Bram Willemsen

KTH Royal Institute of Technology
Division of Speech, Music and Hearing
Stockholm, Sweden

bramw@kth.se
willemsenbram.github.io

1 Research interests

My current research lies at the intersection of vision and language; more specifically, I am working on problems related to **visually grounded language understanding** for conversational agents. By observing conversations between humans, we aim to improve our understanding of **referring language use in dialogue**, so that we may develop systems capable of engaging in interactions with humans that involve references to a co-observed world.

1.1 Referring language use in dialogue

If we define dialogue as an exchange of information, the act of referring can be seen as a speaker attempting to direct the attention of their addressee to some perceivable information of note, i.e. the referent. Participants in a conversation ordinarily collaborate in the process of producing and grounding references (Clark and Wilkes-Gibbs, 1986). Each party may contribute to the description and the successful identification of a referent:

> A: Have you seen my dog?
> B: Golden? Not particularly bright-looking?
> A: No. He is a black Labrador and I'll have you know he's brainier than most people.

Mentions of the same referent in a discourse are said to corefer. For example, in the above exchange *"my dog"* and *"he"* are coreferences, as they denote the same referent.

If we want to model dialogues that reflect this manner of referencing a co-observed world, phenomena such as described should be represented in the data that is used for training and evaluation. Upon review of existing work, we found that few visually-grounded dialogue tasks and datasets had explicitly accounted for these dialogue phenomena. This led us to introduce a task of our own, a collaborative image ranking task we called **A Game Of Sorts** (Willemsen et al., 2022). In this grounded agreement game (Schlangen, 2019), two players are asked to rank nine images based on a given sorting criterion. The game is implemented as a web application that has players exchange text-based messages to discuss how the scenarios with which they are presented–and in which these sorting criteria are embedded–should affect the rank of each image. Although the players see the same images, the position of the images on their screens is randomized. This forces them to refer to each image based on its content rather than its position on screen. We define task success as the players managing to reach an agreement on which rank to assign to each image *and* actually assigning the agreed upon ranks to the same images; the latter is not a given due to the players not being able to see each other's perspective. As the game is played over multiple rounds with the same set of images, we effectively guarantee repeated mentions of the same referents. Analysis of dialogues collected with our task showed it managed to induce mixed-initiative interactions in which the phenomena of interest were present.

1.2 Visually grounded language understanding

Recent advances in multimodal representation learning have led to significant improvements on vision-language benchmarks. Vision-language models (VLMs), such as CLIP (Radford et al., 2021), that have been pretrained on hundreds of millions of image-text pairs, learn to jointly embed both modalities via contrastive objectives. The learned representations have shown to be useful for downstream tasks that involve matching images and text.

Nevertheless, we recognize a few limitations of the current paradigm when we consider interactive settings in which pretrained models encounter new information. Incorporating new information in already trained models remains a challenge (see e.g. Parisi et al., 2019). Retraining from scratch in light of new data is currently not a feasible solution due to the resource-intensive nature of the process. Moreover, these VLMs are trained on large image captioning datasets or similar data from web-based sources that contain images paired with (high-level) descriptions. Although models trained on this data learn to associate (visual representations of) things with the words and phrases that are commonly used to describe them, this general language use may not align with how humans in a conversation would describe those same things. Take, for instance, mentions of referents that are non-descriptive in terms of visually perceivable attributes, such as names of pets: no pretrained model can reasonably be assumed to know, *out of the box*, that a particular dog goes by the name of *Sir Gideon Ofnir,*

Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems, pages 21–23
September 11–12, 2023.

the All-Knowing. For these reasons, we experimented with rapid domain adaptation based on a simple model that learns to transform VLM embeddings to better align the representations with the expected language use without updating the parameters of the base model (Skantze and Willemsen, 2022). Although this approach does not provide a fundamental solution to continual learning with VLMs, as the newly acquired knowledge is not incorporated into the base model, it does provide an opportunity–albeit limited–to adapt to users during an interaction.

A further challenge is the handling of longer texts. Given that most VLMs have been trained to optimize for matching relatively short, high-level descriptions with their associated images, they do not learn to process discourse-like inputs. This limits their zero-shot performance on tasks that require image-text matching based on conversational inputs. Reference resolution in visually-grounded dialogue, by which we mean the grounding of mentions to their exophoric referents, can be formulated as such a task. We proposed an approach to this task (Willemsen et al., 2023) that addressed the discourse processing limitations of pretrained VLMs by fine-tuning a causal large language model (LLM) to function as an auxiliary discourse processor: the pretrained LLM learns to generate definite descriptions of referents based on the (co)referential information in the dialogue; the generated descriptions are then used by the pretrained VLM for zero-shot identification of referents.

2 Spoken dialogue system (SDS) research

In the coming years, I expect much emphasis to be put on learning to integrate modalities end-to-end. Problems that are inherently multimodal, which aside from SDSs also includes visually grounded language understanding, ultimately require solutions that respect the interactions between modalities. For example, even though we can use an automatic speech recognition system to transcribe speech and use an LLM as the natural language understanding component of the SDS to processes the transcription, we would miss out on extralinguistic context, such as the prosodic cues that were present in the speech signal, that may be vital to the interpretation of the message: in trying to understand what message someone is attempting to convey, we do not simply take note of the uttered words; we also pay attention to how those words were uttered.

3 Suggested topics for discussion

- Drawbacks of LLMs: What are the potential negative consequences for end users of the unchecked use of LLMs in SDSs?
- Influence from industry: To what extent should corporate interests be allowed to dictate the direction of academic research?
- Governance of AI: How can we expect SDS research to be affected by looming regulations?

References

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22(1):1–39. https://doi.org/10.1016/0010-0277(86)90010-7.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks* 113:54–71. https://doi.org/10.1016/j.neunet.2019.01.012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*. PMLR, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. https://proceedings.mlr.press/v139/radford21a.html.

David Schlangen. 2019. Grounded Agreement Games: Emphasizing Conversational Grounding in Visual Dialogue Settings. *CoRR* abs/1908.11279. http://arxiv.org/abs/1908.11279.

Gabriel Skantze and Bram Willemsen. 2022. CoLLIE: Continual Learning of Language Grounding from Language-Image Embeddings. *J. Artif. Int. Res.* 74. Place: El Segundo, CA, USA Publisher: AI Access Foundation. https://doi.org/10.1613/jair.1.13689.

Bram Willemsen, Dmytro Kalpakchi, and Gabriel Skantze. 2022. Collecting Visually-Grounded Dialogue with A Game Of Sorts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 2257–2268. https://aclanthology.org/2022.lrec-1.242/.

Bram Willemsen, Livia Qian, and Gabriel Skantze. 2023. Resolving references in visually-grounded dialogue via text generation. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czechia, pages 457–469. https://aclanthology.org/2023.sigdial-1.43.

Biographical sketch

Bram Willemsen is a PhD student at KTH Royal Institute of Technology at the Division of Speech, Music and

Hearing (TMH) working towards visually grounded language understanding for conversational agents in the context of the WASP-funded RoboGround project. Before starting his doctoral studies in Sweden in 2019, he studied and worked at Tilburg University, completing his MSc degree (cum laude) in 2017 and working as a Junior Researcher (full-time) on the Horizon 2020-funded L2TOR project until 2019. He enjoys thought-provoking discussions that induce existential dread and long walks on the beach.