low-cost automated evaluation methods is important.

3 Suggested topics for discussion

I would like to discuss the following topics:

- Human and AI collaboration: What capabilities are needed in dialogue systems for humans and dialogue systems to collaboratively undertake difficult tasks, such as creative activities or discussions?
- Everyday use of dialogue systems: In everyday scenarios, what will be the most common situations in which dialogue systems are used?
- Reflecting individual preference in LLMs: Many current LLMs are trained on data deemed good for many people using methods like reinforcement learning from human feedback (Ouyang et al., 2022). Under such a circumstance, what methods are promising for reflecting individual preferences in LLMs?

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 23H00493.

References

David K. Elson. 2012. *Modeling Narrative Discourse*. Ph.D. thesis, Columbia University.

Sota Horiuchi and Ryuichiro Higashinaka. 2021. Learning to ask specific questions naturally in chat-oriented dialogue systems. In *Proc. of IWSDS*.

Sota Horiuchi and Ryuichiro Higashinaka. 2023. Learning to guide questions in chat-oriented dialogue by using combination of question-guiding corpora. In *Proc. of IWSDS*.

Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. 2022. Partner personas generation for dialogue response generation. In *Proc. of NAACL-HLT*. pages 5200–5212.

Stephanie Lukin, Kevin Bowden, Casey Barackman, and Marilyn Walker. 2016. PersonaBank: A corpus of personal narratives and their story intention graphs. In *Proc. of LREC*. pages 1026–1033.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* .

Roger C. Schank. 1990. *Tell me a story: A new look at real and artificial memory*. Charles Scribner's Sons.

Natalia Tomashenko, Christian Raymond, Antoine Caubrière, Renato De Mori, and Yannick Estève. 2020. Dialogue history integration into end-to-end signal-to-concept spoken language understanding systems. In *Proc. of ICASSP*. pages 8509–8513.

Yuiko Tsunomori and Ryuichiro Higashinaka. 2024. I remember you!: SUI corpus for remembering and utilizing users' information in chat-oriented dialogue systems. In *Proc. of LREC-COLING*. pages 9285–9295.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proc. of ACL*. pages 5180–5197.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In *Proc. of PACLIC*. pages 852–861.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proc. of ACL*. pages 2204–2213.

Biographical sketch

Yuki Zenimoto is PhD student at the graduate school of informatics, Nagoya University. He is supervised by Prof. Ryuichiro Higashinaka. He received his B.S. and M.E from University of Tsukuba in 2022 and 2024. His current research interests include dialogue systems for interviewing and narrative understanding.

Shun Katada

SANKEN, Osaka University, Japan

katada@sanken.osaka-u.ac.jp

1 Research interests

One of my research interests lies in **multimodal processing**. From a research perspective, multimodal is the science of heterogeneous and interconnected data (Liang et al., 2022). The multimodal processing methods mentioned here include linguistic, audio, visual, and biological information processing, which are important for understanding human behaviour. Computational representations and summarizing this information to reflect heterogeneity and interconnections are popular topics in this area. The author's current work focuses on creating multimodal spoken dialogue systems (SDSs) that can recognize human emotions/sentiments. The ultimate goal is to create an adaptive SDS that can change its behaviour by adapting a user's emotion based on multimodal processing.

Affective computing is another research interest that is not mutually exclusive. Affective computing relates to, arises from, or influences emotions (Picard, 2000). Although text modality has a dominant role in expressing emotion/sentiment during dialogue, nonverbal-based emotion recognition, such as facial expression and prosody, has been studied since the 1970s. Moreover, biosignals such as electroencephalograms (EEGs) and electrodermal activity (EDA) signals are often used in affective computing to detect emotional changes. Hence, the second topic focuses on these heterogeneous and interconnected data from an affective computing point of view, which is based on previous studies.

Biosignals have been used in previous works of the author; therefore, it is a candidate for the research topic but will be included in the above two topics.

1.1 Multimodal processing

First, the proposed method in previous work related to multimodal processing is shown (Katada et al., 2022). Language understanding has dramatically progressed through using large language models (LLMs), such as BERT and chatGPT, and has achieved excellent performance in emotion/sentiment estimation; however, using only linguistic information still has limitations. One of the issues is that sentiment is not necessarily expressed by users in human-agent interactions. To solve this issue, previous studies have proposed integrating token sequences derived from user utterances and time-series physiological (electrodermal) signals by multimodal processing. It was expected that integrating physiological signals into the language model can detect sentiment changes that are not expressed by user utterances. The Transformer architecture was applied to fuse text and physiological signals. As a result, our proposed methods significantly outperform the previous result, which is based on the simple early or late fusion method.

Second, a newly created multimodal dialogue corpus, called Hazumi2306, for developing an SDS with multimodal processing will be introduced, although it is not directly related to a new multimodal processing technique. The novelty of Hazumi2306 is that this dataset includes not only text, audiovisual, and physiological data but also frontal EEG data during human-agent interactions. The reason for collecting EEG data is that it has been the subject of focus in affective computing regions to capture unexpressed emotional changes in a controlled experimental environment. Approximately 500 minutes of chat dialogue were collected from thirty participants aged 20 to 70 years in total. The preliminary results of multimodal sentiment estimation based on conventional multimodal processing were also reported. It improved sentiment estimation performance when used with other modalities, although the simple EEG sensor used in this study has only three channels. This work has been published, and the corpus will be publicly available within the year (Katada et al., 2024). The analysis of this dataset by researchers will contribute to developing the SDS.

1.2 Affective computing

Multimodal analysis of human-agent interactions also sheds light on the emotional perception of humans. Basically, sentiment estimation based on multimodal processing considers only human observable signals such as linguistic, audio, and visual information. However, the contribution of the multimodal fusion of biosignals, which are unobservable by humans, has not been explored. In previous work (Katada et al., 2023), differences in the effect between observable (linguistic, audio, visual) and unobservable (physiological) signals were investigated in two different types of sentiment estimation, i.e., estimating sentiment labels annotated by the user and by a third party. Intuitively, a multimodal model based on the observable signal would be effective for estimating labels annotated by a third party since those labels are based on human observation (emotional perception). Addition-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 30–31
September 16–17, 2024.

ally, a multimodal model based on the unobservable signal would be effective for estimating labels annotated by the users since those labels would include unexpressed sentiment. These assumptions are evaluated empirically, and the obtained results generally agree with these assumptions (Katada et al., 2023). The results suggest that physiological features are effective and that the fusion of linguistic representations with physiological features provides the best results for estimating self-sentiment labels. In contrast, the fusion of linguistic, audio, and visual features is effective for estimating sentiment labels based on third party, which can be derived from the corresponding signals that are observable by humans.

2 SDS research

Text-based dialogue systems have rapidly evolved in the past 10 years with the advent of deep learning, Transformer, BERT, and other LLMs. The number of model parameters and parallel computations continue to increase, and these efforts have enabled dialogue systems to produce accurate responses.

One simple perspective is that, unlike text-based dialogue systems, the SDS uses auditory data. Automatic speech recognition (ASR) may include some research topics related to LLMs. In a nonstationary noisy environment, the ASR performance degrades, and user utterance words that include word errors may be sent to an SDS equipped with an LLM. In this case, a dialogue breakdown may occur if the LLM cannot address the word error. Thus, handling word errors in ASR with LLM may be a research topic.

3 Suggested topics for discussion

Related to the abovementioned research, the need of working with signals that can be less invasive is one of the suggested topics. There are non-invasive techniques that may be useful for emotion recognition such as microgesture recognition, thermal imaging, sensors in mobiles, etc. It may be worth discussing what techniques with a multimodal spoken dialogue system would be valuable and practical.

References

Shun Katada, Shogo Okada, and Kazunori Komatani. 2022. Transformer-based physiological feature learning for multimodal analysis of self-reported sentiment. *International Conference on Multimodal Interaction (ICMI)* pages 349–358.

Shun Katada, Shogo Okada, and Kazunori Komatani. 2023. Effects of physiological signals in different types of multimodal sentiment estimation. *IEEE Transactions on Affective Computing* 14(3):2443–2457.

Shun Katada, Ryu Takeda, and Kazunori Komatani. 2024. Collecting human-agent dialogue dataset with frontal brain signal toward capturing unexpressed sentiment. *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)* pages 3518–3528.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430* .

Rosalind W Picard. 2000. Affective computing. *MIT press* .

Biographical sketch

Shun Katada received a Ph.D. degree in life science from Tsukuba University, Japan, in 2014. He also received a Ph.D. degree in information science from JAIST, Japan, in 2022. He is currently a specially appointed assistant professor at SANKEN, Osaka University. He is a member of IPSJ and ACM.

Sadahiro Yoshikawa

Equmenopolis, Inc.
Japan Advanced Institute of Science and Technology
yoshikawa@equ.ai
s2230033@jaist.ac.jp

1 Research interests

My research interests lie in the area of **how users feel when using spoken dialogue systems (SDSs)**, including measuring **user satisfaction** and scoring **naturalness of voice conversation** such as speed, tone, response timing, and turn-taking events. In my ongoing master's thesis, I am working on response timing estimation.

As a research engineer in a company, I have been working on quality assurance for the Intelligent Language Learning Assistant (InteLLA) system, a virtual agent providing English proficiency assessments through oral conversations (Matsuyama et al. (2023)). The quality assurance for this system is towards consistently making users feel "Wow!". In the conversation with a virtual agent, the quality of the animation as well as the voice is important. I am trying to define the metrics for each critical point one-by-one based on the user satisfaction. There are also the viewpoint of **cost efficiency** when building **SDSs on a large-scale**. Building a framework that optimizes costs while maintaing user satisfaction is critical to long-term SDS operation.

1.1 Response Timing Prediction

Response timing has important role for SDS for not only the impression but also the intention of the utterance. For instance, the experience by Roberts and Francis (2013) showed the perceived willingness begins to drop after 600ms, and then clearly and significant steps down from 700 to 800 ms, and the corpus analysis by Kendrick and Torreira (2015) suggests the proportion of dispreferred actions is significantly greater than that of preferreds in case of the responses after approximately 700 ms and the gaps longer than the norm (>300 ms) decrease the likelihood of an unqualified acceptance.

Researchers built models to predict the actual response timing using LTSM (Roddy and Harte (2020)), with syntactic completeness prediction model (Sakuma et al. (2023)). Although most response timing estimation models are regression models, even if the error is the same at 200 ms, the influence of the error at 400 ms and 1500 ms is different. Furthermore, it is hard to confirm how much the error will affect human perceptions. I would like to deal with the difficulty of the response timing perception of humans utilizing deep neural network models.

Besides, several previous studies have indicated that the distribution of response timing varies depending on the conversation situation, such as the nature of conversations such as task-oriented or not (Levinson and Torreira (2015)) and the speaker's language (Stivers et al. (2009)). Therefore, when applying timing estimation models to SDS, we must also consider where the application will be located.

1.2 Future Turn-taking Prediction

The faster turn-taking event prediction with high accuracy, the more inference cost can be used for the quality for the actions of SDSs at the turn-take. If enough time can be used for inference, the inference cost may also be used not only for the actions but also for user adaptation or adjustment of response timing. Therefore, future turn-taking prediction is crucial for SDSs.

Ekstedt and Skantze (2022) proposed Voice Activity Projection (VAP) model forward to predicting future voice activity. The predictive task of VAP uses VAP window, which is discretized into a fixed number of bins as each bin indicates the probability whether the voice is active or not. When a VAP window is set to predict future voice activities, the performance of the task indicates the ability for future prediction. In the paper, the VAP model (referred to as the Discrete model in the paper) enumerates each possible configuration of a VAP window as separate states. In the model, a VAP window can be viewed as sequence of bits where the total number of states grows exponentially as 2^{n_bits}. For instance, the number of bins to 4 for each speaker was in 8 total bits in the paper, thus the output dimension of the model is 256, indicating 256 different possible states. This discrete method resulted in high performance on the task related to turn-taking events in near future (S-pred).

There are extended researches for the VAP model, such as the CPU inference (Inoue et al. (2024)) and multimodal VAP (Onishi et al. (2023)). However, there is some challenges at real-time inference of turn-taking events in practice. Raux and Eskenazi (2009) indicates the lower latency, the more user interruption is caused in an exper-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 32–34
September 16–17, 2024.

iment with the users of an automated call system for bus information. Moreover, SDSs need an algorithm to actually trigger a turn-taking cue using the predicted probability. Ruede et al. (2017) proposed an implementation using local maximum value within a window of an user utterance to trigger a backchannel. Although this is one of the solution for this issue, the window is only used for the model producing the local maximum value curve such as LSTM, not VAP, and a delay occurs because there is a margin between the maximum value in the window and the end of the window. Lala et al. (2019) showed an implementation using consecutive positive predictions as a turn-taking cue. However, this model aims at turn-taking that combines filler and eye-gaze, it thus needs to be verified whether it can be applied to VAP.

1.3 Allowable Threshold for Overlap

Skantze (2021) explains overlap has two types: cooperative and competitive overlap. There is so far very little work on how to produce cooperative overlapping speech, and there is a system regarding overlap in DeVault et al. (2009), in order to help the user to complete the sentence, possibly overlapping with the user's speech. However, the system often resulted in the agent being perceived as barging in and interrupting the user's speech. Unlike cooperative overlaps, competitive overlaps need some kind of resolution mechanism (to determine who should get the floor).

I wondered how much the competitive overlap of SDSs is bad. How much does overlap affect SDS user experience (UX)? Is it enough to add resolution mechanism even if the overlap occurs many times? Is there a way to get users to allow the competitive overlap? In a large-scale usage of SDS, unexpected competitive overlap is inevitable. Therefore, I would like to make the metrics how the effect of the overlap for SDS UX compared with other violations. Besides, I'm exploring a turn-taking strategy that get users to allow overlaps.

1.4 Allowable Threshold for Disturbed Video

InteLLA, which is our virtual agent communicates with the users through video streaming for easy usage to the person who has less computing resources, so we are managing the computing resources for the animation and the network resources. Moreover, InteLLA provides English proficiency assessments, thus the stable conversation is required. In a large-scale usage of SDS, if the excess optimization of computing resources or shortage of network resources is occurred, the agent turn-taking becomes unstable such as disturbed, choppy, delayed, etc. Therefore, optimizing the costs also requires managing computing resources and network latency to ensure video quality for the stable conversation.

However, it is unclear to what extent animation quality affects proficiency ratings and user satisfaction. Moreover, to the best of my knowledge, there is no solid indicators for the video quality of SDSs. If we measure the relationship between the resources and the quality precisely, the users can be use InteLLA at less network resources and InteLLA can work with less computing resources. Therefore, I'm exploring the metrics for precisely measure of the animation to ensure proficiency ratings and user satisfaction.

There is a reference for quality control indicators regarding the video itself. Min et al. (2024) indicates there is full-reference (FR) or no-reference (NR) analysis. For assessing corrupted video quality, my research will be started with FR.

2 Spoken dialogue system (SDS) research

I think the field of dialogue system will become closer to front-end or game engineering if SDSs are easy to custom and publish to internet like a cloud service by an individual in 5 to 10 years. Software for SDSs will be more complex, and be OS-dependent like browsers, and well-known software will be utilized without many people understanding the detailed mechanisms, and some SDSs will be designed by designers. Over the next 5 to 10 years, the number of software that can use for SDSs is expected to increase explosively. At that time, I think what researchers of SDSs should do is understanding the mechanisms of each modules of SDSs and defining the evaluation criteria for SDSs to guide engineers to build stable, secure, and reproducible SDSs. After the next 10 years, the core technologies for SDSs will be expanded by other modalities such as virtual reality and sensing technologies. Therefore, even if the SDSs we called today will be generalized, researchers will be required to extend another modalities for SDSs.

3 Suggested topics for discussion

- How to evaluate the user satisfaction regarding turn-taking events?
- How to implement future turn-taking prediction in practice?
- How to evaluate the quality of animation behaviors in conversations?

References

David DeVault, Kenji Sagae, and David Traum. 2009. Can i finish? learning when to respond to incremental interpretation results in interactive dialogue. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, USA, SIGDIAL '09, page 11–20.

Erik Ekstedt and Gabriel Skantze. 2022. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Proc. Interspeech 2022*. pages 5190–5194. https://doi.org/10.21437/Interspeech.2022-10955.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Real-time and continuous turn-taking prediction using voice activity projection. In *ArXiv*. https://arxiv.org/abs/2401.04868.

Kobin H. Kendrick and Francisco Torreira. 2015. The Timing and Construction of Preference: A Quantitative Study. *Discourse Processes* 52(4):255–289. https://doi.org/10.1080/0163853X.2014.955997.

Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *2019 International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, ICMI '19, page 226–234. https://doi.org/10.1145/3340555.3353727.

Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology* 6:731. https://doi.org/10.3389/fpsyg.2015.00731.

Yoichi Matsuyama, Mao Saeki, Hiroaki Takatsu, Ryuki Matsuura, Fuma Kurata, and Shungo Suzuki. 2023. Intella: Dialog-based english speaking assessment agent that elicits learnerapos;s language ability. *THE JOURNAL OF THE ACOUSTICAL SOCIETY OF JAPAN* 79(3):162–169. https://doi.org/10.20697/jasj.79.$3_1$62.

Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. 2024. Perceptual video quality assessment: A survey. https://arxiv.org/abs/2402.03413.

Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. 2023. Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation. In *Proceedings of the 11th International Conference on Human-Agent Interaction*. Association for Computing Machinery, New York, NY, USA, HAI '23, page 13–21. https://doi.org/10.1145/3623809.3623837.

Antoine Raux and Maxine Eskenazi. 2009. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09*. Association for Computational Linguistics, Boulder, Colorado, page 629. https://doi.org/10.3115/1620754.1620846.

Felicia Roberts and Alexander L. Francis. 2013. Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America* 133(6):EL471–EL477. https://doi.org/10.1121/1.4802900.

Matthew Roddy and Naomi Harte. 2020. Neural Generation of Dialogue Response Timings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 2442–2452. https://doi.org/10.18653/v1/2020.acl-main.221.

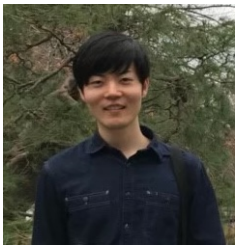
Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing Backchannel Prediction Using Word Embeddings. In *Proc. Interspeech 2017*. pages 879–883. https://doi.org/10.21437/Interspeech.2017-1606.

Jin Sakuma, Shinya Fujie, and Tetsunori Kobayashi. 2023. Response Timing Estimation for Spoken Dialog Systems Based on Syntactic Completeness Prediction. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. pages 369–374. https://doi.org/10.1109/SLT54892.2023.10023458.

Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67:101178. https://doi.org/10.1016/j.csl.2020.101178.

Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106(26):10587–10592. https://doi.org/10.1073/pnas.0903616106.

Biographical sketch



Sadahiro Yoshikawa is a Research Engineer at Equmenopolis. He is also a master's mature student at the Graduate School of Computer Science, Japan Advanced Institute of Science and Technology. Formerly, he was a freelancer as a Data Engineer. His interest is voice response quality of SDSs and defining metrics for SDSs towards quality assurance.

Atsumoto Ohashi

Nagoya University
Nagoya, Aichi
Japan
ohashi.atsumoto.c0
@s.mail.nagoya-u.ac.jp
https://ohashi56225.github.io/

1 Research interests

My research interests concern the field of task-oriented dialogue (TOD) systems, which aim to help users accomplish specific dialogue goals (for example, customer service and booking services) while responding to user requests. To realize a practical TOD system deployable in a wide variety of applications, I particularly focus on **optimizing the task completion ability** using reinforcement learning (RL) and **developing language resources and exploring multilinguality**.

1.1 Optimizing task completion ability

TOD systems process a single input utterance from the user through a pipeline of multiple modules, such as natural language understanding, dialogue state tracking (DST), dialogue policy, and natural language generation (NLG), before generating the final response.

TOD systems have evolved rapidly in recent years, and benchmark scores for measuring the accuracy of each module have improved significantly with the introduction of deep learning-based methods. However, it is known that when these modules are combined to form a pipeline for generating responses and introduced in actual multi-turn interactions, the task completion performance is generally unsatisfactory (Takanobu et al., 2020). One of the main reasons is that several models are trained solely on static data and lack robustness to domain shifts and irregularities that occur in real-world interactions. Even dialogue systems comprising recent large language model (LLM)-based modules lack sufficient task completion performance (Hudeček and Dusek, 2023).

To optimize the real-world task completion ability, RL-based approaches that learn from successes and failures in actual explorations are considered suitable. Several studies have used RL to fine-tune some modules in a dialogue system during simulated interactions and improve their task-completion ability. To achieve a more general optimization method, I proposed an approach that optimizes post-processing networks (PPNs) through RL to modify the output of each module post-hoc, rather than fine-tuning the modules (Ohashi and Higashinaka, 2022, 2023). The PPN-based approach does not require training each module, making it possible to optimize the task completion ability even for systems that include modules that are impossible or difficult to train, such as API-based, rule-based, and LLM-based. See (Ohashi and Higashinaka, 2022, 2023) for details on PPNs.

1.2 Language resources and multilinguality

Large-scale TOD datasets are essential for the research and development of deep learning-based TOD systems. For English, several large-scale TOD datasets, such as MultiWOZ (Budzianowski et al., 2018), have been constructed and have driven the development of English dialogue systems. Recently, some large-scale datasets have also been constructed in Chinese (Zhu et al., 2020). However, compared to English, TOD datasets in other languages are limited. Therefore, the capabilities of multilingual TOD systems are not on par with those in English. I believe that it is important to build corpora in other languages and use them to study fundamental technologies such as DST and NLG.

With this background, I focused on the lack of Japanese datasets and constructed JMultiWOZ, the first large-scale Japanese TOD dataset (Ohashi et al., 2024). The dialogue topics and databases used in JMultiWOZ are designed to be culturally natural for Japan, and the dialogues are collected using real-time interactions of Japanese speakers. We expect that this collection can avoid naturalness issues such as "translationese" and lack of cultural adaptation (Ding et al., 2022), which are common in corpora created by translating MultiWOZ.

In the future, I would like to first investigate how the system based on JMultiWOZ differs in performance compared to that built by translation-based corpora. Moreover, I aim to advance research on linguistically general dialogue models and cross-lingual transfer learning for TOD systems.

2 Spoken dialogue system (SDS) research

Dialogue research in 5 to 10 years In the next 5 to 10 years, I expect the development of dialogue systems capable of more human-like spoken interaction. Specifically, models with incremental processing for real-time input and response are anticipated. Rather than the current approach where the system waits for complete utter-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 35–36

ances from the user, more human-like turn-taking with overlapping speech and interruptions will be possible. Furthermore, the development of dialogue systems capable of human-like grounding based on non-linguistic modalities such as gestures, facial expressions, and environmental context is anticipated. With the acquisition of such human-like dialogue capabilities, dialogue systems will be more readily introduced into human society, and I would like to work on a dialogue system that can evolve similarly to humans by utilizing the rich experience in human society through RL.

Differences between academia and industry in SDS The industry currently invests vast resources to develop NLP systems using scaling approaches and deploy them in applications. Dialogue systems are a prime example of such applications, and this trend is expected to continue. However, it is challenging for academia to secure such vast resources. Therefore, in my opinion, academia should focus on more fundamental research, such as theoretical exploration and novel algorithmic study, rather than ready-to-use practical applications and up-scaling. Collaboration between industry and academia leveraging their respective strengths will lead to breakthroughs in SDS.

3 Suggested topics for discussion

I would like to discuss the following topics:

- In current TODs, labels such as dialogue state can be defined in detail, but how should we deal with a more complex dialogue in which it is difficult to define labels?
- The optimization of dialogue systems through RL typically relies on user simulators, which are often costly to implement. Is there an effective approach to utilize RL without a user simulator?
- Is a domain-specific dataset for each language necessary to build multilingual TOD systems? Or can LLMs generalize dialogue abilities to other languages?
- How can we efficiently collect speech and multimodal TOD data?

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 19H05692 and JST Moonshot R&D Grant number JPMJMS2011.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proc. EMNLP 2018*. pages 5016–5026.

Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. In *Proc. ACL 2022*. pages 1639–1657.

Vojtěch Hudeček and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proc. SIGDIAL 2023*. pages 216–228.

Atsumoto Ohashi and Ryuichiro Higashinaka. 2022. Post-processing Networks: Method for Optimizing Pipeline Task-oriented Dialogue Systems using Reinforcement Learning. In *Proc. SIGDIAL 2022*. pages 1–13.

Atsumoto Ohashi and Ryuichiro Higashinaka. 2023. Enhancing Task-oriented Dialogue Systems with Generative Post-processing Networks. In *Proc. EMNLP 2023*. pages 3815–3828.

Atsumoto Ohashi, Ryu Hirai, Shinya Iizuka, and Ryuichiro Higashinaka. 2024. JMultiWOZ: A large-scale Japanese multi-domain task-oriented dialogue dataset. In *Proc. LREC-COLILNG 2024*. pages 9554–9567.

Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. 2020. Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation. In *Proc. SIGDIAL 2020*. pages 297–310.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *TACL* pages 281–295.

Biographical sketch

Atsumoto Ohashi is a PhD student at the Graduate School of Informatics, Nagoya University. He is supervised by Prof. Ryuichiro Higashinaka. He is interested in task-oriented dialogue systems, reinforcement learning, and natural language processing using deep learning.

Sicong Huang

University of California, Santa Cruz
Silicon Valley Campus
Santa Clara, CA
United States

shuan213@ucsc.edu

1 Research interests

My primary research interests lie in **evaluating and improving the faithfulness of language model-based text generation systems**. Recent advances in large language models (LLMs) such as GPT-4 (OpenAI et al., 2024) and Llama (Touvron et al., 2023) have enabled the wide adoption of LLMs in various aspects of natural language processing (NLP). Despite their widespread use, LLMs still suffer from the problem of hallucination (Huang et al., 2023), limiting the practicality of deploying such systems in use cases where being factual and faithful is of critical importance. My research specifically aims to evaluate and improve the faithfulness, i.e. *the factual alignment between the generated text and a given context*, of text generation systems. By developing techniques to reliably **evaluate**, **label**, and **improve** generation faithfulness, we can enable wider adoption of dialog systems that need to converse with human users using accurate information.

1.1 Evaluating the Faithfulness of Dialog Summarization Systems

Evaluating generated text is often considered a task that is as difficult as generating text per se. Besides gold-standard human evaluation, long-standing automatic metrics such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) have been shown to poorly correlate with human judgements in evaluating faithfulness (Maynez et al., 2020). More recent LM-based metrics such as CTC (Deng et al., 2021) and BARTScore (Yuan et al., 2021) that are designed to target faithfulness exhibit higher correlation with human judgements but often do not account for the types of errors dialog summaries tend to make, resulting in lower performance when evaluating summaries in the dialog domain. To address this issue, I developed technique to improve upon the existing BARTScore metric, tailoring them specifically to account for the unique challenges of dialog summarization, such as colloquial speech, ellipses, and coreference errors.

In-domain Fine-tuning As highlighted in Huang et al. (2022), metrics that perform well on news summarization often fail to transfer effectively to dialog summarization. I investigated techniques to adapt BARTScore to the dialog domain by (1) fine-tuning on other dialog data along with automatically generated summaries (2) fine-tuning on the training set of the evaluation data directly. Results show both approaches can improve metric performance on dialog summaries with fine-tuning directly on the evaluation data being the most effective.

Learning from Synthetic Negative Samples To capture the common types of errors that come with dialog summaries, I investigated methods to generate negative sample summaries that reflect common error types in dialog summaries (Tang et al., 2022), such as entity swapping, mask and regenerate, and totally irrelevant hallucination. Then I applied unlikelihood training to the negative samples, which minimized the probability of generating negative tokens, thus assigning lower score to unfaithful summaries. Results show learning from negative samples can further improve BARTScore's correlation with human judgements, and using all error types yields the highest performance gain.

Through this research, I developed techniques to improve the performance of BARTScore (a high performing metric at the time) at evaluating dialog summaries, enabling more reliable assessment of dialog summarization systems.

1.2 Improving the Faithfulness of Abstractive Summarization Systems

In this research, I attempted at improving summarization faithfulness by investigating methods to properly leverage span-level hallucination information.

Span-level Hallucination Labeling To identify the spans of text that contain hallucinated information, I leveraged GPT-4 as an automatic labeler. I created a dataset of summaries with span-level hallucination annotations by prompting GPT-4 to label information in generated summaries that is inconsistent to the source document.

Comparison of Training Methods This research compared different training approaches that can leverage negative samples to reduce unfaithfulness, including *gra-*

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 37–39

dient ascent (Yao et al., 2024), *unlikelihood training* (Welleck et al., 2020), and *task vector negation* (Ilharco et al., 2023). The results indicate that unlikelihood training is particularly effective in reducing unfaithful information in LLM-generated summaries. The reduction of hallucinated content is also confirmed by human annotations on a subset of generated summaries.

Through this research, I found an effective method to improve summary faithfulness, i.e. span-level annotation and unlikelihood training, and the improvement is consistent across both news and dialog domain. These findings pave ways to reduce hallucinations in text generation more generally.

1.3 Fine-grained Annotation of Generated Text

As Section 1.2 has shown that span-level hallucination annotation can provide valuable information that can be leveraged to improve summary faithfulness, obtaining reliable span-level annotation becomes a critical step in improving faithfulness. Moreover, most text generation metrics only provide scalar value scores, revealing no information on the reasoning and the part of the text that resulted in such scores. Due to the uninterpretability of these metrics, they provide little guidance on how to improve the generated text. Thus, I am currently looking into developing a metric that is based on span-labeling, providing not only scores but also the reasons that resulted in the scores. I believe that interpretability is a crucial feature in the next generation of evaluation metrics.

2 Spoken dialogue system (SDS) research

The field of Spoken Dialog Systems (SDS) research is poised for significant advancements in the coming years, driven by the rapid progress in large language models and the increasing demand for more natural and reliable human-computer interactions. In my opinion, developing more context-aware and factually consistent systems along with reliable evaluation metrics should be important themes of SDS research for the next 5 to 10 years. Advancements in these areas will enable wider adoption of SDS in various scenarios:

- Healthcare: Assisting in patient triage, mental health support, and chronic disease management.
- Education: Providing personalized tutoring and language learning assistance.
- Customer Service: Handling complex queries and providing more empathetic interactions.

Our generation of young researchers has the potential to make significant contributions in several areas:

- Developing highly faithful and contextually appropriate response generation techniques.
- Creating more robust and interpretable evaluation metrics that can reliably assess system outputs of desired quality, such as coherence, engagingness, and faithfulness.
- Improving the handling of ambiguity and implicit information in conversations.

To achieve these goals, we may need to answer some key questions:

- How can we effectively combine the strengths of rule-based systems with the flexibility of neural approaches to achieve faithful responses?
- How can we make SDS and its evaluation metrics more interpretable?
- What are the best ways to incorporate real-world knowledge and common sense reasoning into SDS?

As we advance in this field, it will also be important to address challenges related to privacy, bias mitigation, and maintaining the balance between automation and human oversight. The goal should be to create SDS that not only understand and respond accurately but also enhance human capabilities and improve quality of life across diverse user groups and applications.

3 Suggested topics for discussion

- Are autoregressive LLMs limited by their token-by-token nature, thus unable to plan their outputs and produce fully faithful generations?
- What does it mean for a language processing system to "understand" language?
- Bisk et al. (2020); Bender and Koller (2020) have suggested that training on predicting the next word alone is unable to capture meaning which requires grounding. Are vision-language models (or LLMs trained on more modalities of data) enough to capture meaning? Or are symbolic representations required?

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, pages 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 7580–7605. https://doi.org/10.18653/v1/2021.emnlp-main.599.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Sicong Huang, Asli Celikyilmaz, and Haoran Li. 2022. Ed-faith: Evaluating dialogue summarization on faithfulness. https://arxiv.org/abs/2211.08464.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=6t0Kwf8-jrj.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, pages 74–81. https://aclanthology.org/W04-1013.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and Ilge Akkaya et al. 2024. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, pages 5657–5668. https://doi.org/10.18653/v1/2022.naacl-main.415.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, and Yasmine Babaei et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SJeYe0NtvH.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large language model unlearning.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*. https://openreview.net/forum?id=5Ya8PbvpZ9.

Biographical sketch

Sicong Huang is a first-year Ph.D. student at University of California, Santa Cruz advised by Prof. Ian Lane. He completed an AI Residency at Meta where he devised ways to improve summarization metrics' reliability. Prior to this, he did a Master of Science in Computational Linguistics at the University of Washington, during which he interned at a startup company, Seasalt AI, where he built and deployed a meeting summarization service in production. Before this, he completed a Bachelors in Electrical and Computer Engineering also at University of Washington. During his undergraduate study, for 6 months, he was an exchange student at Tokyo Institute of Technology where he started learning and researching about natural language processing under Prof. Manabu Okumura.

Taiga Mori

National Institute of Advanced Industrial Science and Technology, AI Research Center, Tokyo, Japan

Chiba University, Chiba, Japan

mori.taiga.0412@gmail.com

1 Research Interests

My research interest lies at the intersection of **cognitive science** and **dialog system research**; more specifically, I am interested in the cognitive process of listener response generation and aim to implement my model in a dialogue system to validate its effectiveness and build more human-like dialog systems.

1.1 Listener response studies

In everyday conversation, it is generally the principle that one person speaks at a time (Sacks et al., 1974). However, in reality, listeners do not just listen passively; they respond with short utterances such as "yeah," nods, or laughter. These listener responses are referred to as back-channels (Yngve, 1970), continuers (Schegloff, 1982), response tokens (Gardner, 2001), or reactive tokens (Clancy et al., 1996), and contribute to smooth turn-taking and the deepening of relationships.

It has been found that the frequency of listener responses is especially high in Japanese (Maynard, 1986; Clancy et al., 1996), indicating their important role especially in Japanese conversations. Additionally, Japanese listeners use a variety of responses. Den and Yoshida (2011) extended Gardner's (2001) response tokens to Japanese and categorized Japanese response tokens into six types: responsive interjections, expressive interjections, lexical reactive expressions, repetitions, assessments, and completions.

Dialogue systems that primarily focus on listening to the user's speech and providing appropriate responses are called attentive listening dialogue systems, and they have been the subject of active research (Bevacqua et al., 2012; Lara, 2017). There is also a significant amount of research on detecting the timing for producing listener responses (Ward and Tsukahara, 2000; Morency et al., 2010; Kawahara et al., 2015). However, current attentive listening dialogue systems still face challenges regarding the diversity and consistency of responses. We believe that the cognitive approach is an effective way to address these challenges.

1.2 Cognitive listener response generation model

According to Clark's (1996) grounding model, human communication consists of four hierarchical levels, which he calls *action ladders*. According to this model, at the lowest level of communication, Level 1, the speaker executes a behavior such as vocalization or movement, and the listener pays attention to it. At Level 2, the listener recognizes the signal, such as words or gestures, produced by the speaker. At Level 3, the listener understands what the speaker means. At Level 4, the listener considers the joint action proposed by the speaker. Allwood et al. (1992) also proposed four feedback functions similar to these: *contact*, *perception*, *understanding*, and *attitudinal reaction*.

Based on these theory, we hypothesize that the cognitive process of generating listener responses in everyday conversation also consists of four levels, with different types of responses used depending on the level.
Attention level: Responses at this lowest level indicate that the listener is listening to and paying attention to the speaker's speech, and are typically observed immediately after disfluencies such as fillers or pauses. This is almost synonymous with traditional back-channels. Responses at the attention level include responsive interjections (e.g., "yeah" or "uh-huh" in English).
Word level: This level of responses indicate the listener's understanding or recognition of a certain word produced by the speaker and are observed after devices that induce listener responses, such as rising intonation, lengthening, pauses, or eye contact. This includes responses to try-markers (Sacks and Schegloff, 1979). Responses at the word level include not only responsive interjections but also expressive interjections and repetitions (e.g., "Oh, Mr. Yamada").
Propositional information level: Responses at this level indicate the listener's understanding, empathy, or emotions to a propositional information and are used at a position where the propositional information is complete or predictable. While this partially overlaps with the continuer (Schegloff, 1982), it differs in that it can also be seen within the TCU (Turn Constructional Unit). Responses at this level include responsive interjections,

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 40–42
September 16–17, 2024.

expressive interjections, repetitions, lexical reactive expressions (e.g., "right" or "I see"), and assessments (e.g., "scary" or "interesting").

Activity level: Responses at this highest level also indicate the listener's understanding, empathy, emotions, etc. but are oriented towards activities rather than single propositional information. Since responses at this level are used at the endpoint of the activity, they overlap with sequence-closing devices. Responses at the activity level include responsive interjections, expressive interjections, repetitions, lexical reactive expressions, and assessments.

However, as with Clark's action ladder, these levels are hierarchical, with higher-level reactions encompassing lower-level ones. For example, a response at the conclusion of a storytelling not only serves as a response to the entire story but also retrospectively indicates that the listener has been attentive to the speaker's talk and has correctly understood the individual propositional information and words that make up the story.

Traditional studies on predicting listener responses have primarily focused only on attention level responses (Morency et al., 2010; Kawahara et al., 2015). The lowest attention-level responses can be generated using these traditional prediction methods based on the speaker's speech and body movements as features. However, generating higher-level responses will require matching with the system's knowledge base and some form of reasoning.

1.3 Listener response generation using knowledge graph and LLMs

Currently, we are working on implementing the aforementioned cognitive model as a system. In particular, we are focusing on developing an architecture that generates responses based on the listener's knowledge. Our proposed architecture consists of system knowledge in the form of a knowledge graph and three modules using LLMs.

Information extraction module: This module extracts information from the user's utterance and converts it into structured data using an LLM. The LLM extracts information from the user's utterances and converts it into triples consisting of subject, predicate, and object.

Knowledge comparison module: In this module, the user's knowledge is compared with the system's knowledge, and the system's knowledge state is determined. There are five types of system knowledge states: *complete match* when the system has the same triple of knowledge as the user, *partial match* when the system does not have the same knowledge but has related knowledge that aligns with it, *no knowledge* when the system lacks any related knowledge, *partial conflict* when the system has related knowledge that contradicts the user's knowledge, and *complete conflict* when the system holds contradictory knowledge. Whether related knowledge aligns with or contradicts the user's knowledge is determined by the LLM. For example, even if the system doesn't know the exact temperature, knowing that it is snowing would be considered having related knowledge about the temperature.

Response generation module: This module generates a response using the system's knowledge based on the determined knowledge state. If the knowledge state is a complete match/complete conflict, the module generates an *agreement*/*disagreement* response. If the knowledge state is a partial match/partial conflict, it converts the related knowledge into a natural language sentence using the LLM and generates a *noticing*/*surprise* response. If the knowledge state is a no knowledge, it generates an *acceptance* response.

2 Future of Spoken Dialog Research

Interaction is a topic that spans multiple fields and there is a wealth of knowledge available on it. However, collaboration between these fields has been yet sufficient. One reason is the technical challenge of implementing the higher-order cognitive processing models constructed by linguistics, sociology, psychology and cognitive science into actual systems. However, with the advent of LLMs and other technological innovations, this issue is gradually being resolved. For example, it has become easier for cognitive science researchers like myself to create simple dialogue systems to prove their hypotheses. In the future, further integration is desirable to allow for the effective utilization of each other's insights.

3 Suggestions for discussion

- Multimodality: How can speech be integrated with other modalities such as paralinguistic information, gestures, facial expressions, and eye gaze?
- Explainability: To what extent should the dialogue system be able to explain its own actions? How best to use LLMs?
- Collaboration with other fields: How can we contribute to other fields? What do we expect from other fields?

Acknowledgement

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, Tatsuya Kawahara. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 127-136.

Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Maat, Gary Mckeown, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Etienne De Sevin, Michel Valstar, Martin Wollmer, Marc Shroder, and Bjorn Schuller. 2012. Building Autonomous Sensitive Artificial Listeners. *IEEE Transactions on Affective Computing*, 3(2):165-183.

Emanuel A. Schegloff. 1982. Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In: Tannen, D. (Ed.), *Analyzing Discourse: Text and Talk* (Georgetown University Round Table on Language and Linguistics, 1981). Georgetown University Press, Washington, DC.

Harvey Sacks, Emanuel A. Schegloff. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696-735.

Harvey Sacks, Emanuel A. Schegloff. 1979. Two preferences in the organization of reference to persons in conversation and their interaction. In *Everyday Language: Studies in Ethnomethodology*. New York.

Herbert H. Clark. 1996. *Using language*. Cambridge university press.

Jens Allwood, Joakim Nivre, Elisabeth Ahlsen. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1-26.

Louis-Philippe Morency, Iwan de Kok, Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20:70-84.

Nigel G. Ward, Olac Fuentes, and Alejandro Vega. 2010. Dialog prediction for a general model of turn-taking. In *INTERSPEECH*, Citeseer, pages 2662-2665.

Patricia M. Clancy, Sandra A. Thompson, Ryoko Suzuki, and Hongyin Tao. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics*, 26:355-387.

Rod Gardner. 2001. *When Listeners Talk*. John Benjamins, Amsterdam.

Senko K. Maynard. 1986. On back-channel behavior in Japanese and English casual conversation. *Linguistics*, 24:1079-1108.

Tatsuya Kawahara, Miki Uesato, Koichiro Yoshino, Katsuya Takanashi. 2015. Toward adaptive generation of backchannels for attentive listening agents. In *International Workshop Series on Spoken Dialogue Systems Technology*, pages.1-10.

Victor H. Yngve. 1970. On getting a word in edgewise. In: Campbell, M.A. (Ed.), *Papers from the Sixth Regional Meeting of Chicago Linguistic Society*, Chicago Linguistic Society, Chicago, pages 567-577.

Yasuharu Den, Nao Yoshida, Katsuya Takanashi, Hanae Koiso. 2011. Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, IEEE, pages 168-173.

Biographical Sketch

Taiga Mori is a PhD student at Chiba University and research assistant at the Artificial Intelligence Research Center (AIRC), National Institute of National Institute of Advanced Industrial Science and Technology (AIST). He is generally interested in multimodal interaction and currently working on modeling multimodal listener response generation such as verbal response tokens and head nodding. He uses both quantitative methods such as statistical modeling and qualitative methods such as conversation analysis to build models, and then implement them in dialogue systems to verify the effectiveness and validity of the models.

Benjamin Matthias Ruppik

Heinrich Heine University Düsseldorf
Universitätsstr. 1
40225 Düsseldorf
NRW, Germany

benjamin.ruppik@hhu.de
https://www.ruppik.net/

1 Research interests

My research area is **topological deep learning**, where I apply techniques from **topological data analysis** to **word embedding spaces**. My goal is to use these mathematical methods to understand and improve dialogue systems.

1.1 Word embeddings

Word embeddings associate each word or token in a text corpus with a dense vector in an ambient space, such that words with similar meanings are close together. Figure 1 shows a section of the latent space produced by a pre-trained contextual language model. These language vectors form the backbone of many dialogue system components, especially for natural language understanding, dialogue state tracking and natural language generation. Natural language is thus modelled via **point clouds** in higher dimensional spaces, whose dimensions are usually in the hundreds, if not thousands. From the **shape** of word embeddings, a multitude of features can be extracted, to form the basis for various downstream tasks in dialogue system applications.

We expect representations of real-world datasets in higher-dimensional space to lie in proximity to lower-dimensional sub-manifolds. Typically, one suspects that data manifolds can be described locally by a handful of coordinates, modelled on a low-dimensional Euclidean space. In stark contrast to this **manifold hypothesis**, in previous work, our research group detected singularities in a *static* word embedding resulting from polysemous words (Jakubowski et al., 2020). At words which have multiple meanings, different pieces of the *static* word space appear to converge and coincide. Since modern language models utilize *contextual embeddings*, I have recently focused on applying similar topological tools to *contextual* latent spaces. This comes with new challenges, such as increased computational cost and difficulties in interpretability, but also offers very interesting new perspectives.

1.2 Topological Data Analysis

For us three-dimensional humans, high dimensional data spaces are hard to understand and visualize. **Topological**

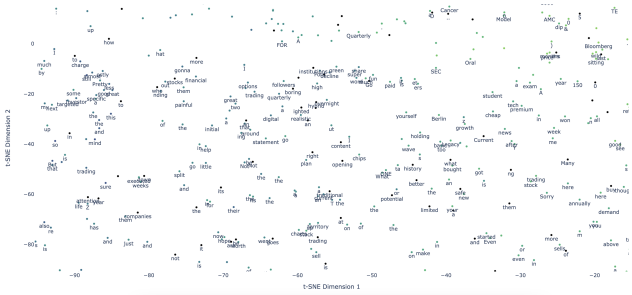


Figure 1: t-SNE projection of a subsection of the contextual embedding space produced by a `roberta-base` model. Topological data analysis is a collection of tools to study point clouds like this both locally and globally.

data analysis (TDA) is a mathematical toolkit which enables glimpses into high-dimensional point clouds by measuring their geometry at various scales. **Topology** is a branch of mathematics which studies the properties of geometric spaces that are invariant under continuous deformations. In data analysis, this involves studying the connected components, non-trivial holes and cavities of spaces derived from the data vectors. The major advantage of **topological features** is their invariance under small deformations and rotations, as opposed to using the embedding vectors directly. This leads to characteristics that are generalizable and not dependent on the exact data set used for training. To that end, I am interested in methods for defining and computing various flavours of **persistent homology** – topological features which can be detected at different size scales.

TDA has already shown its strengths in natural language processing, for instance, in probing model architectures. Kushnareva et al. (2021) build an increasing union of graphs from the attention scores of an input sentence in a pre-trained language model. The persistence features of the resulting objects are utilized for artificial text detection, a supervised classification task. Tulchinskii et al. (2023) topologically estimate the dimension of an embedding text paragraph to decide whether it was generated by an AI.

An important aspect of **topological machine learning**,

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 43–45
September 16–17, 2024.

the fusion of topological tools with deep learning, are vectorization methods. These transform the topological persistence diagrams into a format suitable for training downstream machine learning models. Lately, we made some progress towards including TDA feature extractors in the ConvLab-3 dialogue system code base. We have successfully applied topological features originating from *static* word embeddings to the **term extraction task** for dialogue datasets (Vukovic et al., 2022). Upcoming work extends this by showing that topological features from *contextual* word embedding spaces are even better suited for this task (Ruppik et al., 2024). This is demonstrated by showing that term extractors trained on the topological features of the MultiWOZ dialogue dataset can successfully transfer to another dialogue corpus, which contains different domains than the source datasets.

While these improvements themselves are promising, one should keep in mind that this method also leads to higher computational requirements. Our long-term research goal is to apply topological features for **unsupervised** applications, in particular the possibility of extracting meaningful **concepts** from unlabelled dialogue data to facilitate **ontology learning**.

2 Spoken dialogue system (SDS) research

Since November 2022, with the release of ChatGPT, I have been interacting with dialogue systems almost every day in one way or another: At work, they help me write emails, check grammar (including in this document), translate, answer coding questions, debug, and entertain me at home. These interactions have been mostly text-based, but as the voice mode improves and becomes more authentic, I can see myself using it more for speech in the very near future.

It is only a matter of time until such interactive dialogue agents are integrated into many aspects of our lives, such as customer support, appointment scheduling and booking, education, and all kinds of entertainment (games, AI companions, etc.). The most useful systems will be those which can be easily adapted to different domains. As a research community, we can contribute by investigating domain-agnostic architectures and representations.

Open academic research is more important than ever. We should not leave it to big tech companies to operate a walled garden of closed models, where we depend on them to get restricted access and always need to ask for permission to use state-of-the-art foundation models and build on top of them. This is especially important, since we should not need to trust secretive for-profit companies to be responsible for machine learning model use and deployment, and we should not leave all safety research to them.

One large problem that the academic community faces, including our lab at a public university in Germany, is restricted access to compute resources. This limitation prevents us from training foundation models and competing with large tech companies and their scale of operation. Nevertheless, I believe that our open basic research in academia is invaluable. Even with limited resources, we can come up with new and interesting ideas that a profit-oriented company might never have the motivation to explore. Academia, in particular, provides opportunities for cross-subject collaboration between different departments and people with vastly different backgrounds.

This interdisciplinary collaboration is especially relevant to me as I transitioned from pure mathematics to data science. I would appreciate seeing more results grounded in theoretical mathematics research find their way into practice. Effective collaboration between different subject areas depends on good communication. Mathematicians need to be able to identify possible applications of their methods, and consider implementations of algorithms and their efficiency. Dialogue systems researchers, on the other hand, need to point out places in their systems and pipelines where general methods might apply. They might need to step back from their specific implementations and identify parts that potentially work more generally than the specific setup or given dataset. The future state of dialogue systems research seems harder to predict than ever, but I believe that interdisciplinary collaborations will play a crucial role in it.

3 Suggested topics for discussion

1. Alternative evaluations of large language models: What other ways, apart from benchmarks – that might leak or suffer from overfitting – are there to evaluate language models? For example, can we intrinsically measure the quality of the embedding space produced by a model?

2. Prompting versus introspection: Should we focus on trying to apply the language generation capabilities of autoregressive language models to solve tasks, or should we take their representations and build on top of these? Can we learn everything about a model by observing how it answers the questions we present, or do we need to look into the internals of the models?

3. Reproducibility of research based on black-box language models, and their inaccessibility: How should we deal with the fact that we only have "prompting access" to the most powerful models, often hidden behind an API, and cannot observe and investigate their architecture in full?

References

Alexander Jakubowski, Milica Gašić, and Marcus Zibrowius. 2020. Topology of Word Embeddings: Singularities Reflect Polysemy. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, Barcelona, Spain (Online), pages 103–113. https://aclanthology.org/2020.starsem-1.11.

Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021. Artificial Text Detection via Examining the Topology of Attention Maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 635–649. https://doi.org/10.18653/v1/2021.emnlp-main.50.

Benjamin Matthias Ruppik, Michael Heck, Carel van Niekerk, Renato Vukovic, Hsien-Chin Lin, Shutong Feng, Marcus Zibrowius, and Milica Gašić. 2024. Local Topology Measures of Contextual Language Model Latent Spaces With Applications to Dialogue Term Extraction. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (to appear)*. https://www.arxiv.org/abs/2408.03706.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey I. Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2023. Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. https://dl.acm.org/doi/10.5555/3666122.3667828.

Renato Vukovic, Michael Heck, Benjamin Ruppik, Carel van Niekerk, Marcus Zibrowius, and Milica Gašić. 2022. Dialogue Term Extraction using Transfer Learning and Topological Data Analysis. In Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy Li, Arash Ashrafzadeh, Daniel Hernández Garcia, Malihe Alikhani, David Vandyke, and Ondřej Dušek, editors, *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Edinburgh, UK, pages 564–581. https://doi.org/10.18653/v1/2022.sigdial-1.53.

Biographical sketch

Ben is a postdoctoral researcher in the Dialog Systems and Machine Learning research group led by Prof. Milica Gašić at the Heinrich-Heine-Universität Düsseldorf, which he joined in 2022. In collaboration with the Topology and Geometry group in the Mathematics Department, under the supervision of Prof. Marcus Zibrowius, Ben is developing applications of Topological Data Analysis in Natural Language Processing, focusing on dialogue systems. Before transitioning to machine learning research, Ben was a pure mathematician at the Max-Planck-Institute for Mathematics in Bonn, where he specialized in knotted surfaces in 4-dimensional manifolds. He graduated from the University of Bonn in 2022. Ben is supported by funds from the European Research Council (ERC) provided under the Horizon 2020 research and innovation programme (Grant agreement No. STG2018 804636) as part of the DYMO project.

Photo by Shutong Feng, 2023.

Nicholas Thomas Walker

Otto-Friedrich University of Bamberg
Gutenbergstraße 13
Bamberg
Germany

nicholas.walker@uni-bamberg.de
www.uni-bamberg.de/ds/team/walker/

1 Research interests

In dialogue research, I am especially interested in using knowledge bases and models of dialogue entities to enhance dialogue system output. My research focuses particularly on how graph-structured knowledge underlying a dialogue system can be created, transformed, and extended to ground dialogue system output in world knowledge. In addition, I am interested in how logical rule-based reasoning can play a useful supplementary role in improving dialogue systems' understanding of dialogue context. Although contemporary large language models exhibit ever more impressive abilities, the problem of model hallucinations and groundedness in dialogue context remains an outstanding challenge in many applications (Zhang et al., 2023) that may be addressed in part by approaches that improve dialogue system access to background knowledge (Dinan et al., 2018).

In particular, I am interested graph-based **dialogue management** for **task-oriented dialogue systems**, particularly the use of **dynamic knowledge-bases**. I am especially interested in representations of knowledge combining information about the world with dialogue or user-specific information, e.g. personal knowledge graphs (Balog and Kenter, 2019), and how dialogue-local information can be effectively combined with other sources of knowledge such as logical rule-based reasoning. In this sense, I am interested in approaches which combine contemporary LLMs with other sources of reasoning, in the vein of neurosymbolic AI (Garcez and Lamb, 2023).

1.1 Previous and Current Work

In my previous work, I have investigated graph-based entity-centric models of dialogue management in the form of *conversational entity graphs* for SDS in **Human-Robot Interaction** (HRI). Such graphs are dynamic knowledge graphs centered on dialogue entities, which can be viewed generally as distinct units of information that may be useful to the system. Physical entities in the world that are important in situated dialogue can be conveniently represented within a graph alongside dialogue-specific, abstract or 'virtual" entities (Ultes et al., 2018) such as calendar events or even conversational intents can also be represented in the graph, with relations describing where an event will take place or which entities an intent refers to. As the dialogue proceeds, the graph of the dialogue state can be updated to represent changes in the entities and relations underlying the dialogue.

My previous work in graph-based dialogue management included a combination of probabilistic rule-based logic programming and neural models. Using ProbLog (De Raedt et al., 2007), the dialogue state can be enriched and extended by applying rules expressing common-sense inferences that can subsequently be verbalized for use by a language model (Walker et al., 2023). Thus, this combination of LLMs with logical rules is a hybrid approach allowing for desired common-sense conclusions and relevant information to be provided explicitly to the LLM through verbalization in the dialogue context.

In current research, I am working towards further extension of methods using the conversational entity graph to enrich the graph state and retrieve relevant information for the dialogue system. Where my previous work has investigated retrieval and verbalization methods combined with common-sense rules, I am now investigating ways to extend previous methods to extract relevant information from larger level graph structures beyond triples, such as meta-paths and node neighborhoods. As datasets often lack labels indicating knowledge graph elements that are relevant for a given dialogue turn, I am developing methods of knowledge extraction that do not rely on large quantities of data for supervised learning.

1.2 Future Work

In my future work, I plan to investigate what knowledge representations are most effective in enabling dialogue systems to converse naturally with users. Building upon my previous work, I intend to investigate new models of dialogue management incorporating abstract knowledge about the dialogue such as user and system goals and their interaction with both turn-level information and the dialogue as a whole. The relation of dialogue entities to overarching goals and objectives of the system has long played a role in classic slot-filling methods of dialogue state-tracking, yet these structures provide only a partial view of the interplay between these elements of dialogue

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 46–47
September 16–17, 2024.

(Cohen, 2019). I am therefore interested in investigating methods which provide system decision making and output with stronger grounding in dialogue knowledge to provide more consistent and reliable responses. I believe contemporary LLMs provide an opportunity to combine the strong generative capabilities of these models with a more robust understanding of both user intentions, system purpose and goals, and situational knowledge that can enable more naturalistic and robust dialogue systems.

2 Spoken dialogue system (SDS) research

Within SDS research in general, I believe the latest LLMs that accommodate multimodal input will be a central area of investigation. The degree to which such models are able to reason over diverse sources of multimodal information appears likely to be an important topic (Wang et al., 2024), and it may be of use to integrate external modules to refine their input and output in order to assist in areas in which these models fall short. The use of external tools and integration of different modules with LLMs seems likely to be an important area of research in the future due to the difficulties in fine-tuning and retraining such models. It will also likely continue to be important for research to focus on language model explainability and deeper understanding of the transformer architecture.

3 Suggested topics for discussion

- Neurosymbolic AI: How can the latest LLMs (including multimodal models) interact with logic-based decision making? What role can classical rule-based or logical decision making play in the current era of extremely large language models?

- Intentionality: What ways might we impart more intentionality or illocutionary intent to dialogue systems? Would a system that has some facsimile of such intentionality be likely to perform better at tasks or behave more naturally?

- Adapting System Behavior: What approaches are most effective for adapting system behavior to a task or situation when fine-tuning is not an option? Is prompt optimization sufficient to desired induce system behavior in all contexts? If not, which areas remain a challenge?

References

Krisztian Balog and Tom Kenter. 2019. Personal knowledge graphs: A research agenda. In *Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR)*.

Philip R Cohen. 2019. Foundations of collaborative task-oriented dialogue: what's in a slot? In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. pages 198–209.

Luc De Raedt, Angelika Kimmig, and Hannu Toivonen. 2007. Problog: A probabilistic prolog and its application in link discovery. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence IJCAI-07*. Hyderabad, volume 7, pages 2462–2467.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Artur d'Avila Garcez and Luis C Lamb. 2023. Neurosymbolic ai: The 3 rd wave. *Artificial Intelligence Review* pages 1–20.

Stefan Ultes, Paweł Budzianowski, Iñigo Casanueva, Lina M Rojas Barahona, Bo-Hsiang Tseng, Yen-Chen Wu, Steve Young, and Milica Gasic. 2018. Addressing objects and their relations: The conversational entity dialogue model. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. pages 273–283.

Nicholas Thomas Walker, Stefan Ultes, and Pierre Lison. 2023. A retrieval-augmented neural response generation using logical reasoning and relevance scoring. In *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, Full Papers*.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805* .

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv:2309.01219* https://arxiv.org/pdf/2309.01219.pdf.

Biographical sketch

Nick Walker is a researcher at Otto-Friedrich University of Bamberg. He completed his bachelor's degree in linguistics and master's degree in human language technology at the University in Arizona, and will be defending his PhD at the University of Oslo in October of 2024.

Xulin Zhou

Nagoya University
Nagoya, Aichi
Japan

zhou.xulin.j3
@s.mail.nagoya-u.ac.jp

1 Research interests

I am interested in dialogue systems where the user and the system collaboratively work on tasks through conversation. Currently, dialogue systems primarily aim to respond to human requests, functioning mainly as assistants or tools. However, if a dialogue system that engages in creative collaboration on an equal footing with humans can be developed, it would lead to more dynamic cooperation between humans and machines, fostering new types of interactions. To this end, I aim to create a dialogue system that focuses on cooperative collaboration and the expansion and development of ideas through dialogue.

1.1 Collection and Analysis of Dialogue Data for Collaborative Dialogue Systems

I analyze dialogues where two parties collaborate through conversation. Specifically, I focus on collaborative work that produces outcomes without a correct answer.

Several tasks have been used in studying dialogue systems that engage in collaborative work through conversation (Mitsuda et al., 2022). In the map task, one party holds a complete map while the other holds a partial map, and they complete the map through dialogue (Meena et al., 2014). In the OneCommon task, each speaker views one of two images cut from the same original image and identifies the common element by discussing the arrangement of dots of different colors and sizes (Udagawa and Aizawa, 2021). These tasks involve minimal creative elements and focus on finding a predetermined answer through collaboration. However, several real-world tasks do not have a single correct answer. I believe that building dialogue systems capable of collaboratively tackling such open-ended tasks will enhance their applicability. Therefore, I focus on tasks where there is no predetermined answer, aiming to develop dialogue systems for collaborative work.

Our group focuses on the collaborative creation of taglines (Zhou et al., 2024). Specifically, we are working on what we call the tagline co-writing task, where participants discuss and collaboratively edit taglines for given products.

I have created a tagline co-writing dialogue corpus that aims to gain insights useful for building collaborative dialogue systems and provide data for fine-tuning large language models. The corpus includes dialogues of humans performing the tagline co-writing task, the state of collaborative work during the conversations, and self-evaluations of the participants regarding the created taglines, their work, and their feelings through questionnaires.

I have analyzed the corpus by clustering utterances, extracting frequently occurring phrases in tasks with high self-evaluations, examining the workflow of utterances and edits over the entire task duration, and analyzing the interplay between utterances and taglines. The results indicate that tasks where utterances and tagline edits are conducted in parallel throughout the task tend to receive higher self-evaluations from the participants. Additionally, expressions of gratitude, positive evaluations, conveying understanding, and seeking agreement were found to be important in the tagline co-writing task.

1.2 Building a System for Collaborative Operations through Dialogue

As a prototype for a co-writing dialogue system, we developed a dialogue system using large language models (Zhou et al., 2023). This system combines a next-utterance generation model and a tagline generation model, both trained on data from the tagline co-writing dialogue corpus with a tagline evaluation model trained on data where third parties evaluated the taglines. We performed a dialogue experiment using the system and found that the system exhibited behaviors that are not normally observed between humans, such as overwriting changes the other party has made to the taglines without stating that to the other party. Additionally, issues such as the lack of coherence between utterances and edits, insufficient diversity in taglines, and hallucinations were observed. As future work, we intend to integrate highly advanced generative models such as GPT-4 to construct dialogue systems and to conduct evaluations on the components necessary for a dialogue system designed for the tagline co-writing task, considering that current issues may be resolved by utilizing models with higher performance. Additionally, we aim to consider new metrics for evaluating the tagline co-writing dialogue systems, such as the similarity to human-human interactions in terms of

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 48–49

conversational dynamics.

The timing of utterances and edits is also considered. In the tagline co-writing dialogue corpus, we asked each participant to make utterances and edit taglines at arbitrary timings to promote natural interactions. In contrast, the prototype system we developed employs a turn-based system where turns for utterances and tagline edits alternate between the human and the system, which creates a gap between the system's behavior and actual human behavior. By aligning the system's behavior more closely with the natural, non-turn-based behavior of humans, we aim to build a dialogue system that can act in a more timely manner.

2 Spoken dialogue system (SDS) research

I believe that in the near future, the domain that will lead to the broader and more widespread use of dialogue systems is customer service interactions. Research on dialogue systems for customer service covers various types, including product recommendations and customer support (Gao et al., 2021; Jia et al., 2022).

One of the current shortcomings of customer service dialogue systems is their lack of understanding of the surrounding environment. Enabling these systems to utilize the information that humans can naturally observe will lead to dialogues grounded in the physical world. Since vision plays a significant role in human cognition, utilizing multimodal information such as video is crucial for the advancement of dialogue systems. For instance, when considering a customer service dialogue system providing tour guidance, it could potentially offer services based not only on general information such as written guide books, but also on changing surroundings.

3 Suggested topics for discussion

I would like to discuss the following topics:

- What are the efficient methods for evaluating constructive discussions?
- What are the important elements required for a dialogue system to collaborate with humans effectively?
- How should we evaluate aspects of dialogue systems where the inter-annotator agreement among human evaluators is not high?

4 Acknowledgments

This work was supported by JSPS KAKENHI Grant number 19H05692 and JST Moonshot R&D Grant number JPMJMS2011.

References

Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open* 2:100–126.

Meihuizi Jia, Ruixue Liu, Peiying Wang, Yang Song, Zexi Xi, Haobin Li, Xin Shen, Meng Chen, Jinhui Pang, and Xiaodong He. 2022. E-ConvRec: A large-scale conversational recommendation dataset for E-commerce customer service. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. pages 5787–5796.

Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2014. Data-driven models for timing feedback responses in a map task dialogue system. *Computer Speech & Language* 28(4):903–922.

Koh Mitsuda, Ryuichiro Higashinaka, Yuhei Oga, and Sen Yoshida. 2022. Dialogue collection for recording the process of building common ground in a collaborative task. In *Proceedings of the 13th Conference on Language Resources and Evaluation*. pages 5749–5758.

Takuma Udagawa and Akiko Aizawa. 2021. Maintaining common ground in dynamic environments. *Transactions of the Association for Computational Linguistics* 9:995–1011.

Xulin Zhou, Takuma Ichikawa, and Ryuichiro Higashinaka. 2023. A prototype dialogue system for co-writing taglines with users. In *Proceedings of the Human-Agent Interaction Symposium 2023*. (in Japanese).

Xulin Zhou, Takuma Ichikawa, and Ryuichiro Higashinaka. 2024. Collecting and analyzing dialogues in a tagline co-writing task. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. pages 3507–3517.

Biographical sketch

Xulin Zhou is a Ph.D. student at the Graduate School of Informatics, Nagoya University. She is supervised by Prof. Ryuichiro Higashinaka. She is interested in dialogue systems that can collaborate with humans.

Yoshiki Tanaka

The University of Electro-Communications
Chofu, Tokyo
Japan

y-tanaka@uec.ac.jp
https://yoshi-chanaka.github.io/about/

1 Research interests

My research interests broadly lie in the influence of artificial intelligence (AI) agents on human decision-making. Specifically, I aim to develop **applications for conversational agents in decision-making support**. As part of this focus, during my master's program, I focused on supporting review writing and developed a system that assists in writing user reviews using an interview dialogue system. In this approach, the conversational agent first gathers product information such as users' impressions and opinions during the interview, to create reviews, facilitating the review writing process. Additionally, I conducted a **comprehensive evaluation** from the perspectives of system users and review readers. Although experimental results have shown that the system is capable of generating helpful reviews, the quality of the reviews still depends on how effectively the agent elicits the information from users. Therefore, I believe that personalizing **the agent's interview strategy** to users' preferences regarding the review writing process can further enhance both the user experience and the helpfulness of the review.

1.1 AI-assisted decision-making

In our daily lives, other people's persuasion or suggestions can affect how we make decisions. While these words can encourage beneficial decision-making, they can also lead to poor choices. For example, a friend's recommendation of a movie may prompt you to watch it, but your experience of the movie might be either satisfying or disappointing.

Decision-making support in AI applications aims to assist humans in making optimal choices by providing information and suggestions. The tasks it targets are wide-ranging, including recommendations (He et al., 2017), reservations (Budzianowski et al., 2018), e-mail writing (Fu et al., 2023), review writing (Bhat et al., 2023), and creative support for screenplays (Mirowski et al., 2023).

1.2 Evaluation method for AI agents

Decision-making tasks often have no absolute correct answers. Therefore, offline evaluations using datasets with ground-truth labels have limitations, making actual user studies crucial. In recent years, user studies have been conducted actively via crowdsourcing platforms such as Amazon Mechanical Turk,[1] where participants typically use the system and provide feedback in exchange for compensation. Post-surveys such as questionnaires and interviews in the user study, allow us to measure user satisfaction.

When designing user study experiments, researchers must appropriately determine the demographics of the subjects and the amount of feedback to collect. For example, my experiment targeted 100 Turkers per condition. Similarly, Mirowski et al. (2023) focused on 15 theatre and film industry professionals who have worked in TV, while Fu et al. (2023) targeted 40 participants per condition. User studies can provide direct feedback on participant satisfaction, but it remains challenging to collect as much data as in offline evaluations. A small sample size can lead to reproducibility issues. I argue that a careful design is necessary to accurately measure effectiveness when the amount of collected data is limited. In addition, it is important to properly analyze the collected data and clearly indicate the scope to which the resulting conclusions can be applied.

1.3 Constructing an interview dialogue system

In interview dialogue, the interviewer aims to elicit information from the interviewee. Existing research has shown that conducting surveys using a chatbot platform can yield higher-quality responses than using a Web survey platform (Kim et al., 2019). In cases where information collection is crucial, such as in my research, interview dialogue systems can be a promising option.

An interviewer skillfully eliciting information from the interviewee is desirable, and researchers have used human evaluation to measure this ability (Zeng et al., 2018; Okahisa et al., 2022). One effective way to develop an interview dialogue system with such capabilities is to incorporate dialogue strategies. For example, follow-up questions are a skill possessed by competent interviewers. This skill is effective in eliciting more detailed information when the interviewee's responses are ambiguous or too concise. Additionally, changing the topic during the interview dialogue allows for the collection of a wide range of information.

[1] https://www.mturk.com

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 50–52
September 16–17, 2024.

Another way to conduct effective interviews is by adapting the interview dialogue system to the interviewee. This personalization allows the system to generate more relevant questions by considering the user's background, preferences, and behaviors. The interview dialogue system I developed elicits feedback and opinions from the interviewee about their experiences using a product. In this case, for example, personalizing the system to the interviewee enables it to elicit evaluations compared to products the interviewee has used in the past. An interesting question is how to extract user information and whether such a dialogue strategy can be generalized to other tasks and situations.

2 Spoken dialogue system (SDS) research

I expect that in the future, Spoken Dialogue Systems (SDSs) will have a more significant impact on human decision-making. Although text-based dialogue systems can control vocabulary, SDSs also can control intonation and tone. This controllability expands the range of expression in system responses, enhancing both approachability and persuasiveness. I believe this will also help improve task success rates in task-oriented dialogue (ToD). Achieving these advancements requires the development of speech synthesis technology that can incorporate intonation and tone, as well as technology that can accurately understand users' emotions and intentions.

3 Suggested topics for discussion

Here are some topics to discuss:

- What new possibilities can be enabled by extending current ToD systems to be multimodal?
- How can we ensure data privacy when personalizing SDSs?
- What methods are available for personalizing non-textual information in multimodal dialogue agents?

References

Advait Bhat, Saaket Agashe, Parth Oberoi, Niharika Mohile, Ravi Jangir, and Anirudha Joshi. 2023. Interacting with next-phrase suggestions: How suggestion systems aid and influence the cognitive processes of writing. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, USA, IUI '23, page 436–452.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pages 5016–5026.

Liye Fu, Benjamin Newman, Maurice Jakesch, and Sarah Kreps. 2023. Comparing sentence-level suggestions to message-level suggestions in ai-mediated communication. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, CHI '23.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, WWW '17, page 173–182.

Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, CHI '19, page 1–12.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, CHI '23.

Taro Okahisa, Ribeka Tanaka, Takashi Kodama, Yin Jou Huang, and Sadao Kurohashi. 2022. Constructing a culinary interview dialogue corpus with video conferencing tool. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 3131–3139.

Jie Zeng, Yukiko I. Nakano, Takeshi Morita, Ichiro Kobayashi, and Takahira Yamaguchi. 2018. Eliciting user food preferences in terms of taste and texture in spoken dialogue systems. In *Proceedings of the 3rd International Workshop on Multisensory Approaches to Human-Food Interaction*. Association for Computing Machinery, New York, NY, USA, MHFI'18.

Biographical sketch

Yoshiki Tanaka is a first-year PhD student at the Graduate School of Informatics, The University of Electro-Communications. He is interested in supporting human decision-making using AI. He has participated in the AIWolf-Dial2024jp competition. He is supervised by Assoc. Prof. Michimasa Inaba.

Renato Vukovic

Heinrich Heine University Düsseldorf
Universitätsstr.1
40225 Düsseldorf
NRW, Germany

renato.vukovic@hhu.de
https://renatovukovic.github.io

1 Research interests

My research interests lie generally in **dialogue ontology construction**, that uses techniques from *information extraction* to extract relevant terms from task-oriented dialogue data and order them by finding hierarchical relations between them.

1.1 Dialogue Ontology Construction

Building ontologies for dialogue systems manually is expensive and time consuming (Budzianowski et al., 2018). The ontology is mainly needed by the dialogue state tracking module of task-oriented dialogue (TOD) systems (Heck et al., 2020; van Niekerk et al., 2021), but also important for user simulation Lin et al. (2023). The ontology covers structured information about domains (e.g. hotel), slots (e.g. price range) and values (e.g. expensive), while other conversational aspects such as user emotion can also be part of the dialogue state (Feng et al., 2023). However, generally TOD systems are limited to the knowledge in the ontology, limiting their application to new domains (Feng et al., 2024). Automating parts of the ontology annotation process can thus increase the scalability of TOD systems by making them easier to apply to new domains and unseen data. This is especially interesting when aiming to continually update a TOD system with new knowledge automatically in continual learning setups (Geishauser et al., 2024). Another possibility might be to improve a automatically constructed ontology using label validation approaches from active learning (van Niekerk et al., 2023), since the constructed ontologies come with a significant amount of noise.

Information extraction (IE) aims at structuring information from text data and there are normally two main steps, named entity recognition (NER) and relation extraction (RE) (Genest et al., 2022). Automatic ontology construction can be considered a form of IE for task-oriented dialogue. Ontology construction is about automating the process of building ontologies, rendering manual annotation unnecessary while saving time and making larger portions of unstructured data usable, so the system is able to include new domains, slots and values to talk about dynamically. The process can be separated into three steps, although you can split them up further into more fine-grained steps as well (Toledo-Alvarado et al., 2012; Cimiano et al., 2006):

1. **Term extraction:** extracting relevant domain, slot and value terminology in the textual data and finding concepts
2. **Relation extraction:** predicting hierarchical relations between the concepts, organising them into domains, slots and values
3. **Disambiguation:** ordering the found concepts based on their context such that words with similar meaning or domain end up in the same group, e.g. “expensive” and “high-end”

There are approaches focussing on different steps of the construction process, such as term extractors relying on frequency (Sclano and Velardi, 2007) or induce slots in a data-driven fashion (Qiu et al., 2022). Others extract relevant slots and inducing an ontology hierarchy (Hudeček et al., 2021), which then can be directly used to train a model on a downstream-task, like dialogue state tracking and response generation (Yu et al., 2022) based on the induced slot-schemas.
Apart from that there are also approaches that aim at making state tracking models more versatile and able to handle unseen data (Heck et al., 2022). Furthermore the advent of large language models (LLMs) enables even better generalisation to unseen tasks and domains, such as dialogue state tracking in a zero-shot fashion (Heck et al., 2023).

1.2 Term Extraction

The first step of ontology construction, term extraction aims at capturing all regions of interest in textual data, maximising recall (Nakagawa and Mori, 2002; Wermter and Hahn, 2006). The problem in this first step is that the precision is quite low, which makes additional processing or filtering of non-relevant terms necessary before proceeding with the next step (Frantzi and Ananiadou, 1999). In my research so far I mainly focus on improving the extraction process in terms of precision while keeping the recall at a high level so that less filtering is necessary.

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 53–56
September 16–17, 2024.

Furthermore my goal is to develop a term extractor with better generalisation capability to use it on different kinds of datasets, which cover a lot of different domains.

For this goal my group investigates potentially domain-agnostic features of the word embedding space that capture the meaning and the relevance of potential terms. This term extractor model should get terms in a way such that the follow-up steps are as easily feasible as possible to be able to improve the whole ontology construction process in the long-term. Investigated features include features obtained from pre-trained masked language models and ones obtained by applying mathematical tools like topological data analysis on the word embedding space to find meaningful structures. The models trained on these features show good zero-shot transferability to the much larger schema-guided dialogue (SGD) dataset (Rastogi et al., 2020) on the term extraction task (Vukovic et al., 2022) when trained on MultiWOZ (Budzianowski et al., 2018) as seed dataset. The performance of the term extractor can be further improved by computing them on a contextual level rather than on a global static level (Ruppik et al., 2024).

1.3 Relation Extraction and Disambiguation

In relation extraction, we consider three hierarchical relations between domains and slots, domains and values, and slots and values respectively that have to be predicted between terms. Furthermore, we consider an equivalence relation between terms of the same category in order to disambiguate semantically equivalent terms, such as "expensive" and "high-end" as price values. In our initial set-up we predict all the given relations jointly with one model, although experimental results might suggest that more emphasis on disambiguation might be needed. Note that another possibility is to infer the ontology hierarchy via clustering (Yu et al., 2022), which is not in line with most information extraction approaches.

By utilising such general structural relations, our goal is to utilise existing annotated datasets in order to extract semantic information in the form of an ontology on unseen data. In our research on ontology relation extraction, we experiment with updated decoding mechanisms for language models, such as constrained generation and chain-of-thought (CoT) decoding (Wang and Zhou, 2024) in order to improve generalisability of few-shot prompted and fine-tuned language models. In a transfer learning set-up we show that constrained chain-of-thought decoding improves performance of a language model trained on MultiWOZ as seed dataset and SGD as target dataset (Vukovic et al., 2024).

2 Spoken dialogue system (SDS) research

Understanding and acting upon natural language is one of the earliest challenges for artificial intelligence (AI), as it is part of the Turing test (Turing, 1950). Spoken dialogue system research evolved from the goal of solving the Turing test to solving more specific problems related to language, which in my opinion is one of the most important means for human communication and learning. I think SDS will become more and more incorporated in everyday life as you can already see in personal assistants, such as Siri or Amazon's Alexa. As long as they add value to the life of their user by making it more comfortable or make a personal secretary affordable to broader parts of the society, as they are much less expensive than paying real humans for the more and more tasks the systems are capable of.

Language is one of the main means human learning after mastering their mother tongue, since even movements to learn are normally accompanied by descriptions in language. This observation makes me assume that AI capable of understanding and acting upon language in a human-like manner might learn from the same sources humans learn from (Lynn and Bassett, 2020). This accomplishment would make large amounts of textual data usable for training large models with general knowledge and abilities. Altogether this assumption shows the great potential which lies in SDS research.

In my opinion, it is really hard to forecast what will happen in 10 years time, as there can be large jumps of progress if certain milestones are reached.

3 Suggested topics for discussion

- What degree of supervision is needed to extract semantic information from text and build an ontology?
- How can we leverage existing annotated data to structure information on unseen data?
- How to update the ontology of a model dynamically while interacting with users?
- What can you take from human learning and interaction from and with speech to adapt in spoken dialogue systems?
- Which architectures and training approaches are best suited for ontology relation extraction?

Biographical sketch

Renato Vukovic is currently a second year PhD student working on dialogue ontology construction for task-oriented dialogue under the supervision of Prof. Milica Gašić at Heinrich Heine University Düsseldorf. He holds a bachelor's and master's degree in computer science from the HHU Düsseldorf.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Brussels, Belgium, pages 5016–5026. https://doi.org/10.18653/v1/D18-1547.

Philipp Cimiano, Johanna Völker, and Rudi Studer. 2006. Ontologies on demand? A description of the state-of-the-art, applications, challenges and trends for ontology learning from text. *Information, Wissenschaft und Praxis* Vol. 57, No. 6-7:315–320.

Shutong Feng, Hsien chin Lin, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2024. Infusing Emotions into Task-oriented Dialogue Systems: Understanding, Management, and Generation. *arXiv preprint arXiv:2408.02417* https://arxiv.org/abs/2408.02417.

Shutong Feng, Nurul Lubis, Benjamin Ruppik, Christian Geishauser, Michael Heck, Hsien-chin Lin, Carel van Niekerk, Renato Vukovic, and Milica Gašić. 2023. From chatter to matter: Addressing critical steps of emotion recognition learning in task-oriented dialogue. In Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani, editors, *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czechia, pages 85–103. https://doi.org/10.18653/v1/2023.sigdial-1.8.

Katerina T Frantzi and Sophia Ananiadou. 1999. The C-value/NC-value domain-independent method for multi-word term extraction. *Journal of Natural Language Processing* 6(3):145–179. https://doi.org/10.5715/jnlp.6.3_145.

Christian Geishauser, Carel van Niekerk, Nurul Lubis, Hsien-chin Lin, Michael Heck, Shutong Feng, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2024. Learning with an open horizon in ever-changing dialogue circumstances. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32:2352–2366. https://doi.org/10.1109/TASLP.2024.3385289.

Pierre-Yves Genest, Pierre-Edouard Portier, Elöd Egyed-Zsigmond, and Laurent-Walter Goix. 2022. Promptore-a novel approach towards fully unsupervised relation extraction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. pages 561–571.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gašić. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Toronto, Canada, pages 936–950. https://doi.org/10.18653/v1/2023.acl-short.81.

Michael Heck, Nurul Lubis, Carel van Niekerk, Shutong Feng, Christian Geishauser, Hsien-Chin Lin, and Milica Gašić. 2022. Robust dialogue state tracking with weak supervision and sparse data. *Transactions of the Association for Computational Linguistics* 10:1175–1192. https://doi.org/10.1162/tacl_a_00513.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašić. 2020. TripPy: A triple copy strategy for value independent neural dialog state tracking. In Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes, editors, *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting, pages 35–44. https://doi.org/10.18653/v1/2020.sigdial-1.4.

Vojtěch Hudeček, Ondřej Dušek, and Zhou Yu. 2021. Discovering dialogue slots with weak supervision. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, pages 2430–2442. https://doi.org/10.18653/v1/2021.acl-long.189.

Hsien-Chin Lin, Shutong Feng, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica

Gasić. 2023. Emous: Simulating user emotions in task-oriented dialogues. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, SIGIR '23, page 2526–2531. https://doi.org/10.1145/3539618.3592092.

Christopher W. Lynn and Danielle S. Bassett. 2020. How humans learn and represent networks. *Proceedings of the National Academy of Sciences* 117(47):29407–29415. https://doi.org/10.1073/pnas.1912328117.

Hiroshi Nakagawa and Tatsunori Mori. 2002. A simple but powerful automatic term extraction method. In *COLING-02: COMPUTERM 2002: Second International Workshop on Computational Terminology*. https://aclanthology.org/W02-1407.

Liang Qiu, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Structure extraction in task-oriented dialogues with slot clustering. https://doi.org/10.48550/ARXIV.2203.00073.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*. volume 34, pages 8689–8696.

Benjamin Matthias Ruppik, Michael Heck, Carel van Niekerk, Renato Vukovic, Hsien chin Lin, Shutong Feng, Marcus Zibrowius, and Milica Gašić. 2024. Local Topology Measures of Contextual Language Model Latent Spaces With Applications to Dialogue Term Extraction. *arXiv preprint arXiv:2408.03706* https://arxiv.org/abs/2408.03706.

Francesco Sclano and Paola Velardi. 2007. TermExtractor: A web application to learn the shared terminology of emergent web communities. In Ricardo J. Gonçalves, Jörg P. Müller, Kai Mertins, and Martin Zelm, editors, *Enterprise Interoperability II*. Springer London, London, pages 287–290. https://doi.org/10.1007/978-1-84628-858-6_32.

J. Toledo-Alvarado, Adolfo Guzman-Arenas, and G. Martínez-Luna. 2012. Automatic building of an ontology from a corpus of text documents using data mining tools. *Journal of applied research and technology* 10:398–404. https://doi.org/10.22201/icat.16656423.2012.10.3.395.

A. M. Turing. 1950. Computing machinery and intelligence. *Mind* 59(236):433–460. http://www.jstor.org/stable/2251299.

Carel van Niekerk, Christian Geishauser, Michael Heck, Shutong Feng, Hsien chin Lin, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2023. CAMELL: Confidence-based Acquisition Model for Efficient Self-supervised Active Learning with Label Validation. *arXiv preprint arXiv:2310.08944* https://arxiv.org/abs/2310.08944.

Carel van Niekerk, Andrey Malinin, Christian Geishauser, Michael Heck, Hsien-chin Lin, Nurul Lubis, Shutong Feng, and Milica Gašić. 2021. Uncertainty measures in neural belief tracking and the effects on dialogue policy performance. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 7901–7914. https://doi.org/10.18653/v1/2021.emnlp-main.623.

Renato Vukovic, David Arps, Carel van Niekerk, Benjamin Matthias Ruppik, Hsien-Chin Lin, Michael Heck, and Milica Gašić. 2024. Dialogue Ontology Relation Extraction via Constrained Chain-of-Thought Decoding. *arXiv preprint arXiv:2408.02361* https://arxiv.org/abs/2408.02361.

Renato Vukovic, Michael Heck, Benjamin Ruppik, Carel van Niekerk, Marcus Zibrowius, and Milica Gašić. 2022. Dialogue term extraction using transfer learning and topological data analysis. In Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy Li, Arash Ashrafzadeh, Daniel Hernández Garcia, Malihe Alikhani, David Vandyke, and Ondřej Dušek, editors, *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Edinburgh, UK, pages 564–581. https://doi.org/10.18653/v1/2022.sigdial-1.53.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. *arXiv preprint arXiv:2402.10200* https://arxiv.org/abs/2402.10200.

Joachim Wermter and Udo Hahn. 2006. You can't beat frequency (unless you use linguistic knowledge) – a qualitative evaluation of association measures for collocation and term extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sydney, Australia, pages 785–792. https://doi.org/10.3115/1220175.1220274.

Dian Yu, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent El Shafey, and Hagen Soltau. 2022. Unsupervised slot schema induction for task-oriented dialog. https://doi.org/10.48550/ARXIV.2205.04515.

Bhathiya Hemanthage

School of Mathematical and Computer Sciences.
Heriot-Watt University,
Riccarton, Edinburgh

hsb2000@hw.ac.uk

1 Research interests

My research focus on **Visual Dialogues** and **Generalized Visual-Language Grounding with Complex Language Context**. Specifically, my research aim to utilize Large Language Models (LLMs) to build *conversational agents capable of comprehending and responding to visual cues*.

Visual-Language Pre-trained (VLP) models, primarily utilizing transformer-based encoder-decoder architectures, are extensively employed across a range of visual-language tasks, such as visual question answering (VQA) and referring expression comprehension (REC). The effectiveness of these models stems from their robust visual-language integration capabilities. However, their performance is constrained in more complex applications like multimodal conversational agents, where intricate and extensive language contexts pose significant challenges. These tasks demands language-only reasoning before engaging in multimodal fusion. In response, my research investigates the application of Large Language Models (LLMs) with advance comprehension and generation capabilities to enhance performance in complex multimodal tasks, particularly multimodal dialogues.

In brief, my work in visual dialogues revolves around two major research questions. i) How to redefine visually grounded conversational agent architectures to benefit from LLMs ii) How to transfer the large body of knowledge encoded in LLMs to conversational systems.

1.1 End-to-end multi-modal conversational agents with Symbolic Scene Representation

The SIMMC 2.0 (Kottur et al., 2021) is a multi-modal task oriented dialogue proposed as part of DSTC-10 challenge with the goal of facilitating task oriented dialogue system which can understand visual context in addition to the linguistic context. This is challenging compared to both text-only dialogue datasets (such as ()) and image querying dialogue (such as ()) due to the simultaneous presence of signals from multiple modalities, which a user can refer to at any point in the conversation.

Despite the inherent complexity of multimodal dialogues, our work; (Hemanthage et al., 2023) introduce SimpleMTOD, which recasts all sub-tasks into a simple language model. In (Hemanthage et al., 2023) , we represent the visual information in a symbolic manner. SimpleMTOD combines de-localized object representation with token based spatial information representation.

However, (Hemanthage et al., 2023) and most other works on multimodal dialogue systems (Chen et al., 2023; Long et al., 2023) make a key unrealistic assumption in their inference processes. They assume the availability of a predefined catalog of items that may appear in a scene, and that this catalog remains fixed from training to inference. Our current work (Hemanthage et al., 2024) focus on addressing limitations in real-world applicability due to these unrealistic assumptions.

1.2 Modular multi-modal conversational agents with Pseudo-Labelling

End-to-end modeling with multimodal fusion has demonstrated significant advancements in various visual-language tasks, including phrase grounding (Plummer et al., 2015), referring expression comprehension (REC) (Yu et al., 2016; Nagaraja et al., 2016), and open vocabulary object detection (Gu et al., 2021). However, the applicability of these methods is limited when the language context is sophisticated, as in visual dialogues.

Modular approaches presents several advantages for the more complex multi-modal dialogue task. Firstly, modules can be designed to enable explicit text-only reasoning over the dialogue context, which is crucial in visual dialogue systems. Secondly, the modular approaches can mitigate the challenges posed by lengthy language contexts by summarizing the language context to extract only the essential information for the task before visual-language fusion.

Despite the advantages, a key challenge of the modular approach is the lack of annotated data for training intermediate modules. To address this, our work (Hemanthage et al., 2024), explore semi-supervised learning (SSL) setup where pseudo-labels generated by prompting a Large Language Model (LLM) serve as training targets for intermediate modules. Although our work focuses on Ambiguous Candidate Identification (ACI) in multimodal dialogues, the general approach of modularization with LLM-based pseudo-labelling can be extended to other

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 57–59

complex multimodal tasks with long language context.

2 Spoken dialogue system (SDS) research

Considering the remarkable advancements in artificial intelligence (AI), particularly with the emerge of large language models (LLMs), I anticipate that spoken dialogue systems (SDS) will soon become the preferred and most widespread form of human-machine interaction, overtaking touch and type-based systems. Moreover, I foresee the next generation of dialogue systems shifting their focus towards an embodied setting, moving away from the traditional mobile-phone-based voice assistants. These dialogue-guided embodied agents are expected to have capabilities extending from performing simple household chores like cooking and cleaning to serving as assistants in shopping malls or as receptionists in banks

While being optimistic about the future, it's crucial for us as young researchers to have a thorough understanding of the major limitations of current spoken dialogue systems (SDS) and to focus on overcoming these barriers. In my view, the limited capabilities of current dialogue models to ground multimodal information, especially in the presence of lengthy and sophisticated linguistic contexts, represent a significant obstacle to the progress of SDS.

Furthermore, the data intensive nature of current visual-language models is a key factor that hinders the adaptations of SDS for multi-modal settings. However, the emergence of LLMs and Multimodal LLMs, which can be fine-tuned with limited amount of data, offers a promising avenue for overcoming these challenges.

3 Suggested Topics for Discussion

- How can multimodal dialogue systems benefit from large language models (LLMs)?
- What are the challenges of using LLMs as annotators or pseudo-label generators for unimodal and multimodal dialogues?
- How can knowledge distillation from LLMs contribute to building generalized multimodal dialogue systems?
- End-to-end or Modular: Should we reconsider multimodal dialogue architectures in the era of LLMs?
- How can we use function-calling abilities of LLMs to build multimodal conversational agents?
- How can we develop a multimodal Large Language Model capable of multi-turn dialogues across multiple modalities?

References

Yirong Chen, Ya Li, Tao Wang, Xiaofen Xing, Xiangmin Xu, Quan Liu, Cong Liu, and Guoping Hu. 2023. Exploring prompt-based multi-task learning for multimodal dialog state tracking and immersive multimodal conversation. In Yun-Nung Chen, Paul Crook, Michel Galley, Sarik Ghazarian, Chulaka Gunasekara, Raghav Gupta, Behnam Hedayatnia, Satwik Kottur, Seungwhan Moon, and Chen Zhang, editors, *Proceedings of The Eleventh Dialog System Technology Challenge*. Association for Computational Linguistics, Prague, Czech Republic, pages 1–8. https://aclanthology.org/2023.dstc-1.1.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*.

Bhathiya Hemanthage, Christian Dondrup, Phil Bartie, and Oliver Lemon. 2023. SimpleMTOD: A simple language model for multimodal task-oriented dialogue with symbolic scene representation. In Maxime Amblard and Ellen Breitholtz, editors, *Proceedings of the 15th International Conference on Computational Semantics*. Association for Computational Linguistics, Nancy, France, pages 293–304. https://aclanthology.org/2023.iwcs-1.31.

Bhathiya Hemanthage, Christian Dondrup, Hakan Bilen, and Oliver Lemon. 2024. Divide and conquer: Rethinking ambiguous candidate identification in multimodal dialogues with pseudo-labelling. In *25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 4903–4912. https://doi.org/10.18653/v1/2021.emnlp-main.401.

Yuxing Long, Huibin Zhang, Binyuan Hui, Zhenglu Yang, Caixia Yuan, Xiaojie Wang, Fei Huang, and Yongbin Li. 2023. Improving situated conversational agents with step-by-step multi-modal logic reasoning. In Yun-Nung Chen, Paul Crook, Michel Galley, Sarik Ghazarian, Chulaka Gunasekara, Raghav Gupta, Behnam Hedayatnia, Satwik Kottur, Seungwhan Moon, and Chen Zhang, editors, *Proceedings of The Eleventh Dialog System Technology Challenge*. Association for Computational Lin-

guistics, Prague, Czech Republic, pages 15–24. https://aclanthology.org/2023.dstc-1.3.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, pages 792–807.

B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*. pages 2641–2649. https://doi.org/10.1109/ICCV.2015.303.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, pages 69–85.

Biographical sketch

Bhathiya Hemanthage is a PhD student at the Edinburgh Centre for Robotics Centre for Doctoral Training, supervised by Prof. Oliver Lemon, Dr. Christian Dondrup, and Dr. Phil Bartie of Heriot-Watt University and Dr. Hakan Bilen from University of Edinburgh. His current research focuses on Generalized Visual-Language Grounding with Complex Language Context. He holds a Master's (by research) degree from University of Moratuwa, Sri Lanka. His Master's thesis supervised by Dr. Uthayasanker Thayasivam was on Dialogue State Tracking for Low Resourced Languages. Bhathiya has several years of experience as a Senior Software Engineer.

Jingjing Jiang

Nagoya University
Nagoya, Aichi
Japan
jiang.jingjing.k6
@s.mail.nagoya-u.ac.jp

1 Research interests

The ultimate goal of my research is to develop human-like chat-oriented dialogue systems that establish long-term connections with users by satisfying their need for communication and affection. To achieve this, dialogue systems need to accurately understand the user's mental state and generate appropriate responses. However, most of the current dialogue systems interact with users relying solely on text or speech, which is insufficient for estimating the user's mental state.

Therefore, to enable dialogue systems to accurately capture the user's mental state, we focus on two areas: **construction and utilization of multimodal datasets** in human communication and **real-time multimodal affective computing**.

1.1 Construction and utilization of multimodal datasets

The primary interests of our group include processing and analyzing multimodal information in dialogue.

In face-to-face human communication, we inherently use verbal information, such as language and speech, and non-verbal information, including facial expressions, gaze, and gestures, to convey our intentions and ideas. The combination of these multiple modes of information is termed multimodal information. By comprehensively processing and analyzing multimodal information, human intentions and emotional states can be accurately comprehended during communication.

Several multimodal dialogue datasets have been constructed to date. For example, IEMOCAP (Busso et al., 2008) is a script-based human-human dialogue dataset containing speech, video, and facial motion capture. RECOLA (Ringeval et al., 2013) is a dataset that includes audio, visual, and physiological recordings regarding a collaborative dialogue task. Hazumi (Komatani and Okada, 2021) is a human-agent multimodal dialogue corpus containing audio, visual, and physiological data. However, these datasets lack comprehensive multimodal information during dialogue, which limits the scope and depth of research that can be conducted.

Consequently, our research group has collected a Japanese human-human dialogue dataset comprising a wide range of modalities, including speech, video, physiological signals, gaze, and body movement, as well as subjective evaluations of the interlocutor's emotional valences. All data are synchronized with timestamps. Furthermore, we analyzed the relationships between various physiological signals and subjective evaluations (Jiang et al., 2024). In future work, we plan to extend the analysis beyond physiological signals to understand and model various phenomena that occur in natural human communication.

1.2 Real-time multimodal affective computing

We also focus on developing a model that can detect or predict the interlocutors' emotional state in real-time for spoken dialogue systems.

In the field of affective computing, sentiment analysis and emotion recognition are combined to detect and model human emotional behavior. Most multimodal affective computing approaches in dialogue use text, speech, and video to identify the emotional state of the interlocutor. However, relying solely on this observable information makes it challenging to correctly recognize subtle emotional changes when the interlocutors do not explicitly express their emotions.

To address this limitation, extensive research (Katada et al., 2020; Keren et al., 2017) has been conducted on identifying an interlocutor's emotional state using physiological signals, such as heart rate and electrodermal activity. However, these studies typically conduct affective computing at the utterance or sentence level and do not consider the real-time nature of the user's mental state, which is essential for dialogue systems.

Therefore, we seek an effective method for data processing and multimodal feature fusion to construct a real-time emotion estimation model that leverages audiovisual information and physiological signals. Such a model needs to identify patterns that are generalizable across users. However, emotional expression varies significantly among individuals and is influenced by factors such as culture and personality. These individual differences are critical in affective computing and cannot be disregarded. Future studies should also consider models that adapt to individual users while preserving generalizability across diverse users.

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 60–61
September 16–17, 2024.

2 Spoken dialogue system (SDS) research

Recently, large language models (LLMs) have significantly improved the understanding of user input, retaining long-term contextual information and generating more fluent responses. They have also demonstrated the capability to recognize human emotions (Tak and Gratch, 2023). Building on these developments, multimodal LLMs (MM-LLMs) that process multimodal inputs and outputs across various modalities such as text, audio, and video, have undergone widespread development. Most current MM-LLMs are primarily employed to assist users with specific tasks such as image editing (Zhang et al., 2024); their potential to understand human emotions is also promising.

In the future, MM-LLMs capable of establishing long-term connections with users may be developed, which can access users' intentions and emotional states and respond accordingly. Furthermore, future dialogue systems might be widely integrated into humanoid robots. This integration could significantly enrich user-system interactions, for instance, by incorporating haptics, which has the potential to enhance user immersion and engagement during interactions with dialogue systems (Minato et al., 2023).

3 Suggested topics for discussion

I would like to discuss the following topics:

- What are the possible applications of a dialogue system that can acquire physiological signals?
- What would a dialogue system that builds long-term relationships with humans look like? What kind of appearance, interaction interface, and qualities would it possess?
- Assuming the existence of a dialogue system that can establish a long-term connection with humans, which would be more desirable: a dialogue system that has its personality and emotions, including the possibility of getting angry, or a dialogue system that solely focuses on satisfying all your needs without expressing its own emotions?

Acknowledgments

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42:335–359.

Jingjing Jiang, Ao Guo, and Ryuichiro Higashinaka. 2024. Estimating the Emotional Valence of Interlocutors Using Heterogeneous Sensors in Human-Human Dialogue. In *Proc. SIGDIAL*.

Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. 2020. Is She Truly Enjoying the Conversation? Analysis of physiological signals toward adaptive dialogue systems. In *Proc. ICMI*. page 315–323.

Gil Keren, Tobias Kirschstein, Erik Marchi, Fabien Ringeval, and Björn Schuller. 2017. End-to-end learning for dimensional emotion recognition from physiological signals. In *Proc. ICME*. pages 985–990.

Kazunori Komatani and Shogo Okada. 2021. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *Proc. ACII*. pages 1–8.

Takashi Minato, Ryuichiro Higashinaka, Kurima Sakai, Tomo Funayama, Hiromitsu Nishizaki, and Takayuki Nagai. 2023. Design of a competition specifically for spoken dialogue with a humanoid robot. *Advanced Robotics* 37:1349–1363.

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proc. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. pages 1–8.

Ala N. Tak and Jonathan Gratch. 2023. Is GPT a Computational Model of Emotion? In *Prco. ACII*. pages 1–8.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601* .

Biographical sketch

Jingjing Jiang is a master's student at the Graduate School of Informatics, Nagoya University, under the supervision of Prof. Ryuichiro Higashinaka. She is interested in dialogue systems, multimodal interaction, and affective computing. During her PhD course, she aspires to broaden her perspectives by collaborating with other research institutions on multimodal interaction in dialogue.

Shiyuan Huang

University of California, Santa Cruz
1156 High St
Santa Cruz
California, United States, 95064

shuan101@ucsc.edu
https://shiyuan-eric.github.io/website/

1 Research interests

The rapid growth in the development of generative AI models has made their evaluation as crucial as uncovering their generative capabilities, such as audio text, audio, image and video generation. My research is focused on analyzing these models in terms of their **explainability**, **interpretability**, and **trustworthiness**.

Explainability focuses on the decision-making processes of these models. My research seeks to answer the question: Can the model explain how it made a particular decision? Additionally, it explores what can help the model generate meaningful and understandable explanations about the reasons behind its predictions. Given the nature of neural networks, analyzing parameters in each neuron is often unproductive. Therefore, various methods, such as post-hoc analysis, have been developed to address this question from different angles. However, many methods, such as post-hoc analysis, merely scratch the surface of neural networks. Much further research is needed to address the numerous unresolved problems in this emerging field.

Interpretability involves understanding the inner workings of the models. Given their powerful generative capabilities, it is challenging to determine whether the model has fully comprehended all requirements and generated accurate content, especially when the user is unsure of the correct answer. Thus, I am interested in causal tracing , such as mechanistic interpretability, to gain a deeper understanding of the models.

Both explainability and interpretability aim to achieve the same goal: understanding the generation process and explaining the capabilities of generative models. This understanding will enhance user experience by increasing trust in and effective utilization of the models' outputs, which leads to the aspect of **trustworthiness**.

Given the discussion of research concepts that I am interested in, here are some methods and applications that utilize these concepts:

1.1 Traditional AI Methods + Generative AI

With the long time development of AI, there are many methods being well studied in the field of explainability. For instance, feature attribution is a method that being used to determine how much each feature in a model contribute to the evaluation. My prior work (Huang et al., 2023) tests the performance of language models by using feature attribution method. And it turns out that many traditional AI methods are not well-suited to super-power generative AI. For instance, as a human, sometimes what matters most in a classification problem is not a specific feature but the overall impression from many features. It will be a very potential topic to adapt traditional AI methods to new generative AI model and improve them to better fit this type of high intelligent models.

1.2 Generative AI + Logic

Generative AI is powerful yet unpredictable. Meanwhile, logic is good at eliminating uncertainty. There are many existing works in the field of Neural Symbolic AI that tried to combine logic to deep neural network (Yang et al., 2023; Zhang et al., 2023). Especially, there are many works that use logic to improve the explainability of models such as adding a external knowledge base in the model so that every part of the output can be traced from the knowledge bases (Razniewski et al., 2021; Sun et al., 2021). Slightly different than this, I am interested in trying new methods to incorporate logic with generative AI to improve the explainability and interpretability. For instance, could we use logic as a helpful guide, without strictly adhering to its rules, to enhance the ability of language models to generate better explanations?

1.3 Education Application

I am also interested in the application of generative AI in the field of education (Niousha et al., 2024). There has been a trend that students tend to use generative AI like ChatGPT to seeking for answers instead of searching for answers online. From my point of view, the difference between asking ChatGPT and search engine is very similar to learning via teacher and textbook. The users tend to prefer a more structure, with just right amount of information instead of being overwhelmed. I am interested in applying the power of generative AI to be a personal tutor that not only provides correct solutions but also provides customized feedback for different users.

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 62–63
September 16–17, 2024.

2 Spoken dialogue system (SDS) research

I will never underestimate the rapid development of leading research in dialogue systems. My prediction for dialogue research in the next 5 to 10 years is that we will expand beyond human language, potentially being applied in the field of biology, particularly, studying the dialog system of animals like whales, birds, and dogs. With this field of research being explored, there will be chance that in the future, there will be no language barrier between animals and humans.

Based on the current trajectory of SDS research, I have observed an increasing focus on developing systems with high accuracy and effective penalization mechanisms. With the success of language models, I am confident that SDS will emerge as a significant trend in the coming years. Language models offer a robust foundation by generating the necessary content, but the next challenge lies in ensuring proper verbal delivery. For example, how should an SDS respond when a user interrupts it mid-response? Should it continue or start over? An SDS must not only deliver accurate answers but also communicate in a natural manner.

3 Suggested topics for discussion

- With the advancement of generative AI models, their capabilities have expanded significantly. These models now support inputs and outputs across various modalities, including text, voice, image, and video. However, a major challenge in this field is the lack of a definitive ground truth for evaluation. Traditional AI models often rely on a single correct answer to assess performance. In contrast, generative models produce a range of possible outputs, making it nearly impossible to pinpoint just one correct solution for a given task. Therefore, developing innovative evaluation metrics, particularly in the field of SDS, will be a crucial area of research.

- Generative AI has shown some impressive abilities in many areas. But often, its full potential is not being completely utilized or it's being used in ways that don't quite fit. For instance, in the field of education, people tend to use generative AI as a knowledgeable search engine, which does not really take advantage of what it can do. The same goes for SDS, where figuring out the best way to unlock all their power is a challenge worth discussing.

- Another pressing concern is the wrongful usage associated with dialogue systems that include voice recreation which can be exploited for scams by mimicking someone's voice or using one singer's voice for another song. Looking ahead, regulations need to be refined to better handle problems such as the misuse of content creation that infringes on copyrights or is used in scams. Exploring how these regulations can be improved to address such issues will be crucial for the responsible development and deployment of dialogue systems.

References

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207* .

Rose Niousha, Muntasir Hoq, Bita Akram, and Narges Norouzi. 2024. Use of large language models for extracting knowledge components in cs1 programming exercises. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*. pages 1762–1763.

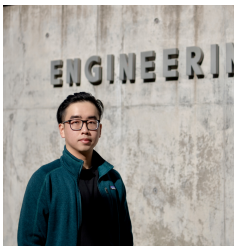
Simon Razniewski, Andrew Yates, Nora Kassner, and Gerhard Weikum. 2021. Language models as or for knowledge bases. *arXiv preprint arXiv:2110.04888* .

Haitian Sun, Patrick Verga, Bhuwan Dhingra, Ruslan Salakhutdinov, and William W Cohen. 2021. Reasoning over virtual knowledge bases with open predicate relations. In *International Conference on Machine Learning*. PMLR, pages 9966–9977.

Zhun Yang, Adam Ishay, and Joohyung Lee. 2023. Coupling large language models with logic programming for robust and general reasoning from text. *arXiv preprint arXiv:2307.07696* .

Hanlin Zhang, Jiani Huang, Ziyang Li, Mayur Naik, and Eric Xing. 2023. Improved logical reasoning of language models via differentiable symbolic programming. *arXiv preprint arXiv:2305.03742* .

Biographical sketch



Shiyuan Huang is a first-year PhD student at the University of California, Santa Cruz. His research focuses on the explainability and interpretability of language models. Prior to his PhD studies, Shiyuan completed his master's degree at the University of California, Santa Cruz. His master's research project, titled "Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations," explored the use of feature attribution methods with large language models to measure their explainability.

Zachary Yang

McGill University
Montreal, Quebec
Canada
H3A 0G4

zachary.yang@mail.mcgill.ca
https://rstzzz.github.io/

1 Research interests

Ensuring safe online environments is a formidable challenge, but nonetheless an important one as people are now chronically online. The increasing online presence of people paired with the prevalence of harmful content such as toxicity, hate speech, misinformation and disinformation across various social media platforms (Ciftci et al., 2017; Watanabe et al., 2018; Mohan et al., 2017; Döring and Mohseni, 2020) and within different video games (Silva et al., 2020) calls for stronger detection and prevention methods. According to the Anti-Defamation League's 2023 report, toxicity in gaming is "now so pervasive that it has become the norm for many players" (ADL, 2023). Moreover, concerns among experts are rising about the potential for advanced AI to cause significant harm through manipulation, even before ChatGPT. Sophisticated AI-assisted information operations have already emerged as a growing concern (Hitkul et al., 2021; McKay and Tenove, 2021; Tucker et al., 2017). Already in 2022, systems like Cicero, an AI language agent, had demonstrated capabilities in persuasion and deception within social gaming environments (FAIR et al., 2022).

To foster a healthier online community, companies have experimented with various approaches to curb the dissemination of toxic and harmful content. These efforts range from word censorship and player bans to content moderation and flagging controversial posts for review.

My research interests primarily lie in **applied natural language processing for social good**. Previously, I focused on measuring partisan polarization on social media during the COVID-19 pandemic and its societal impacts (Yang et al., 2021, 2024b). Currently, at Ubisoft La Forge, I am dedicated to **enhancing player safety within in-game chat systems** by developing methods to **detect toxicity** (Yang et al., 2023), **evaluating the biases** in these detection systems (Van Dorpe et al., 2023), and **assessing the current ecological** state of online interactions (Yang et al., 2024a). Additionally, I am engaged in **simulating social media environments using LLMs** to ethically test detection methods, **evaluate** the effectiveness of current mitigation strategies, and potentially introduce new, successful strategies.

1.1 Safety With In-Game Chat Systems

Ensuring player safety in online games begins with effectively **detecting and preventing toxicity within in-game chat systems**. As more games feature online multiplayer modes with team and all-chat options, players engage in conversations through both text and speech. While definitions of toxicity and hate speech vary among researchers and industry platforms, we adhere to the definitions outlined by the Fair Play Alliance (Lewington, 2021). My initial focus was on improving the detection of toxicity (Yang et al., 2023). Previous research primarily focused on social media, revealing that incorporating the context of parent posts did not enhance performance. Since in-game conversations are more cohesive, I integrated techniques from dialogue systems, including previous chat lines and speaker segmentation, to model multi-turn conversations. This enabled the creation of a context-aware model capable of detecting toxicity in real-time game chat.

While advancing these LLMs to detect toxicity is crucial, addressing the potential biases inherent in them is equally important. Consequently, our team **measured identity biases** using a game-focused dataset (Van Dorpe et al., 2023). Inspired by reactivity analysis, we had users annotate whether a sentence was toxic. We generated sentences typical of in-game chat while replacing key words with specific attributes (e.g., black, trans), groups (e.g., white, young people, women), and personas (e.g., artist, streamer). This approach allowed us to measure whether detection algorithms reacted differently to certain terms, leading to unfair treatment of specific groups of players, either through over-penalization or under-penalization.

To fully grasp the current state of toxicity within in-game chat systems, we ran our detection system on a full year's worth of chat data (Yang et al., 2024a). This research examined in-game events, the number of players and matches played, and the types of games. We recognize that any *deployed system will naturally elicit reactions from players*. A holistic approach that considers both the technical aspects of toxicity detection and the socio-cultural environment of online gaming commu-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 64–67

nities is essential. By gaining a **comprehensive** understanding of these factors, the player safety team can devise more effective strategies to foster a **safer and more inclusive gaming ecosystem**. Capturing the current ecological state before deployment allows us to measure the impact of this detection system in conjunction with any mitigation strategies deployed.

1.2 Simulating Social Media w/ LLMs

With the rise of generative LLMs, the question arises whether they can be utilized to **simulate high-fidelity reflections of social environments**, creating a sandbox mode that allows us to **ethically test detection and mitigation strategies** for social harms such as manipulation during election discourse, the spread of toxicity, hate speech, misinformation, and disinformation.

LLMs have demonstrated the ability to reflect political attitudes (Argyle et al., 2023), showcase personality traits (Serapio-García et al., 2023), and simulate social interaction (El-Kishky et al., 2022; Törnberg et al., 2023). Researchers have already begun using LLMs on a small scale, such as simulating a small town (Park et al., 2023) and social media (Törnberg et al., 2023). Our current work at my research lab focuses on expanding these simulations to a larger scale using the open-source social media platform Mastodon as the environment. We will attempt to employ personas that reflect reality-matching demographics and activity/network attributes from massive Twitter datasets (Pelrine et al., 2023b; Yang et al., 2021; Pelrine et al., 2023a; Orlovskiy et al., 2024). Additionally, we will then introduce benign agents fine-tuned for varying levels of susceptibility to misinformation, mirroring human populations (Liu et al., 2023), and malicious agents that would replicate severe manipulation threats. This controlled setting will then enable us to quantify manipulation effects and assess the effectiveness of proposed defenses, yielding broad applications across AI safety, social science, and policy.

2 Spoken dialogue system (SDS) research

The research on player safety systems is closely connected to spoken dialogue system research, as players frequently communicate through text and speech. Leveraging LLMs to simulate these social environments allows us to ethically test current prevention methods and understand the effectiveness and potential unintended consequences of various mitigation strategies. Spoken dialogue systems, such as chatbots and virtual assistants, rely on natural language processing and generation, which are also fundamental to LLMs. By studying the manipulation and mitigation of social harms in these simulations, we can develop more robust and ethical dialogue systems capable of detecting and preventing misinformation, hate speech, and other malicious content in real-time interactions. This cross-disciplinary approach enhances the safety and trustworthiness of AI-driven communication technologies in both written and spoken forms, ultimately contributing to a more secure and inclusive digital environment.

3 Suggested topics for discussion

- Understanding and mitigating social harms: Addressing toxicity and misinformation through high-fidelity simulation environments.
- Enhancing safety in online environments: multimodal models, handling multi-lingual conversations (where a sentence can contain more than one language), and addressing accents and region-specific dialogue.
- Personification of LLM agents: Developing coherent responses based on backstory and personality.
- Ethically simulating social media sandbox environments at scale with LLM agents: Including the posting of text, speech, images, and video.
- Re-balancing the playing field between good and bad actors: Strategies for countering societal-scale manipulation.

References

ADL. 2023. Hate is no game: Hate and harassment in online games 2023. *Anti-Defamation League* https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2023.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31(3):337–351.

Tuba Ciftci, Liridona Gashi, René Hoffmann, David Bahr, Aylin Ilhan, and Kaja Fietkiewicz. 2017. Hate speech on facebook. Fourth European Conference on Social Media Research, pages 425–433.

Nicola Döring and M. Mohseni. 2020. Gendered hate speech in youtube and younow comments: Results of two content analyses. *Studies in Communication and Media* 9:62–88. https://doi.org/10.5771/2192-4007-2020-1-62.

Ahmed El-Kishky, Thomas Markovich, Serim Park, Chetan Verma, Baekjin Kim, Ramy Eskander, Yury Malkov, Frank Portman, Sofía Samaniego, Ying Xiao, and Aria Haghighi. 2022. Twhin: Embedding the twitter heterogeneous information network for personalized recommendation. In *Proceedings*

of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, KDD '22. https://doi.org/10.1145/3534678.3539080.

Meta Fundamental AI Research Diplomacy Team FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science* 378(6624):1067–1074.

Hitkul, Avinash Prabhu, Dipanwita Guhathakurta, Jivitesh jain, Mallika Subramanian, Manvith Reddy, Shradha Sehgal, Tanvi Karandikar, Amogh Gulati, Udit Arora, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2021. Capitol (pat)riots: A comparative study of twitter and parler.

Robert Lewington. 2021. Being 'targeted' about content moderation:. *Fair Play Alliance* pages 1–21. https://fairplayalliance.org/wp-content/uploads/2022/06/FPA-Being-Targeted-about-Content-Moderation.pdf.

Yanchen Liu, Mingyu Derek Ma, Wenna Qin, Azure Zhou, Jiaao Chen, Weiyan Shi, Wei Wang, and Diyi Yang. 2023. From scroll to misbelief: Modeling the unobservable susceptibility to misinformation on social media. *arXiv preprint arXiv:2311.09630* .

Spencer McKay and Chris Tenove. 2021. Disinformation as a threat to deliberative democracy. *Political research quarterly* 74(3):703–717.

Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The impact of toxic language on the health of reddit communities. Canadian Conference on Artificial Intelligence, pages 51–56. https://doi.org/10.1007/978-3-319-57351-9_6.

Yury Orlovskiy, Camille Thibault, Anne Imouza, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Uncertainty resolution in misinformation detection. https://arxiv.org/abs/2401.01197.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. pages 1–22.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023a. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint arXiv:2305.14928* .

Kellin Pelrine, Anne Imouza, Zachary Yang, Jacob-Junqi Tian, Sacha Lévy, Gabrielle Desrosiers-Brisebois, Aarash Feizi, Cécile Amadoro, André Blais, Jean-François Godbout, et al. 2023b. Party prediction for twitter. *arXiv preprint arXiv:2308.13699* .

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models.

Bruno Silva, Mirian Tavares, Filipa Cerol, Susana Silva, Paulo Alves, and Beatriz Isca. 2020. Playing against hate speech -how teens see hate speech in video games and online gaming communities. *Journal of Digital Media and Interaction* 3:34–52. https://doi.org/https://doi.org/10.34624/jdmi.v3i6.15064.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984* .

Joshua A Tucker, Yannis Theocharis, Margaret E Roberts, and Pablo Barberá. 2017. From liberation to turmoil: Social media and democracy. *J. Democracy* 28:46.

Josiane Van Dorpe, Zachary Yang, Nicolas Grenon-Godbout, and Grégoire Winterstein. 2023. Unveiling identity biases in toxicity detection : A game-focused dataset and reactivity analysis approach. In Mingxuan Wang and Imed Zitouni, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, Singapore, pages 263–274. https://doi.org/10.18653/v1/2023.emnlp-industry.26.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access* 6:13825–13835. https://doi.org/10.1109/ACCESS.2018.2806394.

Zachary Yang, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2023. Towards detecting contextual real-time toxicity for in-game chat. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 9894–9906. https://doi.org/10.18653/v1/2023.findings-emnlp.663.

Zachary Yang, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2024a. Game on, hate off: A study of toxicity in online multiplayer environments. *ACM Games* Just Accepted. https://doi.org/10.1145/3675805.

Zachary Yang, Anne Imouza, Kellin Pelrine, Sacha Lévy, Jiewen Liu, Gabrielle Desrosiers-Brisebois,

Jean-François Godbout, André Blais, and Reihaneh Rabbany. 2021. Online partisan polarization of covid-19. In *2021 International Conference on Data Mining Workshops (ICDMW)*. IEEE, pages 893–901.

Zachary Yang, Anne Imouza, Maximilian Puelma Touzel, Cecile Amadoro, Gabrielle Desrosiers-Brisebois, Kellin Pelrine, Sacha Levy, Jean-Francois Godbout, and Reihaneh Rabbany. 2024b. Regional and temporal patterns of partisan polarization during the covid-19 pandemic in the united states and canada. https://arxiv.org/abs/2407.02807.

Biographical sketch

Zachary Yang is a PhD candidate at McGill University, supervised by Professor Reihaneh Rabbany, specializing in applied natural language processing for social good. His research focuses on studying toxicity, misinformation, and polarization in games and social media. Zachary is also a member of Mila - Quebec AI Institute and the Centre for the Study of Democratic Citizenship.

Previously, he developed scalable methods to measure partisan polarization on social media during the COVID-19 pandemic, with his work published in IEEE VIS and ICDMW. Currently, his research at Ubisoft La Forge aims to improve and prevent toxicity detection within game chat and create industry-leading player content safety systems. This work has led to publications in EMNLP 2023 and a presentation at the Ethical Games Conference in 2024. After completing his PhD, Zachary plans to join a research lab in industry, bridging academia and industry to deploy more systems with humans-in-the-loop, enhancing safety and ease of use.

Sangmyeong Lee

Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara-ken, Japan

lee.sangmyeong.lo3@is.naist.jp

1 Research interests

My research focuses on understanding **semantic structures** in **multimodal dialogue environments**. I'm particularly interested in using graphs to represent meaning, such as **Scene Graph** for visual information[Johnson et al. (2015)] and **Abstract Meaning Representation (AMR)** for language[Banarescu et al. (2013)]. During my masters, I worked on enhancing vision and language models to better differentiate structurally ambiguous image-caption pairs[Sangmyeong et al. (2023)] using linguistic formalism. For my PH.D., I'm exploring a new task: **Object State Inference from Verbal Instructions**. I plan to use the graph structures mentioned earlier to manage relations and attributes of multiple objects inside a visual scene, and accurately express user instructions.

1.1 Semantic Structures in Multimodal Environments

For real-world applications to understand multimodal environments, models must accurately align visual scenes with their corresponding language descriptions. However, natural language often contains structural ambiguity, where a single sentence can have multiple meanings due to different possible phrase structures. This makes it challenging to match vision and language one-to-one, which can lead to difficulties in conveying user intentions accurately, decreasing usability. During my master's, I worked on using various linguistic formalisms, such as syntax trees and semantic parsed graphs, as inputs into the Contrastive Language Image Pre-trained (CLIP) model[Radford et al. (2021)] to improve its ability to distinguish between ambiguous contexts.

1.2 Object State Inference from Verbal Instructions

In the real world, a visual environment consists of multiple objects with physical attributes and inter-relationships governed by the laws of physics. Understanding how these states change due to external factors, such as user instruction, is crucial for task-oriented dialogue systems like cooking robots. My research interest is in simulating and predicting how object states change based on verbal instructions. This field is significant for two reasons: it enhances the system's ability to comprehensively understand visual contexts and instructions, and it can warn users if their instructions might lead to dangerous situations (e.g. putting an egg in a microwave). Previous research used dictionary data structures to represent individual objects, yielding good results but struggling with representing inter-positional relationships[Zellers et al. (2021)]. My focus is on adapting graph structures for this task to better represent complex visual scenes and user instructions.

2 Spoken dialogue system (SDS) research

The field of SDS is undergoing a significant transformation with the advent of Large Language Models (LLMs), such as Chat-GPT. This development has highlighted a distinction between academic and industry research, as the latter has resolved numerous SDS challenges using vast amounts of data in an end-to-end fashion, which is often unaffordable for academia. Consequently, academia needs to establish its own specific research trends to coexist or even leverage LLMs. One potential area is evaluating LLM performance and analysing their principles to identify limitations in achieving human-level intelligence[Sravanthi et al. (2024)].

Meanwhile, my focus is on the novel role of visual and linguistic structural information in the modern era of SDS. Traditionally, structural information has been used to enhance generation models, providing strict structural details absent in plain texts and pixel-level images[Johnson et al. (2018)]. In the LLM era, structural information continues to be valuable, especially since LLMs are too large for use in all specific tasks[Hua et al. (2023)]. However, as computing power advances, LLMs will likely be applied more broadly. My focus on the use of structural information for SDS is divided into two main areas. First, as the complexity of environments increases, structural information such as scene graphs can effectively manage objects and subspaces, especially when labelled with attributes. Second, graph structures like scene graphs and AMR are robust representations of meaning. Generating these structures demonstrates a system's understanding of its surrounding environment and user instructions, facilitating a shared understanding between the user and the system as dialogue progresses.

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 68–69

3 Suggested topics for discussion

I suggest the following three topics for discussion during the vent, focusing on the new directions the SDS research community should explore:

- Coexist with LLMs: I hope to discuss with fellow researchers the future direction of SDS in light of LLM advancements. We should consider what LLMs can and cannot do, whether their current limitations are temporary, and which tasks our research should prioritise. I'm interested in developing benchmarks to assess and explain LLM comprehension abilities and creating a generation framework where LLMs play specific roles.

- Role of Structural Information: As previously mentioned, the traditional role of structural information in assisting generation models is evolving. LLMs. with their extensive pre-training on large datasets, now possess a high level of semantic knowledge. I want to explore how structural information can be applied in SDS research, such as managing complex situations efficiently or generating a common semantic ground for user-system interactions.

- Disambiguation Strategy: When users' language inputs contain ambiguity, the simplest solution is to confirm the intended meaning with the user. However, sometimes it is better for the system to disambiguate using commonsense-based plausibility. Developing a strategy for disambiguation can make the system's dialogue more human-like, enhancing user comfort.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *LAW@ACL*. https://api.semanticscholar.org/CorpusID:7771402.

Bobby Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. Improving long dialogue summarization with semantic graph representation. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:259859068.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 1219–1228. https://api.semanticscholar.org/CorpusID:4593810.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pages 3668–3678. https://api.semanticscholar.org/CorpusID:16414666.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Lee Sangmyeong, Seitaro Shinagawa, and Satoshi Nakamura. 2023. Improving image discrimination ability through understanding of textual syntactic information in clip. *Meeting on Image Recognition and Understanding (MIRU)* .

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. *ArXiv* abs/2401.07078. https://api.semanticscholar.org/CorpusID:266999533.

Rowan Zellers, Ari Holtzman, Matthew E. Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. 2021. Piglet: Language grounding through neuro-symbolic interaction in a 3d world. *ArXiv* abs/2106.00188. https://api.semanticscholar.org/CorpusID:235266260.

Biographical sketch

Sangmyeong Lee received his bachelor from Korea University in February 2021, where he majored in linguistics and informatics. During the study his primary interest was in theoretical semantics, which led him to continue his study at Nara Institute of Science and Technology (NAIST). During the master's from October 2021 to March 2024, he was affiliated with Augmented Human Communication Laboratory, where he worked on structural disambiguation of vision and langauge model via linguistic formalism. From April 2024, he is affiliated with Intelligence Robot Dialogue Laboratory focusing on leveraging multimodal semantic structural information for object state inference. He was also nominated as recipient of Nara Institute of Science and Technology Support Project Ver.2 for Innovative Doctoral Students in the Field of Multidisciplinary Research in Advanced Science and Technology (NAIST Granite Program).

Nicolas Wagner

Natural Language Generation and Dialogue Systems Group
University of Bamberg
96047 Bamberg
Bavaria, Germany

nicolas.wagner@uni-bamberg.de
www.uni-bamberg.de/en/ds/team/wagner/

1 Research interests

My research interests include **multi-user dialogue systems** with a focus on **user modelling** and the development of **moderation strategies**. Contemporary Spoken Dialogue Systems (SDSs) frequently lack the ability to interact with more than one user simultaneously. Moreover, I am interested in researching on the **Controllability of Language Generation** using **Large Language Models** (LLMs). Our hypothesis is that an integration of **explicit dialogue control signals** improves the Controllability and Reliability of generated sequences independently of the underlying LLM.

1.1 Multi-User Dialogue Systems

Although group interactions play an essential role in people's daily lives, research on multi-user dialogue systems is rather underrepresented. A reason for that could be the various challenges associated with this topic: Turn-taking strategies, addressee detection, and the lack of suitable training data, just to name a few. Researchers often face the issue that there are generally no guidelines or best practices for developing new multi-user SDSs, since most publications present systems which are strongly focusing on one specific task and are therefore not generalisable.

To gain insights into how users would like to be addressed by an SDS during a group conversation, we conducted a user study (Wagner et al., 2019). Here, we identified which system behaviours were perceived as less obtrusive and beneficial for the dialogue flow. As a next experiment, we evaluated moderation strategies for group chats (Wagner et al., 2022). The moderation strategies were intended to support groups in negotiating joint appointments and were rated as helpful.

Another topic I am interested in is human-robot interaction. We investigated recommendation strategies in a scenario with a household assistant robot (Kraus et al., 2022). In context of multi-user interaction, we examined how users perceived the usability of different dialogue strategies in a quiz game setup (Wagner et al., 2023). We intend to improve the developed strategies through the use of LLMs, which leads to the next section.

1.2 Controllable Language Generation

Task-oriented dialogue systems are designed to assist users in accomplishing specific tasks through natural language interactions. Traditional systems rely on a pipeline architecture with components for Natural Language Understanding, Dialogue Management, and Text Generation (Jokinen and McTear, 2009). Recently, LLMs are applied to substitute these components, as their generated outputs are much more flexible and appear more natural.

Since no task-specific data can be presented during pre-training, it is necessary to adapt models to a downstream task. Contemporary approaches include fine-tuning (Ouyang et al., 2022) and in-context learning (Brown et al., 2020). Although this equips models with certain capabilities to maintain context over conversations, it does not prevent the risk of incorrect responses or hallucinations. Further approaches like retrieval-augmented generation (Lewis et al., 2020) and automated generation of prompt templates (Sánchez Cuadrado et al., 2024) are supposed to provide additional knowledge and task-dependent prompt design.

However, none of these techniques consider the use of explicit control signals to control the dialogue flow, as they rely instead on implicit dialogue modelling within the neural net of transformer-based LLMs. To overcome these limitations, we propose equipping the system with a component for explicit dialogue control similar to the traditional pipeline architecture - the dialogue controller. For this, we have conducted a baseline experiment in which we showed that a dialogue controller improves the controllability of generated outputs (Wagner and Ultes, 2024). Specifically, the generated responses were more likely to correspond to human-annotated references. The proposed controller is designed to extract task-relevant data from a knowledge source and to provide a prompt instruction depending on the user input and intention. The system architecture is depicted in Figure 1. Furthermore, we plan to conduct experiments on constrained decoding as described in (Shin et al., 2021). This way, we expect to further enhance control over the generated output and provide users with more reliable responses.

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 70–72
September 16–17, 2024.

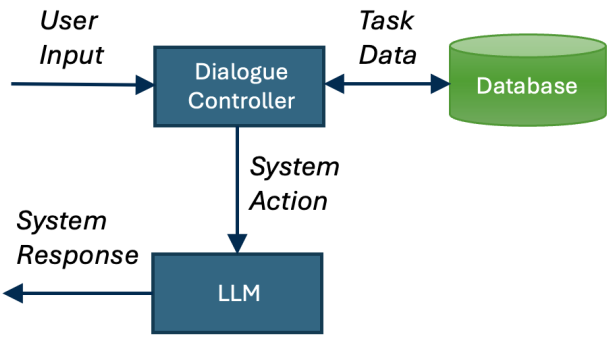


Figure 1: Depiction of the dialogue control architecture.

2 Spoken dialogue system (SDS) research

In my opinion, LLMs will increase the performance and popularity of SDSs and thus have an impact on research. However, I believe spoken language has its greatest potential in domains where information cannot be conveyed more efficiently or conveniently by other means. This includes voice assistants, smart speakers, health sector applications, or smart home environments. Personalisation and context-awareness will also play a vital role in the future, which may enable SDSs to become more and more useful in everyday life. Moreover, the research on multi-user dialogue systems needs to be intensified and common design rules established.

For the young research community, I would welcome the idea of actively participating in the development and application of ethical guidelines for the use of artificial intelligence, as this is where I expect major differences between the aims of industry and academia.

3 Suggested topics for discussion

My suggestions centre on the field of multi-user dialogue systems, and controllable natural language generation using Large Language Models.

- **Multi-User Dialogue Systems**: What are the best practices for development and evaluation? How to deal with the challenges in user modelling and dialogue state tracking? What turn-taking and moderation strategies should be used?
- **Natural Language Generation**: How to design explicit control signals to improve the controllability of language generation? Can constrained decoding enhance the response quality?
- **Actionable Evaluation Metrics**: Which metrics are applicable to measure the perceived naturalness of SDSs? How can they be used for policy optimisation? Can they also be applied for multi-user SDS?

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

Kristina Jokinen and Michael McTear. 2009. *Spoken dialogue systems*. Morgan & Claypool Publishers.

Matthias Kraus, Nicolas Wagner, Wolfgang Minker, Ankita Agrawal, Artur Schmidt, Pranav Krishna Prasad, and Wolfgang Ertel. 2022. Kurt: A household assistance robot capable of proactive dialogue. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI 2022)*. ACM/IEEE.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33:9459–9474.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35:27730–27744.

Jesús Sánchez Cuadrado, Sara Pérez-Soler, Esther Guerra, and Juan De Lara. 2024. Automating the development of task-oriented llm-based chatbots. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. Association for Computing Machinery, New York, NY, USA, CUI '24.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 7699–7715.

Nicolas Wagner, Matthias Kraus, Niklas Lindemann, and Wolfgang Minker. 2023. Comparing multi-user interaction strategies in human-robot teamwork. In *In Proceedings of the 13th International Workshop on Spoken Dialogue Systems Technology (IWSDS '23)*. Springer.

Nicolas Wagner, Matthias Kraus, Niklas Rach, and Wolfgang Minker. 2019. How to address humans: System barge-in in multi-user hri. In *Proceedings of the*

10th International Workshop on Spoken Dialog Systems Technology (IWSDS 2019).

Nicolas Wagner, Matthias Kraus, Tibor Tonn, and Wolfgang Minker. 2022. Comparing moderation strategies in group chats with multi-user chatbots. In *In Proceedings of the fourth ACM Conference on Conversational User Interfaces (CUI '22)*. ACM.

Nicolas Wagner and Stefan Ultes. 2024. On the controllability of large language models for dialogue interaction. In *To be published at SIGDIAL 2024*.

Biographical sketch

Nicolas Wagner received his B.Sc. and M.Sc. degrees in communications and computer engineering with focus on human-machine interaction at Ulm University, Germany, in 2018. From 2018 to 2024, he was a research assistant at the Dialogue Systems Group of Ulm University, Germany. Since 2019, he is also enrolled as a Joint-PhD candidate at the University of Granada, Spain. He is currently pursuing his PhD at both institutions. He joined the Natural Language Generation and Dialogue Systems Group at the University of Bamberg in 2024. His research interests include the development of multi-user dialogue systems and controlled language generation.

Ryuichi Uehara

The University of Electro-Communications
Chofu, Tokyo
Japan

r-uehara@uec.ac.jp

1 Research interests

My research interest involves **persona dialogue systems**, which use the profile information of a character or real person, called a persona, and responds accordingly. Persona dialogue systems can improve the consistency of the system's responses (Li et al., 2016), users' trust (Higashinaka et al., 2018), and user enjoyment (Miyazaki et al., 2021).

My current research focuses on persona dialogue systems, especially dialogue agents that role-play as fictional characters. The first task involves obtaining the dialogue and personas of novel characters and building a dialogue corpus. The second task involves evaluating whether the dialogue agent's responses are character-like relative to the context. The goal of these studies is to allow dialogue agents to generate responses that are more character-like.

1.1 Constructing a Dialogue Corpus for Role-playing

The main focus when assessing a dialogue system that simulates a character is the accuracy with which the system reflects the character's traits in its responses. To compare the system's responses with those of the character, we need a corpus containing the character's dialogue data. Dialogue corpora related to character role-play include ChatHaruhi (Li et al., 2023), CharacterEval (Tu et al., 2024), and TimeChara (Ahn et al., 2024), which were constructed by extracting character dialogue from novels and movies. Another dialogue corpus involving the role-play of historical figures is Character-LLM (Shao et al., 2023), which generates scenes and dialogues based on Wikipedia profile information.

However, these corpora rely on external or preexisting knowledge about the characters' personas and are often limited to well-known works, extracting dialogues directly from them. For characters from popular works, personas can be inferred from external sources such as Wikipedia or assumed based on the model's parameter size. Some datasets, such as the Harry Potter Dialogue Dataset (Chen et al., 2023), include information on relationships with other characters but are restricted to a few major works.

There is a need for dialogue corpora containing characters from minor works to better assess the role-playing capabilities of large language models (LLMs) such as GPT-4, which has vast parameters and extensive training data. Current corpora mainly construct personas using data available on the Web, evaluating role-playing by comparing the model's output to the expected persona. However, such an approach may not accurately assess LLMs trained on comparable sources.

To address this gap, I focus on collecting character dialogues from novels and deriving personas directly from narrative texts and character utterances. My corpus includes not only major works but also minor ones lacking Wikipedia coverage. Persona extraction from novels allows for more authentic character representation as described by the original authors. While I manually acquired utterances and personas at first, ongoing research explores methods for automating this process using LLMs, facilitating corpus expansion. In the future, I aim to use the corpus constructed by my method to evaluate the role-playing performance of a spoken dialogue system. Since there is no definitive "correct" voice for a character in a novel, I am interested in determining the type of voice the system should select to ensure that users perceive it as matching the character's persona.

1.2 Evaluating Responses of Persona Dialogue Systems

In the assessment of the response performance of a persona dialogue system, criteria such as naturalness and fluency are important, similar to those used in open-domain dialogue systems. However, one vital evaluation pertains to whether the responses align with the designated persona.

Several evaluation methods exist for how well personas are reflected in responses. These methods use persona descriptions (Jiang et al., 2020; Zheng et al., 2020), sample monologues (Su et al., 2019; Wu et al., 2020), and evaluations without references (Miyazaki et al., 2021) and involve LLM assessments (Shao et al., 2023; Wang et al., 2024).

The first three types of methods primarily assess individual responses, which may overlook nuances where responses are contextually incongruent with the persona. For example, if a user with a persona stating "I live with

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 73–75
September 16–17, 2024.

my family" asks the system, "Do you live alone?" and the system replies, "Yes," although "Yes" alone does not contradict the persona, in context it implies that the system lives alone, which contradicts the persona.

LLM-based methods involve feeding the model persona information and calculating scores to determine if responses align with the persona. For instance, Character-LLM (Shao et al., 2023), generates prompts based on dialogue history and persona traits to evaluate memorization, values, personality, hallucination, and stability criteria.

However, a significant issue with this method is that the correlation between LLM evaluation and manual evaluation has not been consistently explored. InCharacter (Wang et al., 2024) evaluates the performance of persona dialogue systems using psychological scales focusing on personality and has confirmed a correlation with human evaluations. Nonetheless, the assessment of role-playing performance should consider factors beyond the personality reflected in responses. Aspects such as speaking style and the fidelity of character memories may also need to be correlated with human evaluations.

Human evaluation also has its drawbacks. The first is that the evaluation results vary depending on the evaluator's subjectivity and preferences. The other is that, depending on the popularity of the work, it may be difficult to recruit evaluators who know all the information about the characters (Chen et al., 2024). To address these issues, it is possible to have evaluators learn the evaluation rules and character information, but this would be a very complicated process.

Furthermore, research has indicated that GPT-4 tends to give higher ratings to text generated by the same model (Jiang et al., 2020). Typically, researchers use the best-performing model for dialogue systems and response evaluation. Consequently, when GPT-4 evaluates responses generated by itself, there arises a risk of inaccurate evaluation. Hence, it may be necessary to assess persona dialogue systems using a model other than GPT-4.

I am developing a model that takes both dialogue context and responses as input to determine whether the response aligns with the persona. To train the model, I have built a dataset consisting of pairs of responses that align the persona and those that do not. The responses that align with the persona are extracted directly from a novel, while the non-aligning responses are generated using a LLM based on the former. The dialogue context leading up to each response is also generated using an LLM. The goal is to fine-tune smaller language models so that they can provide evaluations highly correlated with human judgments.

2 Spoken dialogue system (SDS) research

To realize a persona dialogue system in a voice dialogue system, it is important to reflect the persona not only in the speech content but also in tone of voice and emotional expression. Depending on the persona, reflecting dialects and accents in the voice may also be necessary. Studies are already being conducted on changing tone and emotion during speech synthesis. With recent advances in multimodal language models, I believe it will be possible to synthesize speech that suits any persona. However, when setting up speech synthesis from a text-based persona, preventing social bias is important.

Regarding the reflection of dialects and accents, research is being conducted in speech synthesis and text translation for languages and dialects with some level of resources. Studies are also ongoing with respect to low-resource languages to overcome limited resources. In the future, this will allow a spoken dialogue agent to reproduce the dialect of any persona, essentially from any region. However, in cases where the person is from a fictional region where no model exists or is an alien, the methods proposed so far may not address the situation.

Despite some challenges, I believe that realizing a persona dialogue system in a spoken dialogue system (SDS) is a promising endeavor. This will allow for the creation of a more humanlike SDSs and is expected to further deepen the relation between dialogue systems and humans.

3 Suggested topics for discussion

I suggest discussing the following topics:

- When incorporating a persona into an SDS, what content should be considered for speech synthesis?
- Will the evolution of multimodal LLM lead to an SDS that can manage all tasks with a single model?
- What additional features would make a SDS feel more human?

References

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. Timechara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of ACL*.

Nuo Chen, Yang Deng, and Jia Li. 2024. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484* .

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset

for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 8506–8520. https://doi.org/10.18653/v1/2023.findings-emnlp.570.

Ryuichiro Higashinaka, Masahiro Mizukami, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. 2018. Role play-based question-answering by real users for building chatbots with consistent personalities. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Melbourne, Australia, pages 264–272. https://doi.org/10.18653/v1/W18-5031.

Bin Jiang, Wanyue Zhou, Jingxu Yang, Chao Yang, Shihan Wang, and Liang Pang. 2020. PEDNet: A persona enhanced dual alternating learning network for conversational response generation. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), pages 4089–4099. https://doi.org/10.18653/v1/2020.coling-main.361.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597* .

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 994–1003. https://doi.org/10.18653/v1/P16-1094.

Chiaki Miyazaki, Saya Kanno, Makoto Yoda, Junya Ono, and Hiromi Wakaki. 2021. Fundamental exploration of evaluation metrics for persona characteristics of text utterances. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Singapore and Online, pages 178–189. https://doi.org/10.18653/v1/2021.sigdial-1.19.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 13153–13187. https://aclanthology.org/2023.emnlp-main.814.

Feng-Guang Su, Aliyah R. Hsu, Yi-Lin Tuan, and Hung-Yi Lee. 2019. Personalized Dialogue Response Generation Learned from Monologues. In *Proc. Interspeech 2019*. pages 4160–4164. https://doi.org/10.21437/Interspeech.2019-1696.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275* .

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *arXiv preprint arXiv:2310.17976* .

Bowen Wu, MengYuan Li, Zongsheng Wang, Yifu Chen, Derek F. Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2020. Guiding variational response generator to exploit persona. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 53–65. https://doi.org/10.18653/v1/2020.acl-main.7.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, pages 9693–9700. https://doi.org/10.1609/AAAI.V34I05.6518.

Biographical sketch

Ryuichi Ueahra is a master's student at the Graduate School of Informatics and Engineering, University of Electro-Communications. He is interested in persona-aware dialogue systems and role-playing agents. He participated in several competitions on building dialogue systems, including Dialogue System Live Competition 6 and AIWolfDial2024jp. He is supervised by Assoc. Prof. Michimasa Inaba.

Kenta Yamamoto

SANKEN, Osaka University
Ibaraki, Osaka
Japan

kentayamamoto@sanken.osaka-u.ac.jp

1 Research interests

The author is interested in building dialogue systems with **character** and **user adaptaton**. The goal is to create a dialogue system capable of establishing deeper relationships with users. To build a trustful relationship with users, it is important for the system to express its character. The author particularly aims to convey the system's character through multimodal behavior.

Although users currently try to speak clearly to avoid speech recognition errors when interacting with SDSs, it is necessary to develop SDSs that allow users to converse naturally, as if they were speaking with a human. The author focused on user adaptation by considering user personality and proposed a system that adjusts its manner of speaking according to the user's personality. In addition, the author is interested not only in adjusting the system's speaking style to match the user but also in making **the system's listening style** more conducive to natural conversation.

1.1 Character expression for SDSs

The character expression of robots leads to increasing user engagement in dialogue. In this work, "personality" is used as a psychological dimension for classifying users, and "character" notes the impression that the robot gives to the user. Generally, the character (personality) of a system is set based on enumerated personas or the Big Five personality traits. However, the characteristics of a personality that are easier to convey to users will differ depending on the modality. Therefore, the author focused on features in spoken dialogue such as backchannels, fillers, and switching pause length, and constructed a character expression model for a spoken dialogue system (Yamamoto et al., 2022). For example, an extroverted system is programmed to give frequent backchannels, while an emotionally unstable system is controlled to use more fillers. The results of dialogue experiments showed that the system was able to give the impression of executing its role more appropriately in the dialogue by expressing a character according to the task.

1.2 User adaptation based on user personality

The goal with user adaptation of SDSs is to make the systems generate behaviors appropriate to the user, which leads to increasing user satisfaction in the dialogue. There are several approaches when it comes to achieving dialogue suitable for the user such as selecting topics of interest to the user, synchronizing with the user's behavior, and predicting the user's internal state to facilitate dialogue. The author is interested in a system's manner of speaking that makes the user feel comfortable talking. In other words, it should evoke an impression of "getting along well" or "feeling at ease."

The author has previously demonstrated that having an SDS express a character that matches the user's personality can enhance user satisfaction with the dialogue (Yamamoto et al., 2023). In this earlier work, an "extroverted system" and "introverted system" were constructed using the techniques explained in Section 1.1, and the author analyzed the tendency of users to prefer interacting with each system based on the user's personality.

1.3 Control of system behaviors as a listener

There are have been many studies on the operation of spoken dialogue robots, but for users to have a pleasant dialogue, it is also necessary for the robot to behave appropriately as a listener. In the past, the author has been involved in the development of a listening dialogue system Inoue et al. (2020) that ensures users can speak comfortably by appropriately utilizing responses such as backchannels, evaluative feedback, and questions. On the other hand, humans perform various actions and reactions while listening, such as nodding and giving backchannel responses. By appropriately modeling such behaviors, the author aims to adequately express the sense that the agent is actively listening during spoken dialogues.

2 Spoken dialogue system (SDS) research

The author discusses the important topics for future studies on SDSs.

2.1 User adaptation in first-time interactions

Just as it is implausible for humans to like every single person they meet, there is no dialogue system that can be

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 76–77

liked by all people. Therefore, a dialogue system needs to understand the characteristics of the conversation partner from the dialogue and adapt accordingly. However, such user adaptation is typically assumed for dialogue systems used by the same user over a long period.

Conversely, systems designed for first-time interactions, such as with store clerks, should be designed to speak in a manner suitable for the role (satisfying many users). User adaptation based on persona information, such as individual user preferences, makes it difficult to handle first-time interactions.

However, the author believes that it is possible to achieve user-appropriate dialogues even in first-time interactions. This can be done by recognizing the user's personality at the beginning of the conversation and then adjusting the speech style to match it during the conversation. Although human personalities vary, it is believed that preparing several personality groups for specific situations can handle these variations. Another advantage of this approach is that, unlike when using personas, the system does not explicitly communicate the recognized result of the user's personality to the user. Therefore, even if the recognition result is incorrect, the conversation itself can avoid collapsing.

2.2 Understanding human relationships

Two types of human relationships are discussed here: first, the relationship between the user in front of the system and the system itself, and second, the relationships between users. When implementing user adaptation in SDSs, it is necessary to model the relationship between the user and the system. In other words, it is essential to constantly monitor how much trust the user and the system have built. This is because, in scenarios where the user interacts with the system over an extended period, the manner and content of the conversation will change. This change cannot simply be measured by the length of the interaction time because it switches in accordance to changes in the relationship with the user. Therefore, it is crucial to model the relationship between the user and the system and conduct the dialogue accordingly.

Currently, it is sufficient to continue the dialogue by considering only the relationship with the person in front of the system in a one-on-one interaction. Indeed, most of the datasets collected for learning purposes assume one-on-one dialogues. However, when SDSs or robots are used in society, it becomes necessary to conduct dialogues that consider multi-person interaction scenarios and relationships with people not present in the conversation.

In such cases, it is necessary to make utterances that consider the relationships between users. However, there is a lack of data and methods for constructing dialogue systems that handle multi-person interactions or are utilized by multiple users. For example, in multi-person dialogues, predicting the next speaker can vary depending on the relationships between users. It is thus necessary to collect multi-party dialogue datasets in various everyday situations to build models that capture the relationships between users.

3 Suggested topics for discussion

The author suggests three topics for discussion in the discussion panel during the event.

- Is it necessary for spoken dialogue systems to possess human-like dialogue? Are there more appropriate methods for dialogue with dialogue systems?
- How can insights gained from theoretical studies of dialogue, such as conversation analysis, be incorporated into our studies?
- Does a SDS's agreement fulfill the user's need for approval? Can it serve as a substitute for human friends?

Acknowledgement

This work was supported by Grant-in-Aids for JSPS (23K20005, 24K20839).

References

Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In *SIGDIAL*. pages 118–127.

Kenta Yamamoto, Koji Inoue, and Tatsuya Kawahara. 2022. Character expression for spoken dialogue systems with semi-supervised learning using variational auto-encoder (79):101469–101469.

Kenta Yamamoto, Koji Inoue, and Tatsuya Kawahara. 2023. Character adaptation of spoken dialogue systems based on user personalities. In *IWSDS*.

Biographical sketch

Kenta Yamamoto received his Ph.D. in 2023 from the Graduate School of Informatics in Kyoto University, Kyoto, Japan. He was a JSPS Research Fellow (DC1) from 2020 to 2023. Currently, he is an Assistant Professor of SANKEN (The Institute of Scientific and Industrial Research), Osaka University. His research interests include spoken dialogue systems (SDSs) and character expression for SDSs.

Isabel Feustel

Ulm University
Institute of Communications Engineering
Albert-Einstein-Allee 43
89081 Ulm

isabel.feustel@uni-ulm.de
www.uni-ulm.de/nt/staff/
research-assistants/isabel-feustel/

1 Research interests

My research is focused on the field of **explainable AI (XAI)**, which aims to address the challenge of providing transparency to AI systems. I am particularly focused on the development of dialogue systems that enable **natural interaction with explanations**. By employing **computational argumentation** approaches, my objective is to create methods that facilitate meaningful dialogue between users and AI systems, allowing for a greater understanding of the systems' reasoning processes.

1.1 Enabling XAI explanations through dialogue

In recent years, the need for transparency in AI systems has significantly increased, leading to the growing popularity of the field of explainable AI (XAI) (Das and Rad, 2020). Ensuring that AI systems are understandable to users is crucial for building trust and facilitating effective use (Schmidt et al., 2020). One promising approach to achieving this is through dialogue systems, which can enable more dynamic and interactive explanations (Sokol and Flach, 2020).

Dialogue systems offer several advantages for the provision of explanations. These include the ability to segment information into manageable parts, thereby facilitating the comprehension of complex concepts; the capacity to elicit questions based on the specific needs of the user, which results in a more personalized and relevant interaction; and the capability to adapt the system's responses to align with the user's knowledge level and language proficiency, which enhances comprehension and satisfaction.

However, many existing XAI methods are non-conversational, offering explanations that are challenging for non-expert users to comprehend. Current conversational approaches in XAI like Slack et al. (2023), Shen et al. (2023) or Feldhus et al. (2023) often rely on basic question-answering systems and lack sophisticated dialogue management capabilities. This limitation neglects the importance of context in maintaining coherent and meaningful interactions. In order to address these issues, we proposed a generic dialogue architecture that integrates XAI explanations into a dialogue system (Feustel et al., 2023). Subsequently, we implemented a prototype based on this architecture.

Recognizing that effective explanations often require more than just model-specific details, we incorporated a knowledge module containing domain-specific information. This module is essential for providing comprehensive reasoning about the AI's domain, thereby facilitating a more profound comprehension of the foundation of the underlying process.

1.2 Integrating Domain Knowledge

The incorporation of domain expertise prompted the need to ascertain an effective methodology for integrating this knowledge into a dialogue system and establishing a connection with XAI explanations. The proximity of the areas of argumentation and XAI presents an opportunity for exploration, as arguments and explanations share comparable characteristics (Vassiliades et al., 2021). We determined that computational argumentation offers a suitable framework for representing domain facts, as it allows for structured and logical presentations of information.

Utilizing our expertise in argumentative dialogue, we determined that argumentative tree structures could be readily adapted to effectively address this integration challenge (Feustel et al., 2024). We extended our prototype system to include domain specific arguments and conducted a small study to evaluate the system's effectiveness. The results indicated positive trends, suggesting that integrating domain knowledge into the dialogue system has a positive effect on the dialogue.

1.3 Future Directions

In future research we want to explore several key areas to enhance the capabilities of the explanatory dialogue systems.

Firstly, we aim to improve the Natural Language Understanding (NLU) to achieve a more generic understanding of explanation requests, as we observed a high error rate in the current system that was NLU-related, resulting in users not being understood correctly. This involves developing advanced models capable of accurately interpreting and processing a wide range of user queries, re-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 78–80
September 16–17, 2024.

gardless of the specific wording or context.

Additionally, we plan to advance Natural Language Generation (NLG) techniques. Currently, our and other XAI systems rely on template-based system responses, which can result in rigid responses. By exploring more sophisticated NLG methods, such as those powered by large language models, we aim to generate more fluid and contextually appropriate responses. This improvement would also include the ability to paraphrase arguments to better fit the dialogue context, thereby enhancing the coherence and relevance of the information provided to users.

Another important area of focus is the annotation of arguments to enable better selection for specific user requests. By refining the way arguments are annotated and categorized, dialogue systems can more effectively retrieve and present the most pertinent information based on the user's needs. This involves developing detailed and nuanced annotation schemas that capture the essential qualities of arguments, ensuring that the system can make informed decisions about which arguments to present in various contexts.

By focusing on these improvements, we posit that significant advancements can be made towards more sophisticated, transparent, and user-centric dialogue systems.

2 Spoken dialogue system (SDS) research

I believe that in the next 5 to 10 years, the field of dialogue research is expected to see significant advancements in creating more flexible and natural dialogue systems. These systems will be capable of adapting to individual user styles, making interactions more personalized and effective. We will also see the emergence of multilingual and culturally adaptable systems, which can truly focus on users from diverse backgrounds. This will foster global communication and accessibility. Moreover, there will be renewed discussions on human-like systems, exploring the ethical and social implications of developing systems that closely mimics human behavior.

With the integration of large language models (LLMs), there may be a fundamental rethinking of traditional dialogue system frameworks, leading to more intuitive and seamless conversational experiences. We need to think about how LLMs can be integrated into traditional dialogue system architectures to leverage their full potential. However, we also need to be aware of the limitations they bring, such as biases in training data and the potential for generating misleading or inappropriate content.

Additionally, I see a future with more open domain dialogues, allowing users to engage in a wider variety of topics without the constraints of pre-defined domains. I think these open domain applications might function as microservices, where a single speech interface processes the intent and directs the user to the appropriate application to fulfill their request. Virtual agents will increase in prevalence, necessitating a high need for natural speech interaction to ensure user satisfaction and effectiveness across various tasks and applications. Improved assistant systems will further support users in various tasks, from simple queries to complex problem-solving, enhancing productivity and user satisfaction across different applications.

3 Suggested topics for discussion

- **Personalization and User Modelling:** Best practises for tailoring dialogue to individual users. What can be personalized and what should not be personalized? Which aspects of a user can already be modelled and how can we model more complex aspects? E.g. Mental Model
- **Evaluation of Dialogue:** How can we evaluate non-task-oriented dialogues? How can we engage participants to interact with the system without influencing the results?
- **Error-Communication:** There are various aspects where a (dialogue) system can fail (e.g. wrong AI prediction, wrong intent classification, ..). Can we somehow track these failures? How should systems react if the users notices some wrong behavior? Can we implement feedback loops to optimize the dialogue policy?

References

Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* .

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023. InterroLang: Exploring NLP models and datasets through dialogue-based explanations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 5399–5421. https://doi.org/10.18653/v1/2023.findings-emnlp.359.

Isabel Feustel, Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2023. Towards interactive explanations of machine learning methods through dialogue systems. *The 13th International Workshop on Spoken Dialogue Systems Technolog* .

Isabel Feustel, Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2024. Enhancing model transparency: A dialogue system approach to xai with domain knowledge. *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue* .

Philipp Schmidt, Felix Biessmann, and Timm Teubner. 2020. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 29(4):260–278.

Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023. ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-AI scientific writing. In *Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, New York, NY, USA, CSCW '23 Companion, page 384–387. https://doi.org/10.1145/3584931.3607492.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence* https://doi.org/10.1038/s42256-023-00692-8.

Kacper Sokol and Peter Flach. 2020. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz* 34(2):235–250.

Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review* 36:e5.

Biographical sketch

Isabel Feustel is a PhD student at Ulm University supervised by Dr. Dr. Wolfgang Minker and Dr. Stefan Ultes. She obtained a Master's Degree in Media Informatics at Ulm University in 2019, where her research focused on communication styles within dialogues. She began her PhD by exploring argumentative dialogues and focused her studies on explanatory dialogues.

Shutong Feng

Heinrich Heine University Düsseldorf
Universitätsstraße 1
40225 Düsseldorf
Germany

shutong.feng@hhu.de
https://shutongfeng.github.io/

1 Research interests

My research interests lie in the area of **modelling affective behaviours of interlocutors in conversations**. In particular, I look at emotion perception, expression, and management in information-retrieval task-oriented dialogue (ToD) systems. Traditionally, ToD systems focus primarily on fulfilling the user's goal by requesting and providing appropriate information. Yet, in real life, the user's emotional experience also contributes to the overall satisfaction. This requires the system's ability to recognise, manage, and express emotions. To this end, I incorporated emotion in the entire ToD system pipeline (Feng et al., 2024). In addition, in the era of large language models (LLMs), emotion recognition and generation have been made easy even under a zero-shot set-up (Feng et al., 2023b; Stricker and Paroubek, 2024). Therefore, I am also interested in building ToD systems with LLMs and examining various types of affect in other ToD set-ups such as depression detection in clinical consultations and user confidence estimation in tutoring systems (Litman et al., 2009).

1.1 Emotion-aware ToD System

While existing works have explored user emotions or similar concepts in various ToD modelling tasks (Lukin et al., 2017; Guo et al., 2024), none has so far combined these emotional aspects into a fully-fledged dialogue system nor conducted interaction with human or simulated users. Therefore, I propose to incorporate emotion into the complete ToD interaction process, involving understanding, management, and generation.

To achieve this, I first extended the EmoWOZ dataset (Feng et al., 2022) with system emotion labels. With this ToD dataset containing both user and system emotion labels, I could train a both emotionally and semantically conditioned natural language generator, as well as an emotional user simulator (Lin et al., 2023) that both reacts to system emotion and expresses user emotions. Leveraging off-the-shelf dialogue state tracker (van Niekerk et al., 2021) and user emotion recogniser (Feng et al., 2023a), I set up the system around a dialogue policy (Geishauser et al., 2022), which takes dialogue state extended with user emotion as input and outputs action including system emotions. The policy was optimised via reinforcement learning (RL) with the emotional user simulator on the language level. For the reward signal, the policy considered both task success and user sentiment level.

In addition to the above-mentioned modular ToD system, I also took the inspiration from an existing LLM-based end-to-end system (Stricker and Paroubek, 2024). I extended the system to output emotional actions and trained it with the newly collected dataset.

With both systems, I conducted corpus-level evaluation and interactive evaluation with both simulated and real users. Our results show that incorporating emotion into the full ToD pipeline can effectively enhance the user's emotional experience and task success at the same time. This aligns with our hypothesis and intuition that emotion is crucial in ToD systems. I believe this points to a promising direction on improving ToD systems.

The future work would be to combine the advantages of modular systems and end-to-end systems, specifically by incorporating RL with human feedback (RLHF) to LLM-based end-to-end systems. Modular systems are usually centred around a dialogue policy optimised via RL for long-term task success. Yet, they are prone to errors from each small modules. End-to-end models, on the other hand, can leverage the capacity of large pretrained models but existing models are trained on the corpus with supervised learning. This usually leads to suboptimal performance in interactive evaluation. Incorporating RLHF in the training could potentially be a solution and further boost the performance of end-to-end ToD systems. Efficient acquirement of response preference labels and RL training will be my next research efforts.

1.2 Recognising Affect using LLMs

I am also interested in how LLMs can be used to recognise user affects in conversations. My goal was not to build state-of-the-art affect recognition models with LLMs but rather to understand the potential of current LLMs under vanilla set-ups for such a purpose. Specifically, I conducted experiments with a set of LLMs on different types of datasets under an array of prompt-based training set-ups. For datasets, I examined three differ-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 81–83
September 16–17, 2024.

ent types of affects: emotions in ToDs, emotions in chit-chat, and depression. For training set-ups, I looked at zero-shot learning, few-shot in-context learning, and supervised learning with different amount of data. I also considered LLMs as a text-processing back-end in SDS by investigating how automatic speech recognition errors could influence model prediction. With experimental results, I draw insights on LLMs' zero and few-shot ICL ability, data efficiency in task-specific fine-tuning, ability to handle long input sequence, ability to recognise different types of affects, robustness to ASR errors, and so on.

In the future, I will look at how affect recognition and generation can be improved under zero or few-shot set-ups. I will leverage existing resources such as annotator confusion and annotation schemes to elicit reliable reasoning and uncertainty estimation in LLMs.

2 Spoken dialogue system (SDS) research

The emergence of LLMs has great impact on approaches in spoken dialogue modelling. They also bring about opportunities in areas such as unsupervised ontology construction for system design (Vukovic et al., 2024). While LLMs have demonstrated promising abilities in general language modelling tasks and chat applications, smaller models and established modular system set-ups should not be overlooked. Therefore, instead of wishfully using LLMs to replace all SDSs, researchers will understand more about the limitations of LLMs so as to combine the strengths of LLMs and traditional methods.

There will also be more diverse requirements and evaluation criteria for SDSs. In the past, information-retrieval ToD systems focus primarily on task success and inform rate, and chit-chat systems focus on engagement, coherence, and naturalness. As we see more about what more powerful systems can achieve nowadays, we expect more from the system: safety, trust-worthiness, bias, emotion consistency, and many more. We may also expect our dialogue agents to be able to adapt to different challenging scenarios, from out-of-domain requests to cultural shifts. While we see more exciting research opportunities and directions, challenges such as the evaluation of more well-rounded SDSs emerge.

3 Suggested topics for discussion

- **Controllability of LLMs as Dialogue System Back-end:** The issue of hallucination can be especially detrimental in the domain of task-oriented dialogues and in the presence of an ontology and database. How should we make LLMs more controllable for SDS applications?

- **The Future of LLMs**: What ability would the next generation of LLMs have? What would be possible directions of the development in NLP?

- **Affective SDS:** What are risks of building SDSs for affect-related applications, such as emotion support, mental health counseling, more human-like personal assistant, etc.?

References

Shutong Feng, Hsien chin Lin, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2024. Infusing emotions into task-oriented dialogue systems: Understanding, management, and generation. https://arxiv.org/abs/2408.02417.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 4096–4113. https://aclanthology.org/2022.lrec-1.436.

Shutong Feng, Nurul Lubis, Benjamin Ruppik, Christian Geishauser, Michael Heck, Hsien-chin Lin, Carel van Niekerk, Renato Vukovic, and Milica Gasic. 2023a. From chatter to matter: Addressing critical steps of emotion recognition learning in task-oriented dialogue. In Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani, editors, *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czechia, pages 85–103. https://doi.org/10.18653/v1/2023.sigdial-1.8.

Shutong Feng, Guangzhi Sun, Nurul Lubis, Chao Zhang, and Milica Gašić. 2023b. Affect recognition in conversations using large language models. https://arxiv.org/abs/2309.12881.

Christian Geishauser, Carel van Niekerk, Hsien-chin Lin, Nurul Lubis, Michael Heck, Shutong Feng, and Milica Gašić. 2022. Dynamic dialogue policy for continual reinforcement learning. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus,

Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pages 266–284. https://aclanthology.org/2022.coling-1.21.

Ao Guo, Ryu Hirai, Atsumoto Ohashi, Yuya Chiba, Yuiko Tsunomori, and Ryuichiro Higashinaka. 2024. Personality prediction from task-oriented and open-domain human–machine dialogues. *Scientific Reports* 14(1):3868.

Hsien-Chin Lin, Shutong Feng, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2023. EmoUS: Simulating user emotions in task-oriented dialogues. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, SIGIR '23, page 2526–2531. https://doi.org/10.1145/3539618.3592092.

Diane J. Litman, Mihai Rotaru, and Greg Nicholas. 2009. Classifying turn-level uncertainty using word-level prosody. In *Interspeech*. https://api.semanticscholar.org/CorpusID:279660.

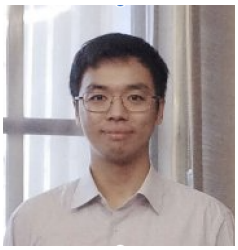
Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. pages 742–753.

Armand Stricker and Patrick Paroubek. 2024. A Unified Approach to Emotion Detection and Task-Oriented Dialogue Modeling. In *IWSDS*. Sapporo (Japon), Japan. https://hal.science/hal-04415809.

Carel van Niekerk, Andrey Malinin, Christian Geishauser, Michael Heck, Hsien-chin Lin, Nurul Lubis, Shutong Feng, and Milica Gasic. 2021. Uncertainty measures in neural belief tracking and the effects on dialogue policy performance. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 7901–7914. https://doi.org/10.18653/v1/2021.emnlp-main.623.

Renato Vukovic, David Arps, Carel van Niekerk, Benjamin Matthias Ruppik, Hsien-Chin Lin, Michael Heck, and Milica Gašić. 2024. Dialogue ontology relation extraction via constrained chain-of-thought decoding. https://arxiv.org/abs/2408.02361.

Biographical sketch



Shutong Feng is a final-year PhD student at the Chair for Dialog System and Machine Learning, Heinrich Heine University Düsseldorf, Germany. He is supervised by Prof. Dr. Milica Gašić and co-supervised by Dr. Nurul Lubis. He is interested in modelling human affect in spoken dialogue systems. Shutong obtained his BA and MEng degrees from the University of Cambridge in 2019. He then worked as an engineer at Huawei Technologies Co. Ltd. before starting his PhD study in 2020.

Zhiyang Qi

The University of Electro-Communications
Chofu, Tokyo
Japan
qizhiyang@uec.ac.jp
https://zhiyangqi.github.io/

1 Research interests

In the modern field of Natural Language Processing (NLP), large language models (LLMs), such as GPT-4 (OpenAI, 2023), have become the key technologies that potentially break the traditional boundaries. These models can generate idiomatic high-quality text, successfully addressing many of the NLP challenges and drive rapid technological advancements. Within the context of LLMs, my research interests are: (1) utilizing the powerful text generation capabilities of the LLMs in terms of **customized dialogue data augmentation** in data-scarce tasks, and (2) applying the LLMs to the **psychological counseling dialogues**. Moreover, I hope to combine these two themes in the future.

1.1 Customized dialogue data augmentation

Spoken dialogue systems (SDSs) often rely on the interaction data between real humans for training. However, different people have different speaking styles and strategies, influenced by factors such as the dialogue topic, age, regional and local language variation, context, identity, preferences, and personality of the speaker, among others. In real life human conversations, individuals may adjust their responses based on the other party's strategy, such as seeking clarification when the other party speaks unclearly. For SDS, those with unique dialogue strategies form a minority group, resulting in relatively scarce dialogue data. Consequently, the SDS cannot adapt to the speaking strategies of others as effectively as humans, particularly when encountering individuals with unique speaking styles.

The scarcity of the annotated data and the challenge of data imbalance are persistent issues in various artificial intelligence domains (Shi et al., 2020; Ahmad et al., 2021; Hedderich et al., 2021). To address those effectively, various data augmentation techniques have been employed, as demonstrated in prior research on different tasks (Feng et al., 2021; Bayer et al., 2022; Kim et al., 2023). For instance, Schick and Schütze (2021) generated text similarity datasets from scratch by instructing a large pre-trained language model (PLM). Similarly, Liu et al. (2022) and Chen and Yang (2021) enhanced the data by manipulating individual utterances within dialogues—in ways such as adding, deleting, changing their order, or regenerating them—while preserving the original meaning, which improved the model's performance in the dialogue summarization tasks.

My research focuses on the dialogues that involve users of different age groups. Inaba et al. (2024a) have found that speakers of various age group exhibit distinct speaking strategies. For example, compared to other age groups, minor interlocutors are less likely to express their opinions. Consequently, the other speaker often seeks confirmation or asks additional questions to make the conversation flowing smoothly. Considering the unique speaking styles of minors and the inherent difficulties in obtaining data from them (Aydin et al., 2021), my recent research employs a framework that combine the LLM and PLM. This approach customizes the generation of dialogue data for minors, enhancing the performance of SDS in situations when data from minors is scarce.

1.2 Psychological counseling using LLM

Mental health is one of the critical issues in today's society. According to the World Health Organization (WHO), nearly 1 billion people worldwide suffer from mental disorders, yet 70% of them do not receive any treatment, such as counseling[1]. There is a significant gap between the existing mental health support and the needs of patients. In recent years, the emergence of online counseling platforms, such as 7cups[2] has made psychological counseling more accessible. However, due to the lack of experience of some counselors, the effectiveness of these services is not always ideal. Additionally, training professional counselors requires considerable effort.

In recent years, AI research related to psychological counseling has been increasing. Inaba et al. (2024b) collected counseling dialogue data using role-playing methods, and the evaluations by professional counselors indicated that the responses generated by GPT-4 were competitive compared to those generated by human counselors. Zhang et al. (2024) enriched the counseling dialogue dataset by using LLM to generate dialogues based on reports from online counseling platforms. Young et al. (2024) investigated the popularity of human and LLM-generated responses across various counseling top-

[1] https://news.un.org/zh/story/2022/06/1104712
[2] https://www.7cups.com/

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 84–86

ics. Their results showed that LLM responses were more popular for topics like interpersonal relationships and physical health, while human responses were preferred for topics related to suicide.

Those studies indicate that LLMs can play the role of counselors, generating high-quality psychological counseling dialogues. However, due to the uncontrollable nature of their generated content, there is a potential risk when interacting with users who have suicidal tendencies or extreme emotions. Consequently, the aim of related research is not to have AI act as counselors directly but to use their powerful text generation capabilities to assist counselors with dialogues. Sharma et al. (2022) developed HAILEY using PLM to help peer supporters on online counseling platforms provide more empathetic responses. Similarly, Hsu et al. (2023) used PLMs to offer real-time response strategies and sentences during counseling dialogues, assisting counselors in their work. This approach mitigates safety and ethical risks while also helping inexperienced counselors develop their professional skills.

My research interest lies in utilizing LLMs to assist counselors with psychological counseling dialogues. Specifically, this study employs LLM to provide various forms of real-time support for the mental health counselors during their sessions with their patients, in terms of dialogue strategies, example responses, and refinement of drafted replies. Ultimately, the usefulness of the support system and the most preferred type of support by counselors will be analyzed through a questionnaire survey.

2 Spoken dialogue system (SDS) research

I believe that future SDSs need to have the ability to adapt to different individuals. For example, people's personalities vary; some enjoy engaging in conversation, while others are better listeners and appreciate different aspects of the dialogue. Additionally, some people are comfortable answering any questions, while others may be more restrained and prefer not to be asked very personal questions. The goal is for SDS to infer the users' personalities through various potential multimodal cues during conversations and adapt their responses accordingly. This adaptability would significantly enhance the evaluation of dialogue systems.

I also hope that SDSs will become increasingly active in the field of psychological counseling. The number of people suffering from psychological problems is enormous, and most of them do not receive adequate support due to a lack of someone to talk to, among other reasons. This situation needs improvement. The powerful capabilities of LLMs can provide significant help in psychological counseling.

Ultimately, applying the user adaptability to psychological counseling will enable SDSs to create more flexible and effective counseling dialogues when interacting with different users.

3 Suggested topics for discussion

I suggest discussing the following topics:

- **Multimodal Dialogue Systems for Individuals with Disabilities**: As multimodal dialogue systems evolve, more information becomes available for dialogue generation. Can we leverage these technologies to facilitate daily life activities for individuals with disabilities? What are the key technologies when building such dialogue systems, and what considerations should be made?

- **LLM's Personality Adaptation**: Humans typically exhibit a single personality type, possibly engaging comfortably in conversations with only a few other personality types. In contrast, LLMs are trained on extensive textual data from conversations involving various personality types. Thus, LLMs can theoretically adapt to any personality, potentially enhancing the conversational experience for all of the users by adopting different conversational styles to match the user's personality.

- **How long can the trend of LLMs last? What are the key technologies for future SDS?**

References

Tanveer Ahmad, Dongdong Zhang, Chao Huang, Hongcai Zhang, Ningyi Dai, Yonghua Song, and Huanxin Chen. 2021. Artificial intelligence in sustainable energy industry: Status quo, challenges and opportunities. *Journal of Cleaner Production* 289:125834.

Selami Aydin, Leyla Harputlu, Özgehan Uştuk, Şeyda Savran Çelik, and Serhat Güzel. 2021. Difficulties in collecting data from children aged 7–12. *International Journal of Teacher Education and Professional Development (IJTEPD)* 4(1):89–101.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Comput. Surv.* 55(7). https://doi.org/10.1145/3544558.

Jiaao Chen and Diyi Yang. 2021. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 6605–6616. https://doi.org/10.18653/v1/2021.emnlp-main.530.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, pages 968–988. https://doi.org/10.18653/v1/2021.findings-acl.84.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, pages 2545–2568. https://doi.org/10.18653/v1/2021.naacl-main.201.

Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2023. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback.

Michimasa Inaba, Yuya Chiba, Zhiyang Qi, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao, and Takayuki Nagai. 2024a. Travel agency task dialogue corpus: A multimodal dataset with age-diverse speakers. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* https://doi.org/10.1145/3675166.

Michimasa Inaba, Mariko Ukiyo, and Keiko Takamizo. 2024b. Can large language models be used to provide psychological counselling? an analysis of gpt-4-generated responses using role-play dialogues. In *The 14th International Workshop on Spoken Dialogue Systems Technology*. https://arxiv.org/abs/2402.12738.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 12930–12949. https://doi.org/10.18653/v1/2023.emnlp-main.799.

Yongtai Liu, Joshua Maynez, Gonçalo Simões, and Shashi Narayan. 2022. Data augmentation for low-resource dialogue summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, United States, pages 703–710. https://doi.org/10.18653/v1/2022.findings-naacl.53.

OpenAI. 2023. Gpt-4 technical report.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 6943–6951. https://doi.org/10.18653/v1/2021.emnlp-main.555.

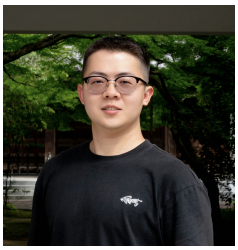
Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2022. Human-ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support.

Wenzhong Shi, Min Zhang, Rui Zhang, Shanxiong Chen, and Zhao Zhan. 2020. Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sensing* 12(10). https://doi.org/10.3390/rs12101688.

Jordyn Young, Laala M Jawara, Diep N Nguyen, Brian Daly, Jina Huh-Yoo, and Afsaneh Razi. 2024. The role of ai in peer support for young people: A study of preferences for human- and ai-generated responses. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, CHI '24. https://doi.org/10.1145/3613904.3642574.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling.

Biographical sketch



Zhiyang Qi is currently pursuing a PhD at The University of Electro-Communications in Japan, within the Graduate School of Informatics and Engineering. Associate Professor Michimasa Inaba is his advisor. Qi is interested in dialogue system competitions and dialogue systems for the Werewolf game. He enjoys playing board games, with his favorite being Catan.

Zi Haur Pang

Kyoto University, Kyoto, Japan
pang.haur.42a@st.kyoto-u.ac.jp
https://zihaurpang.github.io/

1 Research interests

The author's research advances human-AI interaction across two innovative domains to enhance the depth and authenticity of communication. Through **Emotional Validation**, which leverages psychotherapeutic techniques, the research enriches SDSs with advanced capabilities for understanding and responding to human emotions. On the other hand, while utilizing **Embodied Conversational Agents (ECAs)**, the author focuses on developing agents that simulate sophisticated human social behaviors, enhancing their ability to engage in context-sensitive and personalized dialogue. Together, these initiatives aim to transform SDSs and ECAs into empathetic, embodied companions, pushing the boundaries of conversational AI.

1.1 Emotional Validation in SDSs

Emotional expressiveness is pivotal in fostering relationships between humans and artificial intelligence within Spoken Dialogue Systems (SDSs). Traditional methods in SDSs focus on recognizing user emotions (Poria et al., 2019) or generating empathetic responses (Fu et al., 2023). However, these conventional approaches often fall short for individuals who suppress emotions due to stress or traumatic experiences. For instance, the mere recognition and mimicry of user emotions can be insufficient, and simplistic empathetic responses such as "I am so sorry to hear that" may not adequately address the users' deeper needs for emotional support.

At the core of our emotional well-being, as outlined by Maslow's hierarchy of needs (Gorman, 2010), lies the necessity for love, belongingness, and acceptance. This layer underscores the significance of interpersonal relationships and the inherent human desire to be valued and accepted by the community. It is within this context that conventional SDSs responses frequently fail to provide genuine emotional support. A more personalized approach, such as acknowledging a user's feelings with affirmations like "It is okay for you to feel this way," can significantly enhance the interaction by validating the user's emotional experience.

Motivated by the significant impact of emotional validation on user experiences, the author explored a psychotherapeutic communication technique known as *validation*, which involves recognizing, understanding, and acknowledging the emotional states, thoughts, and actions of others. This investigation led to an analysis of validating responses within a human-human emotional story spoken-dialogue corpus (Pang et al., 2023). Building on this, the author utilized the theory of levels of validation (Linehan, 1997) to develop a system capable of generating appropriate responses in attentive listening settings. This system has demonstrated its effectiveness in enhancing emotional expressiveness in both written and spoken dialogues (Pang et al., 2024).

By integrating these validation techniques into human-robot interaction systems, the author aims to meet the inherent human need for emotional support, thereby laying a foundation for trust and rapport through meaningful social dialogue. Ultimately, the author hopes to transform the SDSs from a simple interactive tool into a companion AI that, like a family member or friend, builds lasting relationships and fosters genuine rapport.

1.2 Social Embodied Conversational Agents (ECAs)

Social Embodied Conversational Agents (ECAs) form the core of the author's research endeavors, aimed at deepening the interaction between humans and SDSs in ways that closely mirror human social behaviors. This research encompasses a variety of ECAs—including autonomous androids, virtual agents, and teleoperated humanoid robots—each designed to simulate nuanced social interactions. The author focuses on developing these agents to incorporate sophisticated verbal exchanges as well as expressive non-verbal communication, such as facial expressions and body language, enhancing their ability to engage in lifelike social interactions. With the assistance of ECAs, interaction experiences can be significantly enriched through both verbal and non-verbal behaviors. For instance, virtual agents can display dynamic facial expressions onscreen, while physical robots can provide tangible interactions through touch and responsive gestures or body movements.

Another focus is on the capability of these social ECAs to adapt to everyday social environments. For example, an agent might initiate a light-hearted discussion about a common hobby or a shared interest observed in the user's environment, thereby fostering a more engaging and personalized interaction. This approach underscores the importance of context-aware communication in enhancing interaction quality and integrating these technologies more seamlessly into human social spheres.

The author's research is specifically aimed at investi-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 87–89
September 16–17, 2024.

gating how these interactions, particularly in providing emotional support during personal exchanges like problem or worry sharing, can enhance the user experience in social settings. This targeted exploration seeks to determine how ECAs can augment SDS to offer more natural and human-like experiences, focusing on acute emotional support and establishing prolonged social relationships. This nuanced approach aims to refine the integration of ECAs in scenarios where empathetic engagement is crucial, optimizing the balance between effective support and efficient interaction.

2 Spoken dialogue system (SDS) research

In the upcoming years, potential SDS research directions could include advancing the provision of deeper emotional support and examining the evolving social relationships between humans and SDSs.

2.1 Deeper Level of Emotional Support

To achieve a deeper level of emotional support, it is crucial to move beyond the current scope of response generation and emotion recognition. Present studies either generate responses or recognize emotions based on isolated utterances or entire dialogues (Jiao et al., 2019). However, these approaches fall short of offering true emotional support. Emotional states are not static; they fluctuate dynamically throughout a conversation. Therefore, current methods that treat emotions as static entities are inadequate for fulfilling the need for genuine emotional support.

Moreover, while current response generation methods can provide a range of supportive responses (Xie and Pu, 2021), they often lack the depth required for meaningful emotional assistance. For instance, when a user faces difficult times, generic empathetic responses such as "I am so sorry to hear that" might offer some comfort, but they are often insufficient. Users may seek more substantial support, such as encouragement or validation, which requires a strategic selection of responses tailored to the specific context and emotional state. Current response generation methods fail to address this need, as they focus on producing responses rather than strategically selecting the most appropriate form of support based on the situation.

Advancing research in this area necessitates an interdisciplinary approach, incorporating insights from psychology and social sciences alongside engineering and computational techniques. This broader perspective will enable the development of SDSs that can genuinely understand and respond to the dynamic emotional states of users, providing deeper and more meaningful emotional support.

2.2 Social Relationship Between Human and SDSs

With the advancement of large language models (LLMs), a variety of companion-based SDSs, such as Replika and Character.AI, are emerging in the public domain, prompting users to form diverse relationships with these AI entities. Studies have even shown that female users are beginning to develop romantic relationships with characters in *otome games* (female-oriented mobile games) (Gong and Huang, 2023). This phenomenon necessitates a reevaluation of the social relationship between humans and SDSs. Should these systems be designed to be more human-like to foster deeper rapport and connection, or should they be maintained as mere tools?

If we aim to establish more rapport-driven relationships with SDSs, it is imperative to consider the precautions needed to prevent any negative societal impacts. Conversely, if SDSs are to be treated solely as tools, we must find a balance between enhancing their human-like qualities and retaining their utility as functional assistants. Addressing these questions is essential as we navigate the evolving landscape of human-AI interaction.

Examining these dynamics requires a multidisciplinary approach, integrating insights from psychology, ethics, and technology. This comprehensive perspective will ensure that the development and deployment of SDSs promote positive outcomes and mitigate potential risks associated with their increasing human-like presence in users' lives.

3 Suggested topics for discussion

The author would like to propose the following topics for discussion.

- Should SDSs incorporate human negative traits to achieve a higher level of human-likeness?
- Should SDSs be designed to build rapport or even romantic relationships with humans?
- How should SDSs be designed to balance between providing support and avoiding emotional dependency?

Acknowledgments

This work was supported by KAKENHI (19H05691) and JST Moonshot R&D Goal 1 Avatar Symbiotic Society Project (JPMJMS2011). The author deeply appreciates the constructive comments and valuable feedback provided by the reviewers.

References

Yahui Fu, Koji Inoue, Chenhui Chu, and Tatsuya Kawahara. 2023. Reasoning before responding: integrating commonsense-based causality explanation

for empathetic response generation. *arXiv preprint arXiv:2308.00085* .

An-Di Gong and Yi-Ting Huang. 2023. Finding love in online games: Social interaction, parasocial phenomenon, and in-game purchase intention of female game players. *Computers in Human Behavior* 143:107681.

Don Gorman. 2010. Maslow's hierarchy and social and emotional wellbeing. *Aboriginal and Islander Health Worker Journal* 34(1):27–29.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R Lyu. 2019. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. *arXiv preprint arXiv:1904.04446* .

Marsha M Linehan. 1997. Validation and psychotherapy. *American Psychological Association* .

Zi Haur Pang, Yahui Fu, Divesh Lala, Keiko OCHI, Koji INOUE, and Tatsuya KAWAHARA. 2023. Prediction of validating response from emotional storytelling corpus. In 人工知能学会全国大会論文集 第 *37* 回 *(2023)*. 一般社団法人人工知能学会, pages 2O5OS2a03–2O5OS2a03.

Zi Haur Pang, Yahui Fu, Divesh Lala, Keiko Ochi, Koji Inoue, and Tatsuya Kawahara. 2024. Acknowledgment of emotional states: Generating validating responses for empathetic dialogue. *arXiv preprint arXiv:2402.12770* .

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access* 7:100943–100953.

Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. *arXiv preprint arXiv:2105.06829* .

Biographical sketch

Zi Haur Pang is currently pursuing a Ph.D. degree at the Department of Intelligence Science and Technology, Kyoto University, Kyoto, Japan. He received his Master's degree from the same department at Kyoto University in 2024. Prior to his Master's studies, he worked as a data scientist at AirAsia in 2020. His research interests include human-agent interaction, affective computing and conversational AI.

Author Index