To Your Left: A Dataset and a Task of Spatial Perspective Coordination

Mattias Appelgren
FLoV and CLASP
University of Gothenburgh
mattias.appelgren@gu.se

Simon Dobnik
FLoV and CLASP
University of Gothenburgh
simon.dobnik@gu.se

Abstract

In dialogue speakers speak about the same scene while looking at it from different points of view. Who's view is utilised in utterances shifts inside the same conversation and is coordinated by participants as part of their common ground. However, current AI systems are generally trained on a single perspective or multiple random perspectives and are incapable of such coordinations. In this paper we propose a novel artificial dataset that we are developing as a part of our ongoing work with the purpose of evaluating the current state of the art on their ability to learn to recognise and generate spatial descriptions where the speaker and listener have different points of view.

1 Introduction

When humans communicate with each other we have to consider whose Point of View (POV) or Frame of Reference (FoR) (in this paper we use these terms interchangeably) a description is given from (Levinson, 2003). For example, "The tiger is hiding in the bushes to the right of the child" in this example there are at least three different POVs to consider: the speakers, the listeners, and the child's. The listener would need to infer which POV to use in order to complete its intended task, e.g. aiming a tranquilizer at the correct bush. Furthermore, if the listener later becomes the speaker in the same conversational and situational context, what perspective they would take in their utterance? Current state of the art models struggle with spatial relations on their own (Kelleher and Dobnik, 2017; Liu et al., 2023), and very few consider FoR explicitly (some notable exceptions include Lee et al. (2022); Hua et al. (2018); Steels and Loetzsch (2006)). However, Dobnik (2009) found that even when participants are asked to use a fixed FoR they would shift it in response to different situations. Dobnik et al. (2020) further study this phenomanon in human dialogues and find that people will shift FoR throughout extended dialogues, often without explicitly marking the shift.

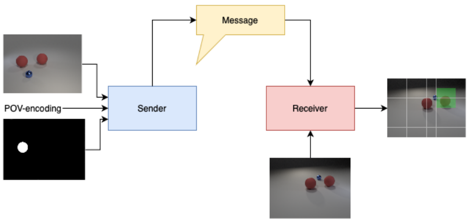

Figure 1: The speaker sees the image, a mask to identify the target, and the listener's POV encoded as a 1-hot vector. It produces a message referring to the target object. The listener sees the same scene from a different POV and receives the message and must predict the region which contains the described object.

In order for robots and other AI systems to communicate successfully with humans they need the capability to generate and interpret referring expressions from different FoRs and in continuous conversational and situational contexts. In this paper we propose an artificial dataset and task which will diagnose systems' ability to consider FoR in spatial descriptions and test conditons under which FoR can be learned by them. We describe work in progress, which means we have not completed the development of this data nor any experiments.

2 Dataset and task

2.1 Task

In our task two agents must communicate about a scene which they are viewing from different POVs. The agents take on the roles of speaker or listener. The speaker is shown a visual scene and an object within the scene that it must refer to. The listener sees the scene from a different POV. The speaker must generate a message describing the target object and the listener must interpret the message and predict which region of the image the object is in.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11–12, 2024, Trento, Italy.

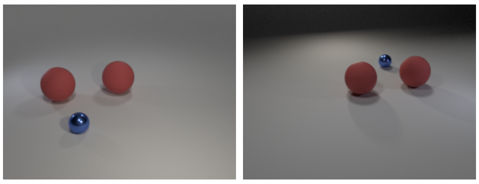

Figure 2: Two views of the same scene.

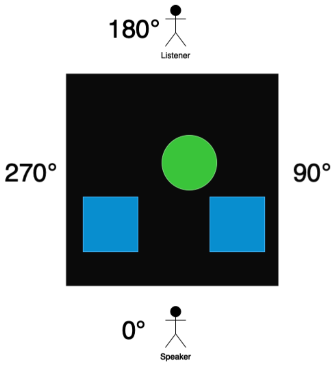

Figure 3: The listener could view the world from four different angles relative to the speaker

Figure 1 shows a diagram of the task set-up.

2.2 Data

We opt for artificial scenes so that we can control precisely the contextual attributes of the interaction environment. The first consideration is that the target object must not be uniquely identifiable from its visual attributes. In Figure 2, if the target was the blue sphere it would be enough to describe it as such to identify it. However, if the target is one of the two red spheres spatial descriptions would have to be used, e.g. “the leftmost red sphere”. As such, each image will contain a target object and one or more distractors that share all of the same visual features as the target, in addition to landmark objects which have different visual features to the target, such as the blue sphere in Figure 2. We will capture the scene from four directions, as shown in Figure 3. In different sub-tasks we will experiment with showing the speaker and listener different combinations of views, for example to give the agents the ability of egocentric perspective shifts (Levinson, 2003).

We will use the code that generated the CLEVR dataset (Johnson et al., 2016) to generate the images, potentially extending it to more general objects as done by Lee et al. (2022), both use the Blender graphics software to render images of objects. Figure 2 shows an example of the same scene from two opposite perspectives.

2.3 Experiments

We will implement the speaker and listener in the EGG toolkit (Kharitonov et al., 2019) which is designed to train emergent-language agents from language games. We use the emergent language setting to evaluate current model architecture’s ability to learn to communicate while restricting certain contextual properties, like viewing scenes from different POVs. Given we allow the agent’s to create any language it is important that we design the task in such a way that they actually have to solve the intended task.

We intend to answer the following questions:

1. Can current model architectures learn to communicate with differing POVs
2. Can we improve models’ ability to learn through special pre-training
3. Given contextual priming, do the emergent languages show properties of human language

After these initial experiments we want to see if we can transfer these learnings to models which use human language. We can do this by generating labels for our underlying data.

3 Related Work

Spatial Relations have been studied on without FoR e.g. Cheng et al. (2024); Kelleher and Dobnik (2017); Fu et al. (2024); Liu et al. (2023); Kuhnle and Copestake (2017); Kordjamshidi et al. (2011). Liu et al. (2023) allow annotators to use camera or intrinsic FoR but do not model them explicitly. Lee et al. (2022) model intrinsic FoR, e.g. “plane left of elephant” from the elephant’s FoR. This is complementary to our data which poses different challenges to models. Steels and Loetzsch (2006) have robots view events from different perspectives and perform a language game, creating a similar scenario to ours, however, their model architectures are quite out of date so we are due a new look at the problem. Fu et al. (2024) propose several visual benchmarks for visual language models, one is multi-view reasoning, however the task is simply to identify how the camera has moved (left or right) with no spatial reference task. Dobnik et al. (2020) gather dialogues with spatial descriptions from different FoR, however, the number of dialogues is too small to train modern models on and the task is more complex, this proposed data is a first step towards solving this more complex task.

Acknowledgments

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision language model. *ArXiv*, abs/2406.01584.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom.

Simon Dobnik, John D. Kelleher, and C. Howes. 2020. Local alignment of frame of reference assignment in english and swedish dialogue. In *Spatial Cognition*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. *ArXiv*, abs/2404.12390.

Hua Hua, Jochen Renz, and X. Ge. 2018. Qualitative representation and reasoning over direction relations across different frames of reference. In *International Conference on Principles of Knowledge Representation and Reasoning*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997.

John D. Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. In *Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017), Gothenburg, 12 –13 June*, volume 1 of *CLASP Papers in Computational Linguistics*, pages 41–52, Gothenburg, Sweden. Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, CLASP, Centre for Language and Studies in Probability.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. Egg: a toolkit for research on emergence of language in games. *ArXiv*, abs/1907.00852.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing*, 8(3):4:1–4:36.

Alexander Kuhnle and Ann A. Copestake. 2017. Shapeworld - a new test methodology for multimodal language understanding. *ArXiv*, abs/1704.04517.

Jae Hee Lee, Matthias Kerzel, Kyra Ahrens, Cornelius Weber, and Stefan Wermter. 2022. What is right for me is not yet right for you: A dataset for grounding relative directions via multi-task learning. In *International Joint Conference on Artificial Intelligence*.

Stephen C. Levinson. 2003. *Space in language and cognition: explorations in cognitive diversity*. Cambridge University Press, Cambridge.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual Spatial Reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Luc L. Steels and Martin Loetzsch. 2006. Perspective alignment in spatial language. In *Spatial Language and Dialogue*.