Visual Dialogue Needs Symmetry, Goals, and Dynamics: The Example of the MeetUp Task

David Schlangen, Nikolai Ilinykh, Sina Zarrieß

Dialogue Systems Group, Bielefeld University, Bielefeld, Germany

1 Introduction

After achieving impressive success representing image content textually (as done by captioning models [1,2,3,4,5]; and referring expression resolution and generation [6,7,8,9]), the Vision and Language community has recently established "Visual Dialogue" as the more challenging follow up task [10,11]. In that task, a Questioner, prompted by some textual information (a caption) can ask an Answerer questions about an image that only the latter sees. We argue here that this setup leads to an impoverished form of dialogue and hence to data that is not substantially more informative than captioning data, if the goal is to model visual *dialogue*. We describe our ongoing work on the MeetUp setting, where two players navigate separately through a visually represented environment, with the goal of being at the same location. This goal gives them a reason to describe visual content, leading to motivated descriptions, and the dynamic setting induces an interesting split between private and shared information.

2 Visual Dialogue

Fig. 1. The Visual Dialogue Collection Task and an Example Dialogue (from [10])

Figure 1 shows the environment in which the visual dialogue dataset [10] was collected. As the example dialogue on the right indicates, this rather artificial setting ("you have to ask questions about the image") seem to encourage a pairwise structuring of question and answer. That the string of pairs forms a dialogue is only recognisable in the fact that each pair concerns a different aspect of the image, and that later questions may refer to entities previously mentioned. Since there is no way for the questioner to provide feedback on the answers, it is unlikely that a model could learn from data of this type that dialogue is more than a sequence of loosely related question/answer pairs, and that even such sequences typically would have structure in human dialogue. (For reasons of space, we cannot argue this point more deeply here.)

3 The MeetUp Task

In contrast, we designed the MeetUp task to elicit more structured dialogue. The task is based on a dynamic environment with several "rooms" (in the instantiation presented here, represented as images) where two dialogue participants (players) are placed in different rooms and have to find each other. As the players cannot see each other, but can communicate (via text messages), the only way they can solve the task is to establish verbally whether they both currently see the same room/image.

Our set-up extends recent efforts along the following dimensions: 1) the task's main goal can be defined independently of reference, in high-level communicative terms (namely "try to meet up in an unknown environment"), 2) the task is symmetric and does not need a rigid interaction protocol (there is no instruction giver/follower), 3) there is a clear division between private information (that only one player has access to) and public information (facts that have been publicly asserted), and reaching the goal involves moving information from the former state to the latter (i.e., it involves *conversational grounding* [12]), 4) reference can be made to things not currently seen, if they have been introduced into the discourse earlier (see line 59, "I found the kitchen"). We have conducted a pilot data collection which indicates that this setting indeed leads to interesting dialogues. We aim to collect a sufficient number of dialogues (in the thousands) in the upcoming weeks, in order to be able to train agents on this task.

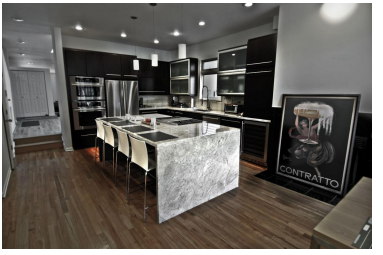

Fig. 2. The scene discussed in the excerpt below

	Time	Private to A	Public	Private to B
31	(01:45)		A: I am now in a kitchen with **wood floors** and **a poster** that says **CONTRATTO**	
			....	
59	(02:50)		B: Wait– I found the kitchen!	
			....	
60	(02:55)	$\xrightarrow{N}$ kitchen		
61	(02:55)	You can go [/n]orth [/e]ast [/s]outh [/w]est		
62	(03:13)		A: I am back in kitchen. It has **a white marble dining table** in center	
63	(03:29)		B: Yes. There are **four chairs** on **the island**.	
64	(03:35)		A: Exactly	
65	(03:37)		B: And **the big Contratto poster**.	
66	(03:48)		B: **Three lights** above **the island**?	
67	(03:53)		A: yep	
71	(04:05)			B: /done
72	(04:07)	A: /done		
73	(04:10)		Well done! You are all indeed in the same room!	

Table 1. (Discontinuous) excerpt from a MeetUp dialogue

References

1. Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J., Zitnick, L., Zweig, G.: From captions to visual concepts and back. In: Proceedings of CVPR, Boston, MA, USA, IEEE (June 2015)
2. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: Language models for image captioning: The quirks and what works. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China, Association for Computational Linguistics (July 2015) 100–105
3. Chen, X., Lawrence Zitnick, C.: Mind's eye: A recurrent visual representation for image caption generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2422–2431
4. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Computer Vision and Pattern Recognition. (2015)
5. Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B.: Automatic description generation from images: A survey of models, datasets, and evaluation measures. J. Artif. Int. Res. **55**(1) (January 2016) 409–442
6. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: ReferItGame: Referring to Objects in Photographs of Natural Scenes. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, Qatar (2014) 787–798
7. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. CoRR **abs/1511.02283** (2015)
8. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L. In: Modeling Context in Referring Expressions. Springer International Publishing, Cham (2016) 69–85
9. Schlangen, D., Zarriess, S., Kennington, C.: Resolving references to objects in photographs using the words-as-classifiers model. In: Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016). (2016)
10. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2017)
11. De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.: Guesswhat?! visual object discovery through multi-modal dialogue. In: Proc. of CVPR. (2017)
12. Clark, H.H.: Using language. 1996. Cambridge University Press: Cambridge (1996) 274–296