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).