Multi-modal Language Models for Human-Robot Interaction

Ruben Janssens
ruben.janssens@ugent.be
IDLab-AIRO, Ghent University - imec
Ghent, Belgium

ABSTRACT

The recent progress in language models is enabling more flexible and natural conversation abilities for social robots. However, these language models were never made to be used in a physically embodied social agent. They lack the ability to process the other modalities humans use in conversation, such as vision, to make references to the environment and understand non-verbal communication. My work promotes the design of language models for physically embodied social interactions, shows how current technologies can be leveraged to enrich language models with these abilities, and explores how such multi-modal language models can be used to improve interactions.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Natural language generation**; **Computer vision**.

KEYWORDS

Human-Robot Interaction, multi-modal dialogue, conversational agent, Natural Language Generation, Natural Language Processing, situatedness, grounding

ACM Reference Format:
Ruben Janssens. 2024. Multi-modal Language Models for Human-Robot Interaction. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3610978.3638371

1 INTRODUCTION

Human-Robot Interaction (HRI) is almost by definition a multi-modal endeavour. Through a wide range of sensors, robots are equipped with the ability to sense the physical and social world around them, and they can make use of both verbal and nonverbal actions to respond. Concerted multi-modal interaction is likely to lead to effective responses and as such has been an ambitious target in the field [1, 13, 20].

Achieving *situated language interaction* is the challenge of making the robot produce and understand language that refers to the social and physical world around it [8]. This is an essential capability of an HRI dialogue system for two reasons. Firstly, if a robot can communicate with a user about the environment they cohabit, their interaction and collaboration will be more effective. Secondly, a naive user will have an expectation that a robot with "eyes", i.e. cameras, has access to the visual environment and has the ability to refer to visual elements in its linguistic interaction. Not doing so effectively will result in disappointment and disengagement.

A related concept is that of *grounding*, the act of attaching semantics and properties to those words that refer to elements in the world. From physically grounding words, such as "red" or "table", to abstract concepts such as "jealousy". Grounding is also seen as an important cognitive skill in HRI and one of the challenges the field poses for Artificial Intelligence (AI) [15, 19]. Grounding in robotics is also a concrete implementation of the wider symbol grounding problem [9].

Recent work in HRI has attempted to grow the capabilities of robots to engage in situated and grounded dialogue. These efforts have especially focused on *goal-oriented dialogue* in collaborative robots [10, 15], assistive robots [6, 17] or companion robots [11, 16].

A social robot also uses additional social signals —the user's pose, facial expressions, gestures, and non-verbal characteristics of speech such as volume, tempo, and prosody— to shape the interaction. Personalising and adapting social interaction is a key ability, both in human-to-human interaction and human-robot interaction. It has been demonstrated that personalisation and adaptivity contribute to the quality and outcomes of interaction. For example, a robot teacher adjusting its delivery when seeing that a pupil is interested, confused, or disengaged [22]. Unfortunately, most attempts at personalisation and adaptivity are limited in scope and often rely on hand-coded responses.

This work aims to answer the following research questions:

(1) How can language models be adapted to produce and understand references to visual information?
(2) How can language models adapt to the user through visually observing non-verbal signals?
(3) How does a robot having these abilities impact the interaction with a user and that user's perception of the robot?

2 MULTI-MODAL LANGUAGE MODELS

In the field of natural language processing, more and more multi-modal language models are being published. These models are able to process visual input alongside textual input, to do tasks such as image captioning (describing an image in a short textual description), visual question answering (answering natural language questions about an image) [7], and visually grounded dialogue (expanding visual question answering into a multi-turn conversation about an image) [5].

However, these tasks and the datasets published for them have limited use for HRI, because they are not made for *embodied* settings: they use visual input as shared conversation topic, instead of

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03.
https://doi.org/10.1145/3610978.3638371

as input showing the environment from one of the conversation partners' perspective.

Some first-person datasets exist, but they are not *social*: they are limited to simple questions or dialogues about explicit visual features, to enable collaboration, e.g. "What is this person doing?" [21], or spatial reasoning and embodied planning for robotics [18].

Cautious first steps are being made in adapting robot or agent behaviour to those social signals, such as by using reinforcement learning (RL) to select verbal supportive behaviours for an educational robot, or employing verbal and non-verbal interventions when detecting that a pupil is disengaged. They saw that students performed better when being stimulated by verbal cues [2, 3]. Leite et al. also detected disengagement in interactions of groups of children with robots and triggered repair interventions where necessary [14].

Recently, Continual Learning is being used to make those social signal detection models more adaptive to individual learners [4].

3 RESEARCH APPROACH AND PRELIMINARY WORK

As a starting point, my work has focussed on investigating how current computer vision and language technologies can be adapted to allow social conversations about visual context. To this end, I proposed a new task called "Visual Conversation Starters", where a robot has to start a social conversation (chit-chat) by asking a question that is based on visual information [12].

I compared various technical approaches to this problem: using text-only models that are provided with a textual description of the visual input and using end-to-end vision-to-text models, as well as comparing smaller (400-700M parameters) models that I fine-tune with large (12-175B parameters) models that are only prompted. Results showed that the end-to-end vision-to-text models tend to be more correct in their references to the visual information, even when fine-tuned on the same size data set, and that larger language models can generate more elaborate questions, that are perceived as more interesting and polite by humans. I also discovered that this elaborateness can be replicated in the smaller language models, by using the larger language model to generate a training set and fine-tuning the smaller model on that data. The combination of the large language model-generated training set with an end-to-end vision-to-text model is powerful as it is good at both referencing visual information and generating elaborate questions. The human perception of the questions was measured through a crowd-sourced evaluation study where 4 raters rated each question on 5 dimensions, using Likert scales.

Furthermore, I collected a data set of ca. 350 responses to the visual conversation starters, by having participants of a science festival interact with the robot. This in-the-wild data provides a useful stepping stone for further work on longer conversations that have to maintain reference to visual information.

4 FUTURE WORK

Next steps will focus on expanding this research to longer conversations, that have to continue being correctly grounded in the visual information over multiple turns. I will start by having language models drive a conversation that follows the visual conversation starter. The responses gathered in the science festival experiment described above will provide initial validation data. Later, I will test these models in real-world interactions, allowing also new references to visual information being introduced by the user later in the conversation, and we will build up knowledge about where the language models go wrong in maintaining visual grounding.

Having built up this knowledge, I will explore various technical approaches to improve the visual grounding, including prompt engineering to carefully inject the visual information into the language models at different steps in the conversation, building on experience gathered in the preliminary work, and fine-tuning vision-to-language models to improve the visual grounding, possibly combining the vision-to-language models with unimodal language models. Using these techniques, I aim to build a visually grounded conversation model for open-domain social conversations (chit-chat).

I will perform a user study where the impact of using this visually grounded language on users' perception of the robot is measured. The robot will start the conversation with a visual conversation starter, followed by an open-domain conversation between the user and the robot, driven by the visually grounded conversation model developed above. I will compare the visually grounded robot with one that does not do this. I will also measure the naturalness of the conversation through the length of the conversation and user perception of the robot using questionnaires. I also aim to measure a second-order effect of the visual grounding on the interaction, e.g. by presenting the user with a choice to which there is no one right answer, having the robot advise the user towards one of the options, and measuring the users' likelihood of following the robot's advice in the two conditions. The interactions taking place in this experiment will be recorded using video and audio and I may use Conversation Analysis techniques to qualitatively analyse what happens during these conversations, especially to investigate where the conversation model fails.

Finally, another aspect of visual grounding, which will not sufficiently be covered by the previous steps of my research, is reacting to non-verbal signals the user sends. These signals are useful to better understand when the robot has made a mistake – a major limitation of current social agents is that they are not aware of when they did not interpret the user's question correctly or when they say something inappropriate. I argue that conversational agents should not only be improved by aiming their output to be completely accurate, but mostly by being able to adapt their output during the interaction, based on implicit or explicit feedback given by the user – this is also how human-to-human interactions work, with many small repair mechanisms when a misunderstanding occurred.

I aim to use these non-verbal signals in order to further improve the conversation models developed above. I will develop a classification model based on Facial Action Units to detect communication mistakes, using Continual Learning techniques to fine-tune models to individual users. This feedback will then be fed back into the model using in-context learning.

ACKNOWLEDGMENTS

This research received funding from the Flemish Government (AI Research Program). I thank my supervisors Tony Belpaeme and Thomas Demeester for their guidance.

REFERENCES

[1] Tony Belpaeme, Paul Baxter, Robin Read, Rachel Wood, Heriberto Cuayáhuitl, Bernd Kiefer, Stefania Racioppa, Ivana Kruijff-Korbayová, Georgios Athanasopoulos, Valentin Enescu, et al. 2012. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction* 1, 2 (2012).

[2] LaVonda Brown and Ayanna M Howard. 2013. Engaging children in math education using a socially interactive humanoid robot. In *2013 13th IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids)*. IEEE, 183–188.

[3] LaVonda Brown, Ryan Kerwin, and Ayanna M Howard. 2013. Applying behavioral strategies for student engagement using a robotic educational agent. In *2013 IEEE Int. Conf. on systems, man, and cybernetics*. IEEE, 4360–4365.

[4] Nikhil Churamani, Minja Axelsson, Atahan Caldir, and Hatice Gunes. 2022. Continual learning for affective robotics: A proof of concept for wellbeing. *arXiv preprint arXiv:2206.11354* (2022).

[5] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 326–335.

[6] Alessandro Di Nuovo, Frank Broz, Ning Wang, Tony Belpaeme, Angelo Cangelosi, Ray Jones, Raffaele Esposito, Filippo Cavallo, and Paolo Dario. 2018. The multi-modal interface of Robot-Era multi-robot services tailored for the elderly. *Intelligent Service Robotics* 11, 1 (2018), 109–126.

[7] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision* 14, 3–4 (2022), 163–352.

[8] Charles Goodwin. 2000. Action and embodiment within situated human interaction. *Journal of pragmatics* 32, 10 (2000), 1489–1522.

[9] Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42, 1-3 (1990), 335–346.

[10] 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 SIGdial Meeting on Discourse and Dialogue*.

[11] Bahar Irfan, Anika Narayanan, and James Kennedy. 2020. Dynamic Emotional Language Adaptation in Multiparty Interactions with Agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[12] Ruben Janssens, Pieter Wolfert, Thomas Demeester, and Tony Belpaeme. 2022. 'Cool glasses, where did you get them?" Generating Visually Grounded Conversation Starters for Human-Robot Dialogue. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 821–825.

[13] Wafa Johal, Gaëlle Calvary, and Sylvie Pesty. 2015. Non-verbal signals in HRI: interference in human perception. In *International Conference on Social Robotics*. Springer, 275–284.

[14] Iolanda Leite, Marissa McCoy, Monika Lohani, Nicole Salomons, Kara McElvaine, Charlene Stokes, Susan Rivers, and Brian Scassellati. 2016. Autonomous disengagement classification and repair in multiparty child-robot interaction. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 525–532.

[15] Séverin Lemaignan, Mathieu Warnier, E. Akin Sisbot, Aurélie Clodic, and Rachid Alami. 2017. Artificial cognition for social human–robot interaction: An implementation. *Artificial Intelligence* 247 (2017), 45 – 69. https://doi.org/10.1016/j.artint.2016.07.002 Special Issue on AI and Robotics.

[16] Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems* 63 (2015), 22 – 35. https://doi.org/10.1016/j.robot.2014.09.031

[17] Peter Mayer and Paul Panek. 2014. Towards a multi-modal user interface for an affordable Assistive Robot. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 680–691.

[18] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021* (2023).

[19] Oliver Roesler, Amir Aly, Tadahiro Taniguchi, and Yoshikatsu Hayashi. 2019. Evaluation of word representations in grounding natural language instructions through computational human-robot interaction. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 307–316.

[20] Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *2011 Ro-Man*. IEEE, 247–252.

[21] Anirudh Sundar and Larry Heck. 2022. Multimodal Conversational AI: A Survey of Datasets and Approaches. In *Proceedings of the 4th Workshop on NLP for Conversational AI*. 131–147.

[22] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27, 12 (2009), 1743–1759. https://doi.org/10.1016/j.imavis.2008.11.007 Visual and multimodal analysis of human spontaneous behaviour:.