Biswesh Mohapatra

INRIA
2 Rue Simone IFF
75012 Paris

biswesh.mohapatra@inria.fr
https://sites.google.com/view/biswesh-mohapatra

1 Research interests

Conversational grounding is an interactive process that has been studied extensively in cognitive science, whereby participants in a conversation check to make sure their interlocutors understand what is being referred to. (Clark and Brennan, 1991) propose the concept of "common ground" which is the mutual knowledge and mutual assumptions accumulated over the course of a conversation between the interlocutors during this interactive process. This common ground is built via words, of course, but also through the use of other modalities: pointing to objects in the environment, nodding to indicate that one has understood, eye-gaze or varying intonation in the speech, as pointed out by (Nakano et al., 2003). One way of thinking about this is that these units have an **underlying uncertainty** which is negotiated and removed by the participants before getting added to the shared information. The uncertainty comes from ambiguities that could be in the form of spatial references like "that car" or event references like "that was funny". When required, these uncertainties are solved through negotiations by providing additional information from the speaker when they sense a lack of understanding from the listener like "the big one next to the Ferrari" or by the listener themselves by asking for clarifications such as "You mean the blue one?". A grounding mechanism deals with removing the ambiguity between speakers while creating a local common understanding among them. This is important both when the model is the speaker and when it is the listener. Without a good grounding mechanism, conversations would not be robust and would often lead to misunderstandings. In fact, this is evident in the state-of-the-art dialogue systems that are increasingly using **Large Language Models(LLM)** as the Natural Language Understanding and Natural Language Generation modules. These LLMs are incapable of retaining and understanding all the information exchanged with the interlocutor during a session of conversation, as shown by Benotti and Blackburn Benotti and Blackburn (2021). They are also shown to be not very effective at making sure that the listener has grounded the information. Moreover, these LLMs do not have specific architectures to take into consideration the possibility of negotiations, clarifications, or cancellation of information during the conversation. They treat the entire dialog history as one unit where utterances are arranged according to the time. However, many dialogs contain overlapping utterances, and multiple independent pieces of information might be exchanged in parallel, or interleaved. This property of natural spoken dialog makes them unique.

While this process is essential to successful communication between humans and between humans and machines, work needs to be done on testing and building the capabilities of current dialogue systems in managing conversational grounding using the recent progress in LLMs such as Llama (Touvron et al., 2023), Palm (Anil et al., 2023) and GPT4 (OpenAI, 2023). Moreover, these models are text-only models and do not take into consideration the multi-modal aspect of grounding in situated environments. These include non-verbal behaviors, para-verbal behaviors and interaction with the environment. Removing the information present in the intonation of speech can lead to the introduction of ambiguities in the models. For example, an utterance that repeats the previous utterance can either be a confirmation of the previous utterance or a question for clarification depending upon the intonation. Grounding in such a context becomes even more challenging than just text-based models. While text-based large language models are able to take advantage of the vast resources of data available publicly for their training, corpora of multimodal data of daily human interactions are scarce and thus need models with the ability to ground the conversations in such low-resource settings.

Thus my research interests include **testing, understanding, and improving** the functioning of current language models with respect to **Multimodal Conversational Grounding**. My Ph.D. work will build on prior research, in modular dialog systems, that dealt with conversational grounding such as (Traum and Allen, 1994; Paek and Horvitz, 2000). However, since the majority of the previous work has been done using symbolic models that are hard to generalize, the work will take advantage of recent developments in pre-trained LLMs that have shown the ability to generalize to new scenarios.

Specifically, I propose to study and develop a

Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems, pages 15–17
September 11–12, 2023. ©2023 Association for Computational Linguistics

framework for incorporating multimodal conversational grounding capabilities into current dialog systems by asking the following questions -

1. How good are current Large Language Models with respect to conversational grounding and where could they be improved?

2. How can multimodal context (for example, a scene that interlocutors are viewing) be inserted into the LLMs to help in conversational grounding?

3. How can we make the models negotiate and align information contained by both participants?

4. How can the grounded information be represented and stored efficiently to use with Large Language Models?

I further elaborate on the above questions and discuss them with respect to the work we are doing, and that we plan to do, in the subsections below.

1.1 Testing Capabilities of Dialog Systems

Since, the advent of LLMs, dialog models have been able to take advantage of their capabilities to generate grammatically and semantically correct utterances. However, their performance in phenomena that are specific to dialogs such as conversational grounding has not been studied. We are currently doing a thorough study of the performance of current LLMs like Llama (Touvron et al., 2023) and GPT4 (OpenAI, 2023). Instances from the multi-modal dataset called Meetup (Ilinykh et al., 2019) are used to test the models on different aspects such as disfluencies, ambiguities, and grounding acts like repair, cancellation, acknowledgment, etc.

1.2 Incorporating Multimodal Context

Multimodal information coming from non-verbals, paraverbals, and the situated environment are very important for removing ambiguities and building common grounds. Looking at ways to provide such additional context to our language models before processing the dialogs thus becomes very important for a model to successfully ground conversations. The Meetup dataset gives us an opportunity to look into ways to incorporate image context information into LLMs. We also plan to look at other spoken dialogs like Switchboard (Godfrey and Holliman, 1993) to incorporate acoustic information as well.

1.3 Negotiating information for Alignment

I am interested in making the model negotiate and align the information contained by the other interlocutor, which is the main purpose of Conversational Grounding. For an effective Spoken Dialog System, we would want the system to ask for a minimum amount of clarifications without compromising on the ability to resolve ambiguities. Hence, the research will look into negotiating common ground by building models that could effectively work with the current LLMs.

1.4 Representing and Storing the grounded information

In order to negotiate and ground the information exchanged during a conversation, we need a good and effective way to represent the grounded information for which we might need to remove the ambiguities, that generally come in the form of additional referring expressions where necessary. Additionally, multimodal information from the conversations also may help us to figure out the intents of the utterances which in turn helps us remove ambiguities before storing the grounded information. Exploring ways to store the common ground effectively in order to use it during dialog generation with less inference time is another important topic that the project will look into.

2 Spoken dialogue system (SDS) research

It seems clear that spoken dialogue systems(SDS) will start incorporating visual elements as Situated Dialogs become more prominent with the rise of use cases such as Embodied Conversational Agents and Social Robots, in the coming years. I also believe that these agents will begin to serve as Personal Assistants, capable of helping users in learning new skills and also managing their daily schedules. Thus, research in the field of extracting accurate information from users over time and using it effectively would be very important as well. Hence, further work on grounding, including clarifying user communicative intentions through multimodal information, and incorporating world knowledge effectively, will be essential for fulfilling the potential of Spoken Dialog Systems.

3 Suggested topics for discussion

- Do we need a different set of architectures for spoken dialog systems that combine the various modalities in better ways or are the current transformer-based models the future?

- How does the advent of models like GPT4 shape the direction of research in Spoken Dialog Systems? What can we learn from these models that can help build better SDS?

- Will an end-to-end dialog system be able to eventually replace modular dialog systems? If not, then what are the key factors that obstruct the current or future end-to-end models from doing so?

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Tachard Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Z. Chen, Eric Chu, J. Clark, Laurent El Shafey, Yanping Huang, Kathleen S. Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Michael Brooks, Michele Catasta, Yongzhou Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, C Crépy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, M. C. D'iaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fan Feng, Vlad Fienber, Markus Freitag, Xavier García, Sebastian Gehrmann, Lucas González, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, An Ren Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wen Hao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Mu-Li Li, Wei Li, Yaguang Li, Jun Yu Li, Hyeontaek Lim, Han Lin, Zhong-Zhong Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alexandra Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Marie Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniela Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Ke Xu, Yu Xu, Lin Wu Xue, Pengcheng Yin, Jiahui Yu, Qiaoling Zhang, Steven Zheng, Ce Zheng, Wei Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *ArXiv* abs/2305.10403.

Luciana Benotti and Patrick Blackburn. 2021. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, pages 515–531. https://doi.org/10.18653/v1/2021.eacl-main.41.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren Resnick, Levine B., M. John, Stephanie Teasley, and D., editors, *Perspectives on Socially Shared Cognition*, American Psychological Association, pages 13–1991.

John J. Godfrey and Edward Holliman. 1993. Switchboard-1 release 2 ldc97s62.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meet up! a corpus of joint activity dialogues in a visual environment. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*. SEMDIAL, London, United Kingdom. http://semdial.org/anthology/Z19-Ilinykh$_{s}emdial_{0}006.pdf$.

Yukiko Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sapporo, Japan, pages 553–561. https://doi.org/10.3115/1075096.1075166.

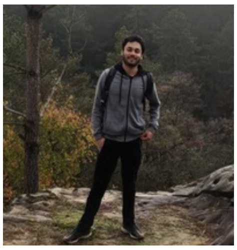
OpenAI. 2023. Gpt-4 technical report.

Tim Paek and Eric Horvitz. 2000. Conversation as action under uncertainty. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, UAI'00, page 455–464.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv* abs/2302.13971.

David Traum and James Allen. 1994. A "speech acts" approach to grounding in conversation .

Biographical sketch



Biswesh is a PhD student at Inria Paris working with the Articulab group under Prof. Justine Cassell and Prof. Laurent Romary. He did an Integrated Master of Technology majoring in Computer Science and Engineering at IIIT Bangalore, India. During his undergraduate degree he interned at IBM Research AI, Siemens Research and was also a Google Summer of Code Scholar (GSOC) in 2018. During his internship at IBM Research, he was exposed to the field of dialog systems and later worked at Inria for a year as a Research Engineer at Articulab, a group focusing on Multimodal Dialog Systems Research. Prior to his endeavor in research, he co-created an online digital simulator called Circuitverse.org, currently having more than 20,000 users worldwide and has also contributed to open-source platforms like OpenStreetMap.