Building Common Ground in Visual Dialogue: The PhotoBook Task and Dataset

Janosch Haber, **Elia Bruni**, and **Raquel Fernández**
Institute of Logic, Language and Computation, University of Amsterdam
{janoschhaber|elia.bruni}@gmail.com
raquel.fernandez@uva.nl

1 Introduction

The past few years have seen an increasing interest in developing computational agents for visually-grounded dialogue, the task of using natural language interaction to communicate about visual content. Current challenges include posing and answering questions about a visual scene (Das et al., 2017a; Das et al., 2017b) or about specific objects in it (De Vries et al., 2017). While these tasks and associated datasets provide a useful starting point to develop multimodal dialogue agents, they have several shortcomings regarding their dialogical properties: (i) the interaction consists of questions followed by answers, which makes the exchanges closer to visual question-answering (Antol et al., 2015) than to dialogue proper, where different dialogues acts can take place; (ii) the tasks are asymmetric: each agent has a predefined role (e.g., questioner or answerer), which determines their contribution in the conversation; and (iii) there are limited opportunities to model how agents accumulate shared information (*common ground* (Stalnaker, 1978; Clark, 1996)) about the visual content they discuss. We present ongoing work on a novel symmetric dialogue setting, the PhotoBook Task, which elicits dialogues that provide rich data for investigating and learning common ground and partner-specific dialogue features in visual environments.

2 The PhotoBook Task

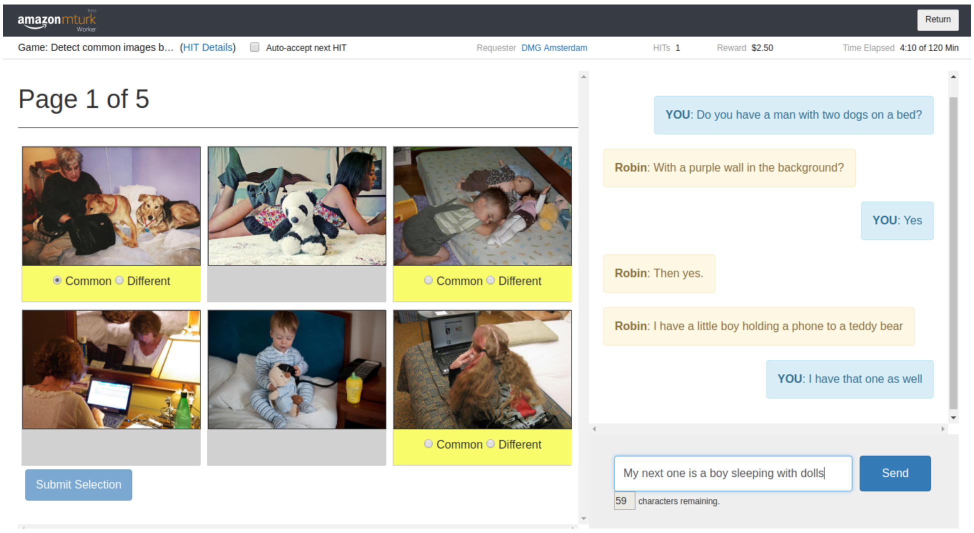
The setup of the PhotoBook Task takes inspiration from experimental paradigms that have been extensively tested within the psycholinguistics literature to investigate partner-specific effects (see (Brown-Schmidt et al., 2015) for an overview). This seminal research has shown that when speakers interact, they typically develop shared ways of referring to entities, which become shorter and more opaque to others over time (Clark and Wilkes-Gibbs, 1986). The key component for eliciting partner-specific effects of this kind is to set up the data collection in such a way that each participant performs a task multiple times with the same partner, building up shared common ground as a result of their interaction history. We incorporate this component in the design of the task to crowdsource the collection of the first large-scale dataset with these features.

In the PhotoBook Task two participants are paired for a conversation game consisting of five rounds. In each round, the participants are shown a set of six similar images, resembling a page of a photo book (see Figure 1). They are then asked to determine which of three highlighted images are shown to both of them by communicating through a text-only chat interface. When all indicated images are marked as either *common* or *different*, the participants are shown a feedback screen and proceed to the next round. During later rounds of the game, a selection of previously displayed images will be visible again, prompting participants to refer to those images based on their visual context as well as previously established referring expressions. Dialogue data collected through the PhotoBook task therefore allows for tracking the evolving common ground between participants.

3 The PhotoBook Dataset

The PhotoBook Task was implemented in the Facebook ParlAI dialogue agent framework (Miller et al., 2017). Up to date, we recorded more than 2,500 games containing 5 dialogue rounds each, stemming from over 1,500 unique participants on crowdsourcing platform Amazon Mechanical Turk. The

Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue, November 8-10, 2018, Aix-en-Provence, France.



Game: Detect common images b... (HIT Details) □ Auto-accept next HIT   Requester DMG Amsterdam   HITs 1   Reward $2.50   Time Elapsed 4:10 of 120 Min

Page 1 of 5

● Common ○ Different

○ Common ○ Different

○ Common ○ Different

YOU: Do you have a man with two dogs on a bed?

Robin: With a purple wall in the background?

YOU: Yes

Robin: Then yes.

Robin: I have a little boy holding a phone to a teddy bear

YOU: I have that one as well

My next one is a boy sleeping with dolls

59 characters remaining.

Figure 1: Screenshot of the AMT user interface of the PhotoBook Task.

resulting data contains a total of over 160k utterances, 130k actions, spans a vocabulary of close to 12k unique tokens, and exhibits a diversity of dialogue acts. A preliminary analysis also shows that the data displays features similar to those observed by (Krauss and Weinheimer, 1966) and (Clark and Wilkes-Gibbs, 1986) for small-scale experiments run in the lab. Participants become more efficient as the game progresses, as evidenced by a significant decrease in completion times and number of words used across rounds while their task success increases. We also observe a simplification in the image descriptions, resulting in an increase in the relative frequency of nouns, while pronouns, determiners, and verbs are likely to be omitted in later rounds. Consider, for example, the following descriptions used to refer to the bottom right image in Figure 1 over different rounds of a game by participants **A** and **B**:

> **B:** Last is a girl with long hair looking at a laptop **A:** Yes, I have that one
> **B:** The girl with long hair looking at laptop **A:** Nope
> **A:** Girl with long hair? **B:** No, not this time
> **A:** Long hair girl? **B:** I don't have the girl

The PhotoBook Task thus provides a means to collect a large-scale dataset focused on central aspects of goal-oriented dialogue. We believe that this dataset can be a rich new repository for developing artificial agents with more consistent, efficient, and natural dialogue abilities in visual environments.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proc. of ICCV*.

Sarah Brown-Schmidt, Si On Yoon, and Rachel Anna Ryskin. 2015. People as contexts in conversation. In *Psychology of Learning and Motivation*, volume 62, chapter 3, pages 59–99. Elsevier.

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

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual Dialog. In *Proc. of CVPR*.

Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proc. of ICCV*.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

Robert M. Krauss and Sidney Weinheimer. 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4(3):343–346.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A Dialog Research Software Platform. In *Proc. of EMNLP: System Demonstrations*.

Robert Stalnaker. 1978. Assertion. In P. Cole, editor, *Pragmatics*. New York Academic Press.