On the role of effective and referring questions in GuessWhat?!

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

Task success is the standard metric used to evaluate referential visual dialogue systems. In this paper we propose two new metrics that evaluate how each question contributes to the goal. First, we measure how *effective* each question is by evaluating whether the question discards objects that are not the referent. Second, we define *referring* questions as those that univocally identify one object in the image. We report the new metrics for human dialogues and for state of the art publicly available models on GuessWhat?!. Regarding our first metric, we find that successful dialogues do not have a higher percentage of effective questions for most models. With respect to the second metric, humans make questions at the end of the dialogue that are referring, confirming their guess before guessing. Human dialogues that use this strategy have a higher task success but models do not seem to learn it.

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.

Referential visual dialogue has a clear task success metric: whether the Questioner is able 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 (Shekhar et al., 2018; Pang and Wang, 2020). But still the models are evaluated without considering how much each question contributes to the goal. We propose two new metrics for evaluating questions. First, a question is *effective* if it rules out at least one possible distractor (Krahmer and van Deemter, 2012). Second, a question is *referring* if it uniquely identifies one object in the image.

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. The figure shows for each question its answer, how many distractors (#D) are left after each answer and whether the question is effective or not effective. The last question is not effective but it is referring, it uniquely identifies the referent. Interestingly, question 2 is also referring but not with respect to the referent, so the dialogue needs to go on.

In the next section we review previous work. Then, we define the metrics formally and calculate them over the Guesswhat?! SOTA models. Finally, we argue that models, differently from humans, do not confirm their guess before guessing.

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. Historically, the performance of VLN models has been evaluated with respect to the objective of reaching the goal location (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

Proceedings of the First Workshop on Advances in Language and Vision Research, pages 19–25

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. in the front?	yes	0	True
5. first cow near us on the right?	yes	0	False

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

steps taken towards the goal for 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. This may be especially useful to allow humans to monitor agents' behaviour, whether for fault detection, assessing performance, or even building trust.

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 two new metrics 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 and referential 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 human as the robot moves. The task we address in this paper is a simplification. In our setup, the view of the robot is static, it is a picture. For our work we use the GuessWhat?! dataset (de Vries et al., 2017).

Recently, Sankar et al. (2019) showed that several end-to-end dialogue systems do not take dialogue history into account. In this paper we are particularly interested in the GuessWhat?! models that generate questions explicitly modelling the dialogue history (Zhang et al., 2018; Shukla et al., 2019; Pang and Wang, 2020).[1]

3 Dataset and Evaluation Metrics

In this section we briefly introduce the dataset and we define the evaluation metrics that we use.

3.1 GuessWhat?!

GuessWhat?! (de Vries et al., 2017) is a cooperative game where two players talk in order to identify an object in an image. The player known as the *Questioner* has to guess the referent by asking yes/no questions. The other player, the *Oracle*, knows the referent object and answers the questions. The GuessWhat?! dataset contains games of different complexity, ranging from easy images with the referent and only one distractor, to images with up to 19 distractors. The dataset is composed of more than 150k human-human dialogues containing an average of 5.3 questions in natural language created by turkers playing the game on MS COCO images (Lin et al., 2014).

3.2 Effective and referring 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. That is, it contains the list of objects annotated in the dataset and given to the Oracle model. Human Oracles did not have access to this list. At each dialogue turn t, $RS(q_t)$ is the set of objects in $RS(q_{t-1})$ such that the answer A to q_t on those objects is the same than the answer to q_t on the referent r. All answers A are computed using the Oracle proposed in (de Vries et al., 2017)

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

whose accuracy on the test set is 79%. Formally:

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

We say that a question q_t is *not effective* iff $RS(q_t) = RS(q_{t-1})$; that is, the question does *not* exclude any distractor. In our definition, an *effective* question excludes at least one distractor; hence, $RS(q_t) \subset RS(q_{t-1})$. The effectiveness of the dialogue is given by the percentage of effective questions it has. In the example given in Figure 1, the last question of the dialogue, namely, *"first cow near us on the right?"* is not effective by our definition. Strictly speaking, it does not exclude any distractor and the human could have guessed after turn 4. This last question verifies the guessed referent by constructing a referring expression for it that is relative to the speaker's position. We say that this question is *referring*.

We say that a question q_t is *referring* wrt the referent r iff $A(q_t, r) =$ "yes" and $A(q_t, o_i) =$ "no" for all other objects o_i in the image. As we do with effectiveness, we calculate A by using the Oracle model (de Vries et al., 2017) repeateadly over all objects. That is, if a question uniquely identifies the referent then its answer is "yes" only for the referent. In the example in Figure 1, the last question is not effective but it is referring, it uniquely identifies the referent. Interestingly, question 2 is also referring but not with respect to the referent, so the dialogue needs to go on. One may expect that referring questions are realized using the definite determiner "the" as in question 2, but this is not always the case as observed in question 5.

4 Experiments and results

In this section we describe the GuessWhat!? SOTA models for which the code or the test set dialogues have been released and we present our results.

4.1 Models and experiments

Models usually implement the Questioner player using two agents: the QGen which generates the questions and the Guesser which takes a finished dialogue and makes a guess for the referent.

We took the dialogues generated by different SOTA models on the test set of the split defined in (de Vries et al., 2017). The Baseline (BL) model proposed by de Vries et al. (2017) is an encoder-decoder architecture conditioned by image and dialogue features for the QGen. Its Guesser is a MLP that embeds the list of candidate objects and chooses the referent conditioned by the dialogue and the image features. The Reinforcement Learning (RL) model (Strub et al., 2017) casts the problem into a reinforcement learning task and trains the previous model with policy gradient. The Visually-Grounded State Encoder (GDSE) models, both Supervised Learning (SL) and Cooperative Learning (CL) (Shekhar et al., 2019) use a visually grounded dialogue state that takes the visual features and each new question to create a shared representation used for both QGen and Guesser. They differ in that SL is trained in a supervised fashion while CL samples new objects from pictures and makes the agents train in a cooperative learning fashion on those artificially generated games. Last, Visual Dialogue State Tracking (VDST) (Pang and Wang, 2020) extends the QGen with a representation of the probability of each object being the referent.

For each of the models, we calculate the reference sets for each question in their dialogues. We calculated the percentage of effective questions in each dialogue comparing failed and successful dialogues. For the last question in each dialogue we calculate whether it is effective and/or referring.

4.2 Results

In this section we first exemplify our metrics over dialogues generated by two models and then present the quantitative results.

Figure 2 shows an example of both metrics on a game on which VDST and CL are successful. Effectiveness is 60 for VDST and 40 for 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 CL dialogue, question 2 is redundant because all cakes in the image are dark brown. There are no referring questions in the VDST dialogue. The CL dialogue finishes with a referring and effective question that is realized using a definite article. The question even includes the connector "so" giving the feeling that the system intends to verify its guess. However, the same system uses a definite determiner in question 3 as if the question was referring but it is not (there are three dark brown cakes).

We report quantitative results for humans and

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 CL in a successful game. Non effective in italics. There are no referring questions in the VDST dialogue. The CL dialogue finishes with a referring and effective question that is realized using a definite article.

for the 5-question and 8-question setups for SOTA models. Table 1 shows average number of questions (#Q), task success (TS) and effectiveness for each of the models and the human dialogues. The table also shows the percentage of dialogues whose last turn is effective and/or referring.

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 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.

Humans and models alike ask non-effective questions mostly at the end of the dialogue. The effectiveness decreases as dialogue progresses for models and humans and reaches its lowest level in the last turn as shown in Table 1. Interestingly, models and humans seem to be using the last turn for different purposes. 26% of human dialogues end with a referring question while the model that reaches the highest value for this metric has only a 7%. We found that human task success for the dialogues that end with a referring question is 95% while it is 80% for the rest.

4.3 Analysis of Oracle accuracy

The computation of both metrics involves using an automatic Oracle. Even though this Oracle achieves high accuracy on the test set, this accuracy is actually measured on human-generated questions. In this section we evaluate this Oracle calculating its accuracy for different types of questions. We also report the different types of questions that the systems produce. The types of questions generated by systems show a distribution shift from those generated by humans. We argue that machine-generated questions are easier and the performance of the Oracle should be equal or higher for them than for the human ones.

Following Shekhar et al. (2019), we classify questions into different types and evaluate the Oracle accuracy for each type. We distinguish between eight types of questions. The first type are those that include a noun representing the category of the referent (e.g., 'is it a dog?'); we use the categories of objects defined in MS COCO (Lin et al., 2014). Then we consider questions about properties usually realized as adjectives or prepositional phrases. We make a distinction between color, shape, size, texture, location, and action questions. The classification is done by extracting keywords for each question type from the human dialogues, and then assigning each question to as many types as it fits. A question may be tagged with several attribute classes if more than one keyword is present. E.g., "Is it the white one on the left?" is classified as both color and location. The list of keywords is available in (Shekhar et al., 2019).

In Table 2 we can see that the distribution of types of questions varies from model to model and differs to the distribution in humans. Humans make more questions about the color, size, shape of the target as well as about the action that the target is performing (e.g. "is she skiing?"). Some models make more questions about the object (e.g. BL, SL and CL) and about the location (e.g. RL and VDST). The table also reports the Oracle accuracy on the human dataset per type of question. The

Model	#Q	TS	Effectiveness			Last Turn	
			All	Failure	Success	Effective	Referring
BL (de Vries et al., 2017)	8	40.7	26.4	27.5	24.7	4.2	4.46
SL (Shekhar et al., 2019)	8	49.7	29.1	31.4	26.9	7.4	5.64
RL (Strub et al., 2017)	8	56.3	32.6	36.5	29.6	3.5	2.60
CL (Shekhar et al., 2019)	8	58.4	30.2	32.3	28.6	7.6	6.08
BL (de Vries et al., 2017)	5	40.8	38.8	39.8	37.4	11.9	5.20
SL (Shekhar et al., 2019)	5	47.8	42.2	44.6	39.9	16.4	7.42
RL (Strub et al., 2017)	5	58.4	48.6	52.9	45.1	23.0	2.68
CL (Shekhar et al., 2019)	5	53.7	44.7	47.8	42.6	18.5	6.32
VDST (Pang and Wang, 2020)	5	64.4	52.9	57.4	51.0	28.7	1.82
Humans (H) (de Vries et al., 2017)	5.3	84.1	56.9	54.7	57.3	33.6	26.01

Table 1: Task success (TS), Effectiveness and Last Turn Effectiveness by model in the test set. Effectiveness is reported for all dialogues and dialogues ending in failure or success. For VDST we only report the results for 5 question dialogues as we only had access to these dialogues. For the last turn of the dialogues we report the percentage of effective and referring questions.

hardest types of question for the Oracle are color and size questions. All models ask fewer of these questions than humans. Also most models, except for VDST and RL, ask more object questions than humans; this is the type of question for which the Oracle has the highest accuracy. The models VDST and RL ask more location questions. However, we have manually observed that the location questions that cause more errors for the Oracle are questions regarding order (e.g. "the third counting from the left?"). Such questions constitute 8% of the human questions and have an accuracy of 58% but are not made by VDST and RL. The type of location questions asked by VDST are illustrated in Figure 2. We have argued that models make questions that are easier for the Oracle than those made by humans. We hypothesize that the Oracle accuracy is then higher for machine-generated questions. We will investigate this hypothesis further in future work.

Type	Acc	BL	SL	RL	CL	VDST	H
Obj	94	49.00	48.08	24.00	46.40	36.44	38.12
Color	63	2.75	13.00	0.12	12.51	0.01	15.50
Shape	67	0.00	0.01	0.00	0.02	0.00	0.30
Size	60	0.02	0.33	0.02	0.39	0.01	1.38
Tex	70	0.00	0.33	0.01	0.15	0.00	0.89
Loc	67	47.25	37.09	74.80	38.54	64.80	40.00
Act	65	1.34	7.97	0.66	7.60	0.30	7.59
Other	75	1.12	5.28	0.49	5.90	0.03	8.60

Table 2: Oracle accuracy per type of question and question distribution for the models. We report BL, SL, RL and CL question type distribution with 8 questions, and VSDT with 5 questions and the human dialogues.

5 Conclusions

We proposed two new metrics 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 both if the question is answered with 'yes' as well as if it is answered with 'no'. We observe that it 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. One of such strategies is captured by our second metric: questions that may not be effective but are referring.

Humans seem to use the last turn to confirm their guess before guessing. Human dialogues that confirm the guess using a referring questions have a higher task success than those which do not. We plan to explore whether models can learn to confirm their guess before guessing. As future work we plan to refine our referring metric. We have observed that some dialogues do not make explicit the object category in the confirmation. E.g. "the one near us on the right?" in Figure 1. By our definition, this question would not be referring because it is also true for the wooden stick.

We believe that our metrics could be heuristics that guide the training of end-to-end models.

References

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.

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. 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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755, Cham. Springer.

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. 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. 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. 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. 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. 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. 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.

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)*.

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 *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.