11-877 Advanced Multimodal Machine Learning **Spring 2024**

Week 11: Interaction 3: Interaction with People

Instructors: Paul Liang and Daniel Fried *Synopsis Leads: Jiya Zhang*

Edited by Paul Liang *Scribes: Simran Khanuja, Jiya Zhang*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2024/

Summary: Multimodal machine learning is the study of computer algorithms that learn and improve through the use and experience of multimodal data. It brings unique challenges for both computational and theoretical research given the heterogeneity of various data sources.

In week 11's discussion session, the class discussed about multimodal interactions with people involving pragmatics definitions, and pragmatics cases in a multimodal setting, enabling human adaptations, and human vs LLMs on annotations or feedback. The following was a list of provided research probes:

1. Humans can provide many different types of feedback to help models accomplish challenging tasks in NLP, robotics, and multimodal tasks (e.g., ranking, scoring, and instructing). What are other types of feedback that can be useful for model training? Can we create a taxonomy of feedback forms, and describe each of their pros and cons? When should we use each type of feedback?
2. In NLP, there's been a trend of replacing human annotations/feedback with large language models. What are some limitations of this approach? What tasks that are currently done by humans cannot be replaced by large foundation models? What abilities might models need to have to be able to fully replace human annotators?
3. One key aspect of computational pragmatics is how context makes language have meaning beyond what's literally said. Give some examples of settings that involve multimodal context where the multimodality changes or enriches the literal meaning of the language.
4. Brainstorm some settings where it would be useful for models to adapt to the people they are interacting with. This adaptation could involve the peoples' language, preferences, and backgrounds. Are these settings within reach of current models? What techniques do you think will be useful to enable adaptation? Are there also societal concerns if these models understand too much of their users?
5. Pick a task that people carry out in pairs or teams, that involves some social or grounded interaction between the people (e.g., pair programming, advising a graduate student, assembling a piece of furniture). How close or far do you think our current AI approaches are from being able to collaborate with the people carrying out this task? What is a research agenda towards enabling human-AI collaboration?

As background, students read the following papers:

1. **(Required)** Pragmatics in Language Grounding: Phenomena, Tasks, and Modeling Approaches [Fried et al., 2023]
2. **(Required)** Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation [Fernandes et al., 2023]
3. (Suggested) Underspecification in Scene Description-to-Depiction Tasks [Hutchinson et al., 2022]
4. (Suggested) Symbolic Planning and Code Generation for Grounded Dialogue [Chiu et al., 2023]
5. (Suggested) Continual adaptation for efficient machine communication [Hawkins et al., 2020]
6. (Suggested) CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation [Li et al., 2023]
7. (Suggested) Robot Learning on the Job: Human-in-the-Loop Autonomy and Learning During Deployment [Liu et al., 2023b]

8. (Suggested) The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue [Haber et al., 2019]
9. (Suggested) Human Learning by Model Feedback: The Dynamics of Iterative Prompting with Midjourney [Don-Yehiya et al., 2023]
10. (Relevant) Draw Me a Flower: Processing and Grounding Abstraction in Natural Language [Lachmy et al., 2022]
11. (Relevant) Continual Learning for Grounded Instruction Generation by Observing Human Following Behavior [Kojima et al., 2021]
12. (Relevant) Continual Learning for Instruction Following from Realtime Feedback [Suhr and Artzi, 2023]
13. (Relevant) Computational Language Acquisition with Theory of Mind [Liu et al., 2023a]
14. (Relevant) Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning [Lazaridou et al., 2020]
15. (Relevant) Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts [Takmaz et al., 2020]
16. (Relevant) Speaking the Language of Your Listener: Audience-Aware Adaptation via Plug-and-Play Theory of Mind [Takmaz et al., 2023]
17. (Relevant) Human-in-the-loop Abstractive Dialogue Summarization [Chen et al., 2022]
18. (Relevant) Pragmatic Image Compression for Human-in-the-Loop Decision-Making [Reddy et al., 2021]
19. (Relevant) Werewolf Among Us: A Multimodal Dataset for Modeling Persuasion Behaviors in Social Deduction Games [Lai et al., 2022]
20. (Relevant) The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes [Kiela et al., 2021]
21. (Relevant) SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents [Zhou et al., 2024]
22. (Relevant) Vid2Robot: End-to-end Video-conditioned Policy Learning with Cross-Attention Transformers [Jain et al., 2024]
23. (Relevant) The Future of Human-in-the-Loop Cyber-Physical Systems [Schirner et al., 2013]

We summarize several main takeaway messages from group discussions below:

1 Pragmatics

Syntax, semantics, and pragmatics constitute the three fundamental aspects of natural language understanding (see Table 1). Human interactions introduce a layer of complexity to pragmatics, as it encompasses contextual elements and non-literal interpretations. Distinguishing between semantics and pragmatics can be challenging, as many concepts that do not neatly fit within the realm of semantics can often be attributed to pragmatics.

To elucidate pragmatics, consider language as a means of sketching ideas. When individuals interpret these sketches, they extrapolate meaning beyond the literal content based on their personal experiences. Pragmatics thus allows for the acquisition of additional information beyond what is explicitly conveyed, owing to the contextual knowledge possessed by the interlocutors.

In essence, languages should exhibit flexibility, adaptability, and grounded in human experience. Over time, linguistic aspects such as sentence construction, phonetics, and character representation may undergo simplification, while concurrently witnessing the emergence of new meanings and vocabulary. It raises intriguing questions regarding whether the overall complexity of languages remains constant amid such evolutionary shifts in both directions.

1.1 How to Make Models Learn Pragmatics

To enhance machine understanding of pragmatics within human language, we can provide data that help explain those context in a machine-understandable way. With an ample dataset and fine-tuning incorporating contextual representations, models can better grasp pragmatic elements. However, despite advancements,

Table 1: Three Aspects of Language

Aspects	Definition
Syntax	Syntax studies the sentence structure and grammar rules.
Semantics	Semantics studies the meanings of linguistic expressions (as opposed to their sound, spelling, etc.)
Pragmatics	Pragmatics studies the meaning of sentences with a certain context.

certain limitations persist, particularly in areas such as common sense comprehension and reasoning. These shortcomings may stem from inherent weaknesses within the model architecture or insufficient data availability. Additionally, leveraging a knowledge base facilitates the retrieval of words and phrases used in past contexts. This approach obviates the need for extensive model training and has been discussed in the "Photobook" paper [Haber et al., 2019].

1.2 Pragmatics in Multimodal Settings

Context exerts varied influences on meanings, particularly within multimodal settings where human communication incorporates multiple modalities. Beyond textual content, human interaction involves gestures, facial expressions, honorifics, and text emojis, imbuing communication with sentiments and cultural nuances that cannot be fully conveyed through pure text alone, posing challenges for language models. Pragmatic principles extend beyond linguistic domains to encompass non-verbal modalities. For instance, models may struggle to discern the significance of a stop sign, whereas humans readily interpret such signals, employing reasoning to anticipate future events and take appropriate actions.

2 Human Adaptation

As humans, the more we interact with someone, the better we understand their culture and context, enabling us to interpret their meaning more accurately. To enable LLMs to generalize knowledge from one person's interaction to other use cases, memory mechanisms should be implemented to retain information over time. This includes remembering personalized preferences and past context. However, existing tools often fail to retain prior context or preferences, which differs greatly from human interactions. There's a risk of over-adaptation in real-life scenarios, making it tricky to determine the boundary for allowing control over adaptive systems. It can be unsettling if the model knows too much about a person. Another concern is the risk of counterfactual presentation bias, where users are limited in topic suggestions, stifling exploration.

2.1 How to design Human Adaptation

Designing a machine learning system capable of incorporating human adaptation involves addressing both system and model-level challenges. At the model level, it's crucial to design data representations that include historic data or embeddings alongside identifiers for individual users. This ensures that the model can effectively leverage past interactions to tailor responses to specific users.

To fine-tune the model with new data and track updated parameters for individual users, the machine learning system must implement mechanisms for continual learning and adaptation. This could involve techniques such as online learning or incremental updates, allowing the model to dynamically adjust based on new information while preserving user-specific preferences.

However, incorporating human adaptation into chatbots or assistants comes with computational overheads, raising questions about feasibility and resource constraints. One potential approach is to leverage techniques like Low Rank Adaptation (LoRA, [Hu et al., 2021]) to fine-tune small weights to adapt to individual user preferences. By employing LoRA fine-tuned versions of chatbots, tailored responses can be served to each user efficiently.

Ultimately, the decision to include human adaptation in chatbots or assistants should consider the trade-offs between computational resources and user experience enhancement. Strategies like LoRA can help mitigate resource constraints while still providing personalized interactions.

3 Replace Human with LLM

The question of whether we can replace humans with LLMs for annotations or feedback is complex. While LLMs can achieve comparable accuracy to humans on selected benchmarks, there are significant limitations. LLM annotations lack clarity and interpretability, unlike human annotations. Additionally, LLMs cannot provide personalized inputs, which are crucial in many scenarios.

However, LLMs excel in certain tasks, such as sentiment classification of documents, where they may outperform humans. Nonetheless, there are concerns about ”test set pollution” as more LLM-generated data becomes prevalent. Training future models solely on LLM-generated data risks losing data diversity and potentially inhibiting the discovery of new insights.

Therefore, while LLMs can complement human efforts in certain tasks, it’s essential to carefully consider their limitations and the potential impact on data diversity and insights.

References

Jiaao Chen, Mohan Dodda, and Diyi Yang. Human-in-the-loop abstractive dialogue summarization, 2022.

Justin T. Chiu, Wenting Zhao, Derek Chen, Saujas Vaduguru, Alexander M. Rush, and Daniel Fried. Symbolic planning and code generation for grounded dialogue, 2023.

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. Human learning by model feedback: The dynamics of iterative prompting with midjourney, 2023.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. Bridging the gap: A survey on integrating (human) feedback for natural language generation, 2023.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches, 2023.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1184. URL https://aclanthology.org/P19-1184.

Robert D. Hawkins, Minae Kwon, Dorsa Sadigh, and Noah D. Goodman. Continual adaptation for efficient machine communication, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. Underspecification in scene description-to-depiction tasks, 2022.

Vidhi Jain, Maria Attarian, Nikhil J Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, Igor Gilitschenski, Yonatan Bisk, and Debidatta Dwibedi. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers, 2024.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes, 2021.

Noriyuki Kojima, Alane Suhr, and Yoav Artzi. Continual learning for grounded instruction generation by observing human following behavior. *Transactions of the Association for Computational Linguistics*, 9: 1303–1319, 2021. doi: 10.1162/tacl_a_00428. URL https://aclanthology.org/2021.tacl-1.77.

Royi Lachmy, Valentina Pyatkin, Avshalom Manevich, and Reut Tsarfaty. Draw me a flower: Processing and grounding abstraction in natural language, 2022.

Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James M. Rehg, and Diyi Yang. Werewolf among us: A multimodal dataset for modeling persuasion behaviors in social deduction games, 2022.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. Multi-agent communication meets natural language: Synergies between functional and structural language learning, 2020.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.92. URL http://dx.doi.org/10.18653/v1/2023.emnlp-main.92.

Andy Liu, Hao Zhu, Emmy Liu, Yonatan Bisk, and Graham Neubig. Computational language acquisition with theory of mind, 2023a.

Huihan Liu, Soroush Nasiriany, Lance Zhang, Zhiyao Bao, and Yuke Zhu. Robot learning on the job: Human-in-the-loop autonomy and learning during deployment, 2023b.

Siddharth Reddy, Anca D. Dragan, and Sergey Levine. Pragmatic image compression for human-in-the-loop decision-making, 2021.

Gunar Schirner, Deniz Erdogmus, Kaushik Chowdhury, and Taskin Padir. The future of human-in-the-loop cyber-physical systems. *Computer*, 46(1):36–45, 2013. doi: 10.1109/MC.2013.31.

Alane Suhr and Yoav Artzi. Continual learning for instruction following from realtime feedback, 2023.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.353. URL https://aclanthology.org/2020.emnlp-main.353.

Ece Takmaz, Nicolo' Brandizzi, Mario Giulianelli, Sandro Pezzelle, and Raquel Fernandez. Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4198–4217, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.258. URL https://aclanthology.org/2023.findings-acl.258.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents, 2024.