Javier Chiyah-Garcia

Heriot-Watt University
Edinburgh, Scotland
United Kingdom

fjc3@hw.ac.uk
https://jchiyah.github.io

1 Research interests

Conversations between humans are based on the collection of mutual knowledge, experiences, beliefs, assumptions and even goals of the interlocutors. We estimate these unconsciously, but they attribute meaning outside words that is crucial to understanding the interaction. This common grounding is what connects the meaning of the physical world with our language abstractions (Harnad, 1990). However, current intelligent systems do not share this mutual understanding and common grounding, heavily impairing the interaction. Users have to simplify queries, be more explicit and repeat information already mentioned so that their language matches the way that dialogue systems communicate.

As previous works have demonstrated, dialogue models trained on large scale datasets are not able to capture meaning beyond words (symbols) (Bisk et al., 2020; Bender and Koller, 2020; Bender et al., 2021) and thus fall short on tasks that require common sense or understanding nuanced meanings. Training on text alone or even text and images may not be enough to continue advancing in the field further. Spoken dialogue systems (SDSs) operate with a different view of the world from us and thus, struggle to draw connections between what can be observed or the result of actions and language, resulting in poor interactions.

1.1 Situated human-robot interaction

My research interests lie in the area of **situated interaction** in environments where robots and humans are co-located as part of my PhD. In these settings, **natural language instructions** given by a human can be rooted in surrounding objects, the dialogue history and even previous events. Therefore, human-robot interactions in situated environments requires agents to maintain appropriate situation awareness, which a dialogue system trained on text alone may not be suited to (Bisk et al., 2020).

As other fields related to interaction, such as computer vision or gesture recognition, start to reach maturity with efficient and high-performing off-the-shelf tools, it is important to incorporate these with SDS. A more holistic approach that combines dialogue, world state and other interaction modalities may yield better interactive systems and solve some shortfalls of the current field, such as the large amounts of training data needed or the disconnection between virtual and physical world. Previous works have proposed training models that combine natural language with visuals and world state to spur grounding (Bisk et al., 2016; Mei et al., 2016; Tan and Bansal, 2018; Suhr et al., 2019a; Shridhar et al., 2020; Padmakumar et al., 2021), yet most of them focus on understanding well-formed instructions sequentially, as opposed to fluid, unpredictable or noisy dialogues as with real-world conditions. In these cases and unlike current SDS, humans are able to collaborate and adapt, asking for clarifications or help when needed.

1.2 Referential ambiguities

Of particular relevance to situated dialogues are **referential ambiguities**, which arise when a referring expression does not uniquely identify the intended referent for the addressee. They signal a potential mismatch between the perspectives of the speaker and hearer (see e.g., Dobnik et al. (2015)) and thus hamper the interaction (e.g., not finding the correct object or resolving an action). Upon detecting such ambiguities, we engage in subsequent meta-communicative clarificational exchanges (Purver, 2004) to repair the miscommunication (Purver et al., 2018).

My current work explores the use of state-of-the-art models to **resolve referential ambiguities** in multimodal dialogues. We use the SIMMC 2.0 dataset (Kottur et al., 2021), where a conversational agent helps a user pick items to shop in a virtual shared environment. The agent needs to answer queries and perform instructions as well as keep track of the items mentioned throughout the dialogue in a multi-modal scene. Due to the high amount of similar-looking objects and long dialogues with dynamic objectives, the user needs to employ rich referring expressions, which commonly cause ambiguities in both the visual and conversational contexts (see Figure 1).

Initial analyses into the clarificational exchanges that arise suggest that models struggle to understand and resolve these ambiguities compared to other coreferences. This follows my work from the past year in vision and language models for detecting these ambiguities and resolving coreferences in multi-modal dialogues that led to Chiyah-Garcia et al. (2022). Vision and language models

Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems, pages 1–4
September 11–12, 2023.

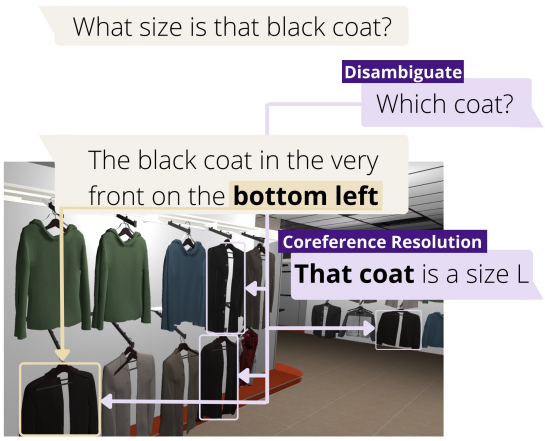

Figure 1: Example dialogue from the SIMMC 2.0 dataset where the system engages in a clarificational exchange to find the correct coat mentioned by the user.

are not enough, as they do not easily carry information across turns and/or are able to ground the information to the objects in the scene.

Future work will focus on learning the signals required to process clarifications and suitable architectures in the context of situated multi-modal interactions. Vision and language models, although promising, lack the relational information needed to fully ground both modalities in complex environments. Models that learn disentangled object representations (Bengio et al., 2013) could be better at exploiting the attributes of potential referential candidates and ultimately be better suited at resolving ambiguities in increasingly unstructured and multi-modal scenarios.

1.3 Past work

My current work on situated human-robot interaction with SDS builds upon my previous work in explainable dialogue systems to operate remote autonomous vehicles (Chiyah Garcia et al., 2018a,b, 2020a), automatically generating natural language explanations of learned robot behaviour (Chiyah Garcia et al., 2021) and analysing the use of crowd-sourced versus lab-collected data (Chiyah Garcia et al., 2020b; Lopes et al., 2020) for bootstrapping human-robot dialogue systems in the domain of emergency response.

2 SDS research

The field of dialogue research should work more closely with other fields related to interaction. Dialogue systems trained on text alone cannot fully understand the nuances of language and how these affect the physical world, hence works are increasingly combining natural language and rich image representations to improve text and vision benchmarks (Das et al., 2018; Suhr et al., 2019b; Zellers et al., 2019; Shridhar et al., 2020; Padmakumar et al., 2021). Improved multi-modal representations or more complete world views may be crucial for SDS to navigate more complex scenarios.

SDS could also become more robust and flexible in the way that they process natural language. Incrementally processing words instead of turns could enable SDSs to better understand and coordinate the conversation with a human (Schlangen and Skantze, 2009; Eshghi et al., 2015), as we often use feedback mechanisms such as backchannels (i.e., 'okay' or 'mhm') or clarifications to signal what has been grounded in a dialogue. Fluid human-robot interactions may require keeping track of the conversation context explicitly in real-time (Hough and Schlangen, 2016) so the SDS can self-repair the state when there are issues or misunderstandings (Hough, 2015).

Finally, the field of SDS could explore new ways of blending the natural language element of interactions with other modalities beyond vision, such as non-verbal communication. Agents that only understand words may not be suitable to interactions outside labs and in unstructured environments.

3 Suggested topics for discussion

Here are some of the topics for discussion:

- **Multi-modality** in SDS design, how to represent other modalities aside from language and how to use this to track the dialogue context.
- The rise of **large language models** such as GPT-4 and the challenges and opportunities that they bring to SDS.
- **Situated human-agent interaction**, where the agent can both observe and modify the world. Human-robot collaboration through natural language is a related sub-topic.
- **Incremental natural language understanding** either through a mix of semantic and statistical or pure methods.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, FAccT '21, page 610–623. https://doi.org/10.1145/3442188.3445922.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *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.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35(8):1798–1828.

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 *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.

Yonatan Bisk, Deniz Yuret, and Daniel Marcu. 2016. Natural language communication with robots. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 751–761. https://doi.org/10.18653/v1/N16-1089.

Francisco Javier Chiyah Garcia, José Lopes, and Helen Hastie. 2020a. Natural language interaction to facilitate mental models of remote robots. In *Proceedings of the Workshop on Mental Models of Robots, HRI'20*. ACM, Cambridge, UK, HRI'20.

Francisco Javier Chiyah Garcia, José Lopes, Xingkun Liu, and Helen Hastie. 2020b. CRWIZ: A framework for crowdsourcing real-time Wizard-of-Oz dialogues. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 288–297. https://www.aclweb.org/anthology/2020.lrec-1.36.

Francisco Javier Chiyah Garcia, David A. Robb, Atanas Laskov, Xingkun Liu, Pedro Patron, and Helen Hastie. 2018a. Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of The 11th International Natural Language Generation Conference*. ACM, Tilburg, The Netherlands, INLG'18, pages 99–108. http://www.aclweb.org/anthology/W18-65.

Francisco Javier Chiyah Garcia, David A. Robb, X. Liu, Atanas Laskov, Patron Patron, and Helen Hastie. 2018b. Explain yourself: A natural language interface for scrutable autonomous robots. In *Proceedings of Explainable Robotic Systems Workshop*. Chicago, IL, USA, HRI'18.

Francisco Javier Chiyah Garcia, Simón C. Smith, José Lopes, Subramanian Ramamoorthy, and Helen Hastie. 2021. Self-explainable robots in remote environments. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, HRI '21 Companion, page 662–664. https://doi.org/10.1145/3434074.3447275.

Javier Chiyah-Garcia, Alessandro Suglia, José David Lopes, Arash Eshghi, and Helen Hastie. 2022. Exploring multi-modal representations for ambiguity detection & coreference resolution in the simmc 2.0 challenge. In *AAAI 2022 DSTC10 Workshop*.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Simon Dobnik, Christine Howes, and John Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*. SEMDIAL, Gothenburg, Sweden. http://semdial.org/anthology/Z15-Dobnik_semdial_0006.pdf.

Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, and Matthew Purver. 2015. Feedback in conversation as incremental semantic update. In *IWCS 2015 - Proceedings of the 11th International Conference on Computational Semantics*. Association for Computational Linguistics, pages 261–271.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42(1-3):335–346. https://doi.org/10.1016/0167-2789(90)90087-6.

Julian Hough. 2015. *Modelling Incremental Self-Repair Processing in Dialogue*. Ph.D. thesis, Queen Mary University of London.

Julian Hough and David Schlangen. 2016. Investigating Fluidity for Human-Robot Interaction with Real-time, Real-world Grounding Strategies. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Stroudsburg, PA, USA, September, pages 288–298. https://doi.org/10.18653/v1/W16-3637.

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://aclanthology.org/2021.emnlp-main.401.

José Lopes, Francisco Javier Chiyah Garcia, and Helen Hastie. 2020. The lab vs the crowd: An investigation into data quality for neural dialogue models. In *Workshop on Human in the Loop Dialogue Systems at NeurIPS 2020*. https://arxiv.org/abs/2012.03855.

Hongyuan Mei, Mohit Bansal, and R. Matthew Walter. 2016. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *30th AAAI Conference on Artificial Intelligence, AAAI 2016*. AAAI press, pages 2772–2778. http://arxiv.org/abs/1506.04089.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramithu, Gokhan Tur, and Dilek Hakkani-Tur. 2021. TEACh: Task-driven Embodied Agents that Chat. In *arXiv:2110.00534 [Cs]*.

Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, University of London.

Matthew Purver, Julian Hough, and Christing Howes. 2018. Computational models of miscommunication phenomena. *Cognitive Science* 10(2):425–451. https://doi.org/10.1111/tops.12324.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Association for Computational Linguistics, Athens, Greece, pages 710–718. https://www.aclweb.org/anthology/E09-1081.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/1912.01734.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019a. Executing Instructions in Situated Collaborative Interactions. 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)*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 2119–2130. https://doi.org/10.18653/v1/D19-1218.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019b. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 6418–6428. https://doi.org/10.18653/v1/P19-1644.

Hao Tan and Mohit Bansal. 2018. Source-target inference models for spatial instruction understanding. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*. AAAI press, pages 5504–5511. http://arxiv.org/abs/1707.03804.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pages 6713–6724. https://doi.org/10.1109/CVPR.2019.00688.

Biographical sketch

Javier Chiyah-Garcia (he/him) is a PhD student at Heriot-Watt University working on human-robot collaboration using conversational agents. Currently, he is exploring methods of interaction with situated robots in smart factories with industry partner Siemens. Previously, he worked on the development of a dialogue system for remote operation of autonomous underwater vehicles, funded by the Ministry of Defence in the UK. He also explored how explanations affect the operator's mental model of the underwater vehicles. One of his goals is to make robots more intuitive to use through speech, and he is very excited about all the amazing things that human-robot teams can achieve together. After the PhD, he plans to join a research lab in industry to act as a bridge with academia and deploy more intuitive robots that make our lives a bit easier.