The VDG Challenge: Response Generation and Evaluation in Collaborative Visual Dialogue

Nikolai Ilinykh and **Simon Dobnik**
Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science (FLoV),
University of Gothenburg, Sweden
{nikolai.ilinykh,simon.dobnik}@gu.se

Abstract

We propose the VDG Challenge: a shared task that addresses and benchmarks the task of utterance generation in collaborative visual dialogue. The task features two challenging datasets, an evaluation protocol and a tentative schedule. Our shared task will allow researchers to unravel problems of modelling multi-modal interaction and fit of the existing approaches in the NLP and NLG communities.

1 Introduction

In the early 70s, the SHRDLU system (Winograd, 1971) was a revolutionary development. Many believed it had the ability to understand human language, as it was able to describe what it saw in an environment based on human queries. However, the illusion of intelligence of SHRDLU was dispelled as it became clear that the system did not know, for example, what a "box" is. Fast forward to today's world, we have made significant progress with more advanced language models. For instance, models like ChatGPT (OpenAI, 2023) and various Transformer-based architectures for language understanding and generation (Devlin et al., 2019; Brown et al., 2020) have demonstrated the ability to understand language to some extent. Moreover, integrating language with other modalities has become essential in improving these models and making them more applicable to real-world scenarios (Bender and Koller, 2020). In fact, language-and-vision research has been making steps in this direction (Bernardi et al., 2016), as its aim is to build language systems that can map language with the world (Wittgenstein, 1953).

Despite the progress made, research efforts in multi-modal NLP have primarily concentrated on more specific tasks, such as referring expression generation (Krahmer and van Deemter, 2012), image description generation (Karpathy and Fei-Fei, 2017), and visual question answering (Antol et al., 2015). This focus is understandable because modelling human-human or human-world interaction is a challenging task, and there are several reasons for this. Firstly, human language use is highly dynamic, adaptive, and difficult to represent computationally. Successful communication relies on understanding the context, which can encompass textual and multi-modal information, as well as grasping the speaker's intentions. Deciding what to say, how to say it, and when to say it are fundamental aspects of effective communication that require a nuanced understanding of language and context. Secondly, building a robust model requires high-quality dialogue data, which is challenging to gather and ensure that it possesses the properties observable in human language use. Simply put, there are multiple components involved in modelling human dialogue, and they must all be integrated harmoniously to create a truly effective conversational model.

Our proposal, the **V**isual **D**ialogue **G**eneration (VDG) challenge, aims to create a platform that addresses the challenges in modelling multi-modal human-like situated dialogue (Clark et al., 1991). Specifically, our setup revolves around a collaborative visual dialogue, where two participants are placed in an environment with individual visual scenes and are asked to solve a specific task through language interaction. Within this setup, we focus on a particular task of **next utterance generation**, which is part of a broader communicative context. The primary goal of the challenge is to build and evaluate (neural) modelling proposals that can generate better responses given specific contexts. These contexts are defined as sets of previously generated utterances and visual scenes that collectively form a single language game. Each language game may serve a different purpose, such as describing, asking, or clarifying. Importantly, our

aim is not to build a conversational agent capable of holding a full-scale dialogue with a human. Instead, we narrow our focus to a single step: generating a response given a particular situation. By doing so, we can concentrate on examining the quality and value of the generated texts, which is important for building a model of conversation.

The challenge will use two datasets: the Cups (Dobnik et al., 2020) and MeetUp (Ilinykh et al., 2019), both of which are multi-modal and rich in various dialogue phenomena, such as clarification requests and turn-taking, crucial for a complete collaborative process (Clark and Wilkes-Gibbs, 1986). In addition, Cups corpus has data in two languages, English and Swedish. While there have been a few proposed visual dialogue models and datasets (Das et al., 2017; de Vries et al., 2017), they suffer from rigidity and a lack of many phenomena frequently observed in natural human dialogues. Our proposal aims to learn from better high-quality dialogue data, even though the datasets are relatively small in size. As a result, this challenge specifically focuses on transfer learning, learning from small data, and benchmarking the ability of existing generative models to generate responses in human-like multi-modal dialogues. An important feature of our data is that the dialogues were produced with specific (and different) tasks in mind, resulting in high-quality interactions. This raises questions about how much interactive knowledge is shared between different contexts and domains, and how much of it is specific to certain situations. Additionally, we aim to learn from the Natural Language Generation (NLG) community about the challenges and issues that arise when building generative multi-modal models, including biases, ethical concerns, and the naturalness of generated responses.

2 Datasets

Both Cups and MeetUp were collected in a task-oriented setting. In Cups (Figure 1), two participants were asked to locate missing cups on a table in a virtually generated scene. It is worth noting that the cups missing for each participant were not necessarily the same ones. These cups varied in colour, type, and location, and each participant could only see a subset of them from a different view. To communicate and identify each other's missing cups, participants used the chat interface. Importantly, there were no restrictions on how the task should be approached, allowing participants

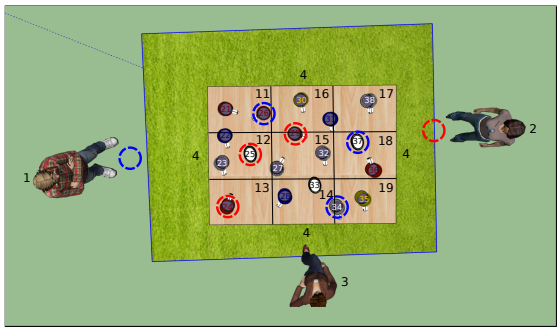

Figure 1: The Cups corpus: an allocentric view of the scene with annotated objects. Participants (labelled as "1" and "2") cannot see objects marked with their colour (either red or blue). Katie (labelled as "3") is a passive observer of the scene.

the freedom to choose their strategies. For additional scenes from the Cups corpus, refer to Appendix A.

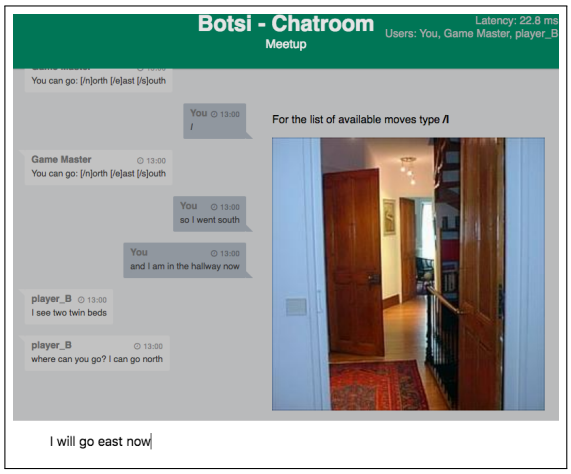

Figure 2: The MeetUp game interface. A view for player A is presented. The image on the right side changes if the player moves to a different room.

In MeetUp (Figure 2), participants are randomly placed in a room, which is shown as a real image. This room is part of a larger house area, consisting of connected real images. Participants are informed that they can move around the house by executing specific actions, for example, using "/s" to move south from the current room shown in the image. Through navigation and interaction in the chat interface, participants are required to ensure that they end up in the same room, where they both see an identical image. It is worth noting that the house layouts often contained multiple rooms of the same type, and participants were frequently asked to meet in a room of this specific type. Similar to Cups, there were no restrictions placed on

the participants regarding how they should solve the task.

3 Task description

The main objective of this challenge is to benchmark and evaluate generative models in the context of multi-modal dialogues. Specifically, the task is to generate the utterance $\mathbf{u}_i$ given the context $\mathbf{C}$, where i represents the utterance number. The context $\mathbf{C}$ is formed by the dialogue history and visual scene(s). An important question arises: how effectively can the models utilise the context for generating the next utterance? To address this, we first split each dialogue in the dataset into conversational games (Dobnik and Storckenfeldt, 2018). Each game corresponds to a specific sub-task such as asking for more information or describing a scene. In the Cups corpus, game annotations for dialogues in Swedish are already available, and we plan to prepare annotations for dialogues in English and MeetUp dialogues as well. These conversational games can be seen as mini-contexts that help achieve a particular goal (Kowtko et al., 1991). The sentences we select belong to the context of a specific conversational game $\mathbf{C}_L$, which, in turn, is part of a broader dialogue context $\mathbf{C}_D$. We provide two types of contexts: (i) conversational game context $\mathbf{C}_L$: this includes the utterances preceding the target utterance $\mathbf{u}_i$ within a single conversational game, and (ii) dialogue context $\mathbf{C}_D$: this provides information on how the target utterance fits into the overall context of all other utterances and games. It is important to include the full dialogue context, as the flow of dialogues is not always linear; sometimes, a single conversational game can be embedded within another game. We emphasise that a single utterance can often be just a fragment of a broader set of utterances that together convey a specific idea.

3.1 State of the data and statistics

The challenge will be conducted on the data available in public repositories. The MeetUp repository [1] contains 430 dialogues, where each dialogue is a sequence of events. An event can be a message from either a bot or a player. This can also be a navigation action executed by one of the players. Each valid navigation action changes position of the player in the house resulting in change of the scenery that the player sees. The images of house environments are taken from the ADE20k corpus (Zhou et al., 2017) and can be referred back to it. The MeetUp dialogues have on average 13.2 turns per dialogue, with each turn consisting on average of 5.1 tokens. There are 28.3 navigation actions performed on average per dialogue which means that there are approximately 2 moves per message. There are a few instances in the dataset when a single participant played the game multiple times, e.g. one worker participated in the game 49 times. Novice players played with each other only in 22 games. This information can be potentially useful for modelling because, participants adopt and change their strategy based on the familiarity with the game and they carry some of that knowledge to new conversations[2].

The Cups corpus[3] consists of dialogues and corresponding individual static views of the same scene. We provide the views for each participant along with the ground-truth top-down view of the scene with no missing objects. We will also provide files with bounding box annotations of object ids as shown in Figure 1. The textual part of the dataset includes annotations of turns, dialogue acts, frame of reference (FoR), repair, and dialogue games (Swedish only) with the goal of capturing situated collaborative referring (Dobnik et al., 2015). Cups also contains annotations of reference and coreference to scene entities Dobnik and Silfversparre (2021) using the CoNLL 2011/2011 annotation scheme (Pradhan et al., 2011). The Cups dataset has a fewer dialogues: 2 dialogues in English and 6 dialogues in Swedish. However, as they can take over an hour they are much longer and are structured in more dialogues games than MeetUp dialogues. There are on average 299 turns per dialogue in the English data and 171 turn per dialogue in the Swedish data.

The information on downloading the data will be available to the participants. We plan to complete annotations of dialogues with conversational games before INLG 2023. In terms of the splits, we are planning to follow the standard 80/10/10 split for training, validation and test data. Note that these splits are not for dialogues themselves,

[1] https://github.com/clp-research/meetup

[2] Based on our observations expert players tend to produce fewer messages, instead relying on the strategy of asking the other player to stay in the room and describe it, while they are looking for it.

[3] https://github.com/sdobnik/cups-corpus

but for target utterances **U**, which are part of annotated conversational games. Our datasets are relatively small. However, they contain rich natural interaction data (rather than short crowd-sourced interactions or artificially generated dialogue data found in some popular datasets). We would like to encourage challenge participants to exploit the possibility of applying transfer learning by training the multi-modal dialogue model first on the other (larger) datasets (Zhang et al., 2018; Galetzka et al., 2020) and then fine-tuning them on our data and evaluate the possibility of such transfer as well as compare the datasets with each other.

Visual dialogue is a task that has previously been addressed in the Visual Dialog Challenge[4] where the goal is to answer a question given an image and a dialogue history. The challenge has attracted several submissions and has been conducted three times. However, the data used in this challenge lacks several linguistic phenomena found in Cups and MeetUp (Byron, 2003) which go beyond simple question-answer pairs (Das et al., 2017; Dong et al., 2021). In appendix B we provide a linguistic analysis of dialogues from both Cups and MeetUp and demonstrate that complexity and richness of dialogue phenomena found in our data.

4 Evaluation campaign

Table 1 presents the preliminary schedule for the proposed challenge. Initially, we will provide a description of the available infrastructure, which will serve as the hosting server for managing system submissions. Participants are expected to adhere to these requirements, and they should specify the use of GPUs, external APIs, and other components in their systems. For the submission and review of papers, we will use the OpenReview platform. For evaluating the generated responses, we will compare the outputs of each model against a held-out test set, using various metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015), BERTScore (Zhang et al., 2020), and BARTScore (Yuan et al., 2021).

We will conduct a human evaluation and develop an evaluation protocol for the system submissions. To collect human judgments, we will leverage Amazon Mechanical Turk, and we might also explore using Prolific as an alternative platform, which could potentially provide more qualified human crowd-workers[5]. The human evaluation procedure will be modelled after the one used in the WebNLG+ 2020 challenge (Castro Ferreira et al., 2020), given that one of the authors of this proposal has experience in running human evaluation. For the human evaluation, crowd-workers will assess the quality of the generated responses based on the dialogue history and visual history (in the case of MeetUp) or scene views (in the case of Cups). These evaluations may be compared against the ground-truth human responses. To rank the responses, a Likert scale from 1 to 5 will be employed, where a higher score indicates better quality of the generated response. In human evaluation we will focus on several aspects including:

1. **Relevance**: Does the response align with the available history of linguistic interaction between participants? Does the utterance sound like something a person would say? We refer to this criterion as **local relevance**. Additionally, we will consider context provided after the generated response (from the ground-truth dialogue) and ask human evaluators to assess if the response fits the overall topic of conversation (**global relevance**).

2. **Usefulness**: Does the response contribute to solving the task? Can people identify the visual elements that the utterance refers to? We will present human crowd-workers with examples of situations where a response is relevant but not useful, and vice versa. The aim is to measure the degree of informativeness of responses, considering the nature of the task.

3. **Correctness**: Is the response well-structured, grammatical, and written in fluent language?

Additionally, we will ask human evaluators to perform fuzzy matching of the generated utterances with the ground-truth responses. This approach takes into consideration that the system's output might not be entirely relevant locally but could still be relevant globally. By doing so, the models will not be penalised by automatic metrics for generating responses that differ from the ground-truth

[4] https://visualdialog.org/challenge/2020

[5] As our focus is not solely on conversational agents, the evaluation will be on assessing the quality of the generated utterances within a given context. Implementing a more sophisticated evaluation setup, where a submitted system actively plays the full game, would require a different type of challenge task.

Period	Phase
September 2023	Announcement at INLG 2023 along with the call for participation. The training and validation data are made available on the challenge website. Release of automatic evaluation scripts. Registration of participants is open.
December 2023	Test data is released, system submission. The baseline model is released along with its results for automatic evaluation.
January 2024	Deadline for system submission.
February 2024	Results of automatic evaluation are announced.
April 2024	Results of human evaluation are announced. Authors are asked to submit their system reports.
May–June 2024	System report reviewing and notification. Camera-ready submission of the system reports.
June 2024	The challenge is completed. Participant reports and challenge report are submitted to INLG 2024 and presented at the conference.

Table 1: Tentative protocol for the challenge. The schedule might change depending on the timeline of INLG 2024.

significantly, as long as they remain relevant to the conversation itself.

To ensure the quality of human evaluation, we will prepare a set of utterances in contexts that clearly represent both low and high points on the Likert scale for each of the aspects mentioned earlier. These examples will be shown to crowdworkers before they begin evaluating the actual outputs of the submitted systems. Conducting a few test rounds for human evaluation will help us understand the workers' performance and the level of guidance they require to perform well in our task. This process will help us build a pool of highly skilled workers who are trained to evaluate challenge submissions[6].

The challenge winners will be selected based on multiple criteria. Instead of focusing solely on models that perform well overall, we will also consider models that excel in specific tasks. For instance, we will look for better transfer learning approaches, multi-lingual models, or uni-modal approaches that perform well across various metrics. By examining individual properties of the submitted systems, we aim to document and benchmark the task of utterance generation in visual dialogue from multiple perspectives.

5 Conclusion

We present the VDG Challenge as a platform to advance research in grounded situated dialogue. We believe that the task of generating the next utterance in collaborative visual dialogue holds significant value for the NLG (Natural Language Generation) community, especially considering the remarkable performance and attention achieved by large language models in the NLP field. Our primary objective is to establish a comprehensive task benchmark, and as such, we welcome novel ideas for multi-modal dialogue modelling. We would be delighted to host the challenge at INLG 2024.

[6]We will also recruit workers from external websites such as https://www.mturkcrowd.com.

Acknowledgments

The challenge that is described in this paper will be 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. We would like to thank reviewers for their valuable and insightful comments.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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*, pages 5185–5198, Online. Association for Computational Linguistics.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.*, 55(1):409–442.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Donna K Byron. 2003. Understanding referring expressions in situated language some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for Real World Interaction*, pages 39–47.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Herbert H Clark, Susan E Brennan, et al. 1991. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 326–335.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, Gothenburg, Sweden. SEMDIAL.

Simon Dobnik, John D. Kelleher, and Christine Howes. 2020. Local alignment of frame of reference assignment in english and swedish dialogue. In *Spatial Cognition XII*, pages 251–267, Cham. Springer International Publishing.

Simon Dobnik and Vera Silfversparre. 2021. The red cup on the left: Reference, coreference and attention in visual dialogue. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Potsdam, Germany. SEMDIAL.

Simon Dobnik and Axel Storckenfeldt. 2018. Categorisation of conversational games in free dialogue over spatial scenes. In *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Aix-en-Provence, France. SEMDIAL.

Tianai Dong, Alberto Testoni, Luciana Benotti, and Raffaella Bernardi. 2021. Visually grounded follow-up questions: a dataset of spatial questions which require dialogue history. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 22–31, Online. Association for Computational Linguistics.

Fabian Galetzka, Chukwuemeka Uchenna Eneh, and David Schlangen. 2020. A corpus of controlled opinionated and knowledgeable movie discussions for training neural conversation models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 565–573, Marseille, France. European Language Resources Association.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meet up! a corpus of joint activity dialogues in a visual environment. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, London, United Kingdom. SEMDIAL.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676.

Jacqueline C. Kowtko, Stephen D. Isard, and Gavin Doherty. 1991. Conversational games within dialogue. In *Proceedings of the ESPRIT Workshop on Discourse Coherence*, University of Edinburgh.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

OpenAI. 2023. Chatgpt: Large-scale unsupervised language modeling for open-domain conversational agents. https://openai.com/blog/chatgpt. Accessed on June 9, 2023.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5503–5512.

Terry Winograd. 1971. *Procedures as a representation for data in a computer program for understanding natural language / by Terry Winograd*. Massachusetts Institute of Technology Cambridge.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ADE20K dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5122–5130. IEEE Computer Society.

A Cups corpus: individual views

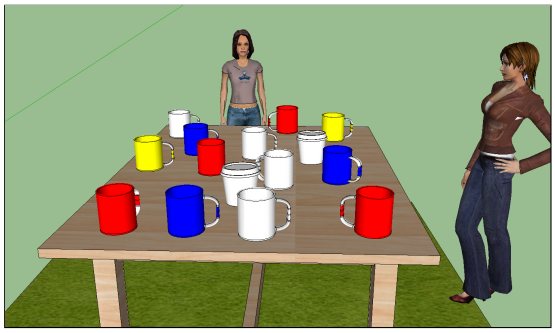

Figure 3: Individual scene view of the speaker 1 from Figure 1.

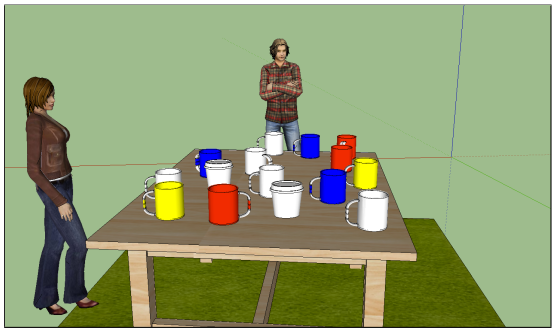

Figure 4: Individual scene view of the speaker 2 from Figure 1.

B MeetUp and Cups examples

Here is an example from the MeetUp corpus:

(1)
a. Game Master: You have to meet in a room of type *utility room*.
b. A: Hi. I'm in a bedroom with pink walls.
c. B: I seem to be in a kitchen.
d. A: I'll go look for a utility room.
e. A (privately): *north*
f. A (privately): *west*
g. B (privately): *east*
h. A: Found a room with a washing machine. Is that a utility room?
i. B: Was wondering as well. Probably that's what it is.
j. B: I'm in the pink bedroom now. I'll come to you.
k. B (privately): *north*
l. B (privately): *west*
m. B: Poster above washing machine?
n. A: Mine has a mirror on the wall.
o. B: yeah, could be mirror. Plastic chair?
p. A: And laundry basket.
q. A: *done*
r. B: Same
s. B: *done*

MeetUp has many high-level descriptions of the scene ((1-b)–(1-c)), which require understanding of

room types. Strategy coordination ((1-d)–(1-j)) and revisions of previously made references ("poster" to "mirror" in (1-m)–(1-o)) occur as well. There is also a need for memory (1-j), which is important for co-reference. Another property of the dataset is the presence of indirect dialogue acts ((1-p)–(1-o)).

Below is an excerpt from the Cups corpus:

(2)
a. A: i see lots of cups and containers on the table
b. B: me too
c. A: some white, some red, some yellow, some blue
d. B: I see six white ones
e. A: i see seven
f. A: but maybe we should move in one direction...
g. B: ok, lets do that
h. A: shall we take it from katies point of view?
i. B: ok
j. . . .
k. B: so what do you see in the "second row" from my perspective?
l. A: i see red, then space, then white and blue (same as katie's)
m. A: no yellow
n. B: is it on the edge of the table?
o. B: on your left
p. A: ok, yes!

We observe reference to the same (or different!) objects using attributes such as colour and identification of object mismatch, e.g. (2-a)–(2-e). In (2-f)–(2-i) participants negotiate interactive strategy. Adjusting a perspective (or frame of reference) for spatial relations is also important in dialogue games, e.g. (2-k)–(2-p).