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

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

Simon Dobnik
CLASP / Address line 1
University of Gothenburgh
simon.dobnik@gu.se

Abstract

Speaking about the same scene from different points of view is a natural part of human dialogue. The point of view being used often shifts inside of the same conversation and is coordinated by participants as a 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) 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 speaker's, the listener's, 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 a listener later becomes a speaker in the same conversational and situational context, what perspective would they 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 FoR. 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.

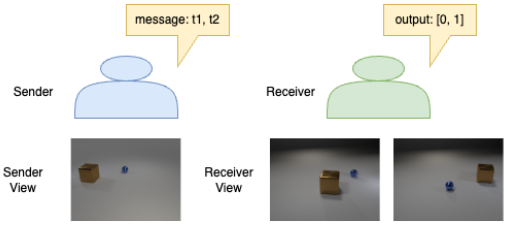

Figure 1: The structure of the language game. The Sender selects a view, observes an image and produces a message. The receiver sees two images and the message and must produce some output, in this case selecting the correct image. the goal of the game is for agents to coordinate on language the messages of which communicate the point of view.

In order for robots and other AI systems to communicate successfully with humans then 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 the signaling game a speaker s sees a set of information i_s, e.g. an image. The speaker must convey a message m to a listener l who in turn has its own information i_l. The listener must then produce an output y based on m and i_l. The interaction is evaluated on whether y matches an expected target output y_t. The goal of the game is to get the agents to converge on a language, the messages of which convey the point of view. The language is com-

pletely made up by the agents, the speaker initially selecting random tokens and the listener selecting output randomly, but through feedback they both converge on a shared understanding. For example, Chaabouni et al. (2021) used this formalism to study how artificial agents would learn to communicate about colours. In their experiments i_s was a single colour, represented as its RGB value. i_l was two colours, the same colour as the sender saw plus an additional distractor colour. The learner produces a 1-hot vector, the output y, where the target, y_t would be the same colour that the sender saw. The message m would have to convey which colour the sender saw, and be specific enough such that the listener would be able to select it when a distractor was present. The authors then used this to observe in what way the agents had "chosen" to conceptualise the colour space.

Havrylov and Titov (2017) used a similar set up to Chaabouni et al. (2021) only they used images instead of colours. In their experiment the speaker is shown the target image img_t and the listener is shown a set of images $imgs$ where $img_t \in imgs$. Again, the listener must select which of the images is the correct one. Havrylov and Titov (2017) found messages which correlated with e.g. images containing pizza. Thus the messages seemed to contain information about the contents of the images.

We wish to investigate whether we can influence what type of information the agents communicate through the language that they are constantly adapting by manipulating the images that we show to them, by carefully picking which images we show and which distractors. We will investigate a curriculum of different scenarios with increasing complexity, with the goal of having the models learn language referring to visual features in each one. We capture this curriculum in Figure 3. Each experiment will use the set-up with the sender seeing one image (and potentially some additional information) and the listener seeing two (or more) images and having to select among them.

The first row is simply a replication of the colour experiment in Chaabouni et al. (2021), so the agent needs to learn to communicate about colour. In the second row we start introducing images which we generate using Blender which are of geometric shapes with different colours and size (using the same method of generation as the CLEVR dataset (Johnson et al., 2016)). The first experiment is simply one where we vary a single attribute between objects, e.g. the colour or the shape. E.g. the target image may contain a big blue sphere while a distractor image contains a big blue cube. The agents would then need to learn how to communicate the chosen attribute. We can increase the complexity of this task by varying more attributes and adding more distractors. E.g. having the target be a big green cube with distractors: small green cylinder and big blue cylinder. Here the agents should learn to describe the objects in more detail.

In the next step we add in spatial relationships. Here the target could be a green square to the left of a blue cylinder. The listener would be shown distractors with objects that share the same visual attributes but a different spatial relation between them, e.g. a green square to the right of a blue cylinder. Thus it would not be sufficient to just describe one of the objects but the relation between them would have to be described as well.

In the final step we keep the same type of images in the previous one but now we show the listener the scene from a different point of view. We capture images of the scene from four different points of view each one with the camera rotated $90°$. Figure 5 shows a scene which has been captured in this way. Figure 2 shows an example of what the sender and receiver may see. There are several versions of this experiment that we could try which would require different strategies from the agents in order to communicate about the images. The first would be to always show the listener the images rotated by a fixed amount, e.g. always show them the $180°$ images. In this case the agents would not have to learn to explicitly communicate about the point of view shift. Instead, assuming the sender learns a word for "left" the receiver would simply have to learn that this means "right" (in our semantics of these words). This is an intuition of how this would work if humans were to learn this task, however, it would be interesting to see if the agents could learn the same thing. In fact, one way to investigate this could be to use the same sender as in the previous experiment (spatial but unrotated images) but train a new listener which sees the rotated images. However, more interesting is to see if the agents can learn to communicate in a situation where the points of views shift. To experiment with this we will show the listener different points of view each time. We will select one of the four points of view in each interaction. Now, it would be impossible for the speaker and listener to coordinate on which image to choose if neither of the agents knew their relative points of view. As such we will encode this

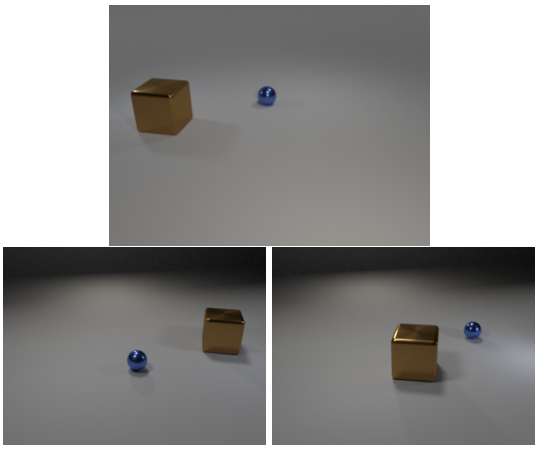

Figure 2: Top:sender view, bottom: listener view. The target image for the listener is the left one, which is a 180° rotation of the top image.

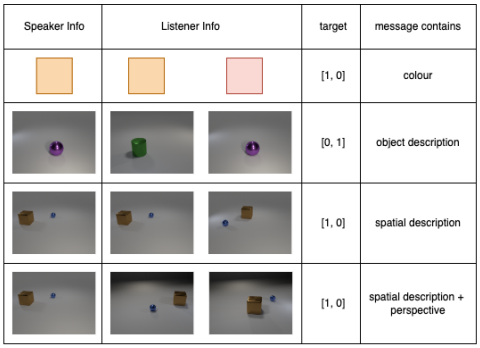

Figure 3: A curriculum of tasks of increasing difficulty.

information as a simple 1-hot vector where each position represents one of the four rotations. We can then provide this information either to the sender, the receiver or both. In principle giving it to the sender would mean the sender would have to adjust its utterance to accommodate the speakers point of view by adjusting the referring expression. If the listener receives the pov encoding the speaker would not know and would therefore simply describe it from its point of view while the listener would have to make the adjustment to its point of view. Finally, if both receive the information the agents would need to learn to coordinate on which method to use (or potentially to explicitly have a word in their shared language for indicating which pov is being used). We would investigate which strategy the agents adopt in our analysis.

Our goal is to investigate model's ability to learn to communicate about spatial relationships when the agents are viewing the scene from different points of view. Our interest lies in the dual tasks of producing and interpreting spatial descriptions. Our experimentation will begin using the paradigm of signaling games (Lewis, 1969; Kharitonov et al., 2019).

2.2 Data

We opt for artificial scenes so that we can control precisely the contextual attributes of the interaction environment and to allow us to capture images of the scene from different directions. The scenes consist of geometric objects placed on a white tabletop with a particular light source. Images are generated in pairs (or n-tuples) where one image is the target and the other are distracting images. In the current set-up we use the same variation in visual features as is present in the CLEVR dataset (Johnson et al., 2016), namely: colour, shape, size, and material. If two objects share the same visual features then they are identical in the scene except for the effects of lighting and object rotation. A scene is generated in order to fulfill certain criteria, as described above. The scenes, which are generated using Blender and a modification of the code used to generate CLEVR, will be captured from four different directions, each 90 degrees rotated from one another. Figure 5 shows an example of these four views.

3 Method

We are implementing this in the EGG framework (Kharitonov et al., 2019) which implements the signaling game framework. Figure 1 shows the overall shape of the interaction. The agents are neural models implemented in PyTorch. Figure 4 shows an example of what the structure of those models. The model would consist of a vision encoder which takes the image and extracts a feature vector, this would then be passed to a language generation model, e.g. an LSTM. The language generator generates a string of tokens, the size of the vocabulary and the maximum length of the messages are hyper-parameters that can be varied and experimented with. The receiver is a decoder model which uses some kind of decoder to decode the message as well as image encoders to encode the images it is given. These can then be fed into some kind of decision network, e.g. a Multi-Layer Perceptron (MLP) (i.e. a fully connected neural network) with a softmax output layer. We then evaluate this using cross entropy and back-propagate the error through the listener and then, since the message is discreet, either use reinforcement learning or Gumbel Softmax to propagate the error to

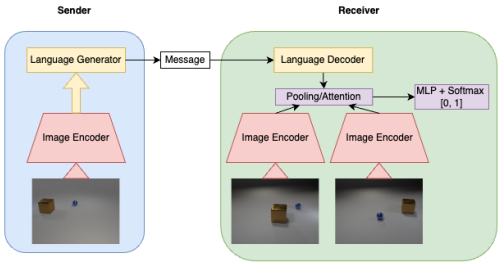

Figure 4: The internal structure of the agents.

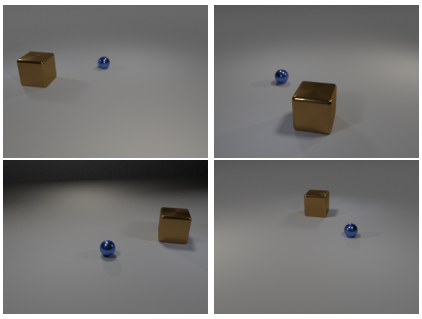

Figure 5: Four views of the same scene.

the sender.

4 Analysis

The success of the agents will be evaluated on the listener's ability to select the correct image. We will measure this accuracy over the training steps which will give us a curve showing the speed of convergence of the agents' learning. Further we will keep a held out validation set. This is often not done in recent trials of this nature. We think it is important to do this since it will show that the agents have learned a communication schema which is generalisable to unseen images and have not just over-fit to the training data and thus learned to memorise the data. We will see our experiments on a task successful if the agents converge on a shared language which has high accuracy on the validation set.

In addition to this we will perform several evaluations of the languages that the agents learn to try to understand what kinds of strategies the models have utilised to communicate about the images. The first method we will use is to measure the likelihood of a particular language token co-occuring with images containing particular visual attributes. E.g. $P(t_1|blue)$ being the probability that token 1 occurs when a blue object is present in the scene. We will investigate several n-gram lengths to see if the tokens are entangled or if each token represents a single concept.

Another way we will analyse the generalisability of the language learned is to perform a type of visual ablations. Our expectation, for example, would be that if the scene contains a blue cylinder left of a green square than the message the sender generates to describe that would be able to be used by the listener to recognise a scene with the same properties but which is not the exact same image. We will perform this test by having the sender generate a message for one image and then insert a different target image in the listeners set of images. If this test fails we can also add this type of scenario into the training data.

We will try to determine in which way the speaker and listener solve the problem of the perspective shift by investigating, for example, if there exists a token which co-occurs with a specific listener perspective (indicating that the telling the listener which point of view its describing it from).

5 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 elephants 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) present a set of dialogues where people speak about objects on a table that they see from different points of view. The work highlights the need for AI systems to communicate about spatial relations from shifting POVs. Our proposed dataset would provide a testbed for that 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

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2021. Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences of the United States of America*, 118.

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.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *ArXiv*, abs/1705.11192.

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.

David Lewis. 1969. Convention. *A Philosophical Study. URL: https://www. princeton. edu/˜ harman/Courses/PHI534-2012-13/Nov26/lewis-convention1.pdf*.

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*.