Effective questions in referential visual dialogue

Mauricio Mazuecos (1), Alberto Testoni (2), Raffaella Bernardi (3) and Luciana Benotti (1)
(1) FAMAF, Universidad Nacional de Córdoba, CONICET, Argentina
(2) DISI, University of Trento, Italy
(3) DISI, CIMEC, University of Trento, Italy
mmazuecos@famaf.unc.edu.ar alberto.testoni@unitn.it
raffaella.bernardi@unitn.it luciana.benotti@unc.edu.ar

Abstract

An interesting challenge for situated dialogue systems is referential visual dialogue: by asking questions, the system has to identify the referent to which the user refers to. Task success is the standard metric used to evaluate these systems. However, it does not consider how *effective* each question is, that is how much each question contributes to the goal. We propose a new metric, that measures question effectiveness. As a preliminary study, we report the new metric for state of the art publicly available models on GuessWhat?!. Surprisingly, successful dialogues do not have a higher percentage of effective questions than failed dialogues. This suggests that a system with high task success is not necessarily one that generates good questions.

1 Introduction

GuessWhat?! (de Vries et al., 2017) is a cooperative two-player referential visual dialogue game. One player (the *Oracle*) is assigned a referent object in an image, the other player (the *Questioner*) has to guess the referent by asking yes/no questions. The GuessWhat?! dataset contains games of different complexity, ranging from easy images with a referent and 1 distractor to images with 19 distractors.

Referential visual dialogue has a clear task success metric: whether the Questioner is able or not to correctly identify the referent at the end of the dialogue. The need of going beyond this metric to evaluate the quality of the dialogues has already been observed. So far attention has been put on the linguistic skills of the models (Shukla et al., 2019; Shekhar et al., 2019) and their dialogue strategies (Abbasnejad et al., 2018; Shekhar et al., 2018). Recently, Sankar et al. (2019) showed that current SOTA dialogue systems do not take dialogue history into account, and new models were proposed to make questions more informative and consistent with the dialogue history (Shukla et al., 2019; Ray et al., 2019; Abbasnejad et al., 2019; Pang and Wang, 2020). But still the models are mostly evaluated without considering how much each question contributes to the goal. We propose a new metric to evaluate dialogue *effectiveness* as the percentage of effective questions it contains. Intuitively, a question is effective if it eliminates at least one possible distractor from the set of objects (Krahmer and van Deemter, 2012). Figure 1 gives a game played by humans as an example. In the image there are 8 candidate objects: the referent object is the cow marked in green and the distractors are the other 6 cows and the wooden stick. The dialogue is highly effective: 80% of the questions eliminate at least one distractor.

Human question	Answer	# D	Effective
1. is it a cow?	yes	6	True
2. is it the big cow in the middle?	no	5	True
3. a cow on the left?	no	3	True
4. *on the right?*	yes	3	False
5. first cow near us?	yes	0	True

Figure 1: Human-human dialogue on the Guesswhat?! referential task extracted from (de Vries et al., 2017). The target is highlighted in green. # D is the number of candidates remaining after the question is answered. Four out of five questions eliminate distractors and, hence, are effective.

2 Previous work

Despite recent progress in the area of vision and language, recent work (Jain et al., 2019) in the navigation task (VLN) argues that current research leaves unclear how much of a role language plays in this task. They point out that dominant evaluation metrics have focused on goal completion rather than how each action contributes to the goal (Anderson et al., 2018). The nature of the path an agent takes, however, is of clear practical importance: it is undesirable for any robotic agent in the physical world to reach the destination by taking a lot of deviation or getting into dangerous zones. Jain et al. (2019) propose alternative metrics that evaluate the intermediate steps in the VLN task.

As argued by Lowe et al. (2019), the vast majority of recent papers on emergent communication show that adding a communication channel leads to an increase in task success. This is a useful indicator, but provides only a coarse measure of the agent's learned communication abilities. As we move towards more complex environments, it becomes imperative to have a set of finer tools that allow qualitative and quantitative insights into the emergence of communication.

Following this idea of not only focusing on goal completion but on evaluating how much each step contributes to the goal, in this paper we propose a new metric for referential dialogue. We agree with Thomason et al. (2019) that incremental evaluation metrics such as ours should look further back into the dialogue history. We believe that language and vision systems should also be evaluated on aspects such as grammatically, truthfulness, diversity and other aspects as done in previous work (Lee et al., 2018; Ray et al., 2019; Xie et al., 2020; Murahari et al., 2019). In this paper we focus on whether a question is effective considering the dialogue history and the visual context.

One of the motivations for referential visual dialogue is to provide robots with the ability to identify objects through dialogue with a humans. The task we address in this paper is a simplification. In our setup, the view of the robot is static (i.e. a picture). For our work we use the GuessWhat?! dataset (de Vries et al., 2017). We are particularly interested in models that generate questions explicitly modelling the dialogue history (Zhang et al., 2018; Shukla et al., 2019; Pang and Wang, 2020).[1]

3 Effective questions

Our definition of *effective* question is based on the set of candidate objects: the *reference set* RS. We compute RS for each question q_t. The reference set before the dialogue starts, $RS(q_0)$, contains all the objects in the image. At each dialogue turn t, $RS(q_t)$ is defined as the set of objects in $RS(q_{t-1})$ such that the answer[2] to q_t on those objects is the same than the answer to q_t on the referent r. Formally:

$$RS(q_t) := \{o_i \in RS(q_{t-1}) \mid Answer(q_t, o_i) = Answer(q_t, r)\}$$

We say that a question q_t is *not effective* iff $RS(q_t) = RS(q_{t-1})$. That is, the question does exclude any distractor. The *effectiveness* of the dialogue is given by the percentage of effective questions it has.

Table 1 reports the average *effectiveness* (Global column) for humans and SOTA models for which either the code or the dialogues with suitable annotations have been released. We also distinguish the effectiveness of dialogues finished in either Failure or Success. The baseline model (de Vries et al., 2017) represents the Questioner as two independent models, the question generator and the guesser, and train them by supervised learning. RL (Strub et al., 2017) further trains this baseline with a reinforcement learning phase. GDSE-SL differs from the baseline by having a joint encoder for the Questioner components and GDSE-CL exploits this joint architecture by letting the two components cooperate with each other (Shekhar et al., 2019). Last, VDST (Pang and Wang, 2020) extends the questioner with a probability distribution of each object being the referent and trains with reinforcement learning.

The results suggest that models make more non-effective questions than one may expect. Surprisingly, successful dialogues generated by models do not have a higher percentage of effective questions. Even for humans, effectiveness is not considerably higher for successful dialogues. Human effectiveness is

[1] Unfortunately, the code or test dialogues of some previous work are not available (Zhang et al., 2018; Shukla et al., 2019).

[2] Answers are provided by the Oracle model proposed in (de Vries et al., 2017) whose accuracy on the test set is 79%. As initial reference set, we take the list of objects annotated in the dataset (de Vries et al., 2017).

VDST		**GDSE-CL**	
1. is it food?	yes	1. is it food?	yes
2. is it in the left?	yes	2. *is it a cake?*	yes
3. is it in the front?	yes	3. *is it the dark brown?*	yes
4. *is it in the top?*	no	4. *is it the entire cake?*	yes
5. *in the middle?*	no	5. so the most left of the brown ones?	yes

Figure 2: Dialogues generated by VDST and GDSE-CL in a successful game. Non effective in italics.

higher in almost every column of the table, the VDST model is close. Humans do not see the list of annotated objects as the Guesser models do. They rely on their sight on the image and they may ask questions that discard objects present in the image but not annotated in the dataset and hence not part of the reference set we calculate. All of these questions are marked as non-effective because they discard objects invisible to our metric and to the models. Hence, human effectiveness could be higher than we have calculated using the GuessWhat?! dataset object annotations.

Model	Max Qs	Task success	Effectiveness		
			Global	Failure	Success
Baseline (de Vries et al., 2017)	8	40.7	26.4	27.5	24.7
GDSE-SL (Shekhar et al., 2019)	8	49.7	29.1	31.4	26.9
RL (Strub et al., 2017)	8	56.3	32.6	36.5	29.6
GDSE-CL (Shekhar et al., 2019)	8	58.4	30.2	32.3	28.6
Baseline (de Vries et al., 2017)	5	40.8	38.8	39.8	37.4
GDSE-SL (Shekhar et al., 2019)	5	47.8	42.2	44.6	39.9
RL (Strub et al., 2017)	5	58.4	48.6	52.9	45.1
GDSE-CL (Shekhar et al., 2019)	5	53.7	44.7	47.8	42.6
VDST (Pang and Wang, 2020)	5	64.4	52.9	57.4	51.0
Humans (de Vries et al., 2017)	∞	84.1	56.9	54.7	57.3

Table 1: Results comparing task success and effectiveness of generative systems for unseen images. Our manual inspections of human dialogues has shown that humans ask non-effective questions mostly at the end of the dialogue to reinforce their belief before guessing. We only show results of the 5 questions setup for VDST as we only had access to those dialogues.

Figure 2 shows an example of both metrics on a game on which VDST and GDSE-CL are successful. Effectiveness is 60 for VDST and 40 for GDSE-CL. Our definition of effectiveness not only accounts for question repetitions, but it also captures paraphrases and context-dependent redundancies. Examples of context dependent redundancy can be seen for both systems. In the VDST dialogue, 4 is redundant because, in this image, there is no cake that is both in the front and in the top. In GDSE-CL dialogue, question 2 is redundant because all cakes in the image are dark brown.

4 Conclusion and future work

We proposed a new metric for evaluating Guesswhat?! dialogues. Effectiveness, as we defined it, evaluates whether the question can rule out at least one possible distractor. We consider a question to be effective if it is able to make the reference set smaller. We observe that effectiveness decreases as dialogues advance and reaches its lowest level in the last turn. We also find that successful dialogues do not have a higher percentage of effective questions. This is surprising, and hints at the fact that there are other strategies to accomplish reference identification other than asking effective questions. We believe that our metric could be a heuristic that guides the training of end-to-end models.

References

Ehsan Abbasnejad, Qi Wu, Iman Abbasnejad, Javen Shi, and Anton van den Hengel. 2018. An active information seeking model for goal-oriented vision-and-language tasks. *CoRR*, abs/1812.06398.

Ehsan Abbasnejad, Qi Wu, Javen Shi, and Anton van den Hengel. 2019. Whats to know? uncertainty as a guide to asking goal-oriented questions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peter Anderson, Angel X. Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir Roshan Zamir. 2018. On evaluation of embodied navigation agents. *CoRR*, abs/1807.06757.

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

Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1862–1872, Florence, Italy, July. Association for Computational Linguistics.

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

Sang-Woo Lee, Yu-Jung Heo, and Byoung-Tak Zhang. 2018. Answerer in questioner's mind: Information theoretic approach to goal-oriented visual dialog. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2579–2589. Curran Associates, Inc.

Ryan Lowe, Jakob Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. 2019. On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 693–701, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Vishvak Murahari, Prithvijit Chattopadhyay, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Improving generative visual dialog by answering diverse questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1449–1454, Hong Kong, China, November. Association for Computational Linguistics.

Wei Pang and Xiaojie Wang. 2020. Visual dialogue state tracking for question generation. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.

Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. 2019. Sunny and dark outside?! improving answer consistency in VQA through entailed question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5860–5865, Hong Kong, China, November. Association for Computational Linguistics.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy, July. Association for Computational Linguistics.

Ravi Shekhar, Tim Baumgärtner, Aashish Venkatesh, Elia Bruni, Raffaella Bernardi, and Raquel Fernandez. 2018. Ask no more: Deciding when to guess in referential visual dialogue. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1218–1233, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat. 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 2578–2587, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, Matthew Turk, and William Yang Wang. 2019. What should I ask? using conversationally informative rewards for goal-oriented visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6442–6451, Florence, Italy, July. Association for Computational Linguistics.

Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron C. Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. In *Conference on Robot Learning*, Osaka, Japan.

Huiyuan Xie, Tom Sherborne, Alexander Kuhnle, and Ann Copestake. 2020. Going beneath the surface: Evaluating image captioning for grammaticality, truthfulness and diversity. In *Workshop on Evaluating AI Systems (AAAI 2020)*.

Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton van den Hengel. 2018. Goal-oriented visual question generation via intermediate rewards. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 189–204. Springer.