The impact of answers in referential visual dialog

Mauricio Mazuecos[1] and **Patrick Blackburn**[2] and **Luciana Benotti**[1]
[1]Universidad Nacional de Córdoba, CONICET, Argentina
[2]University of Roskilde, Denmark
mmazuecos@mi.unc.edu.ar
patrick.rowan.blackburn.gmail.com
luciana.benotti@unc.edu.ar

Abstract

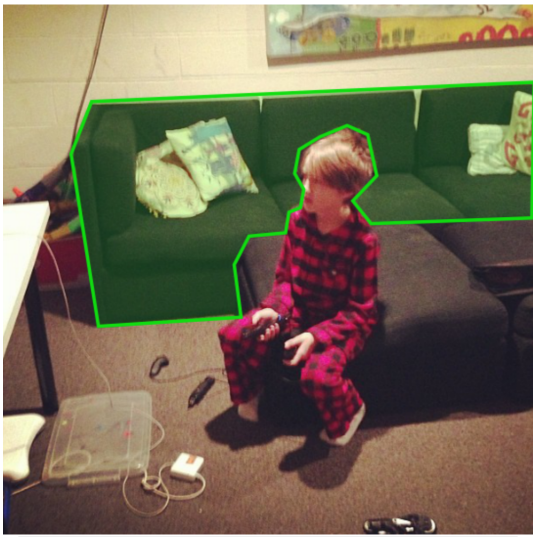
In the visual dialog task GuessWhat?! two players maintain a dialog in order to identify a secret object in an image. Computationally, this is modeled using a question generation module and a guesser module for the questioner role and an answering model, the *oracle*, to answer the generated questions. This raises a question: *what's the risk of having an imperfect oracle model?*. Here we present work in progress on the study of the impact of different oracles in human generated questions in GuessWhat?!. We show that having access to better quality answers has a direct impact on the guessing task for human dialog and argue that better answers could help train better question generation models.

1 Introduction

Collaborative reference resolution is a task that has attracted a lot of attention in recent years with the introduction of GuessWhat?! (de Vries et al., 2017). GuessWhat?! is a cooperative two-player referential visual dialogue game. One player (the *Oracle*) is assigned an object in an image and the other player (the *Questioner*) has to guess the referent by asking yes/no questions. An example of a dialog in the GuessWhat?! dataset can be seen in Figure 1.

In this task, much work has been done on the question generation policies (Strub et al., 2017; Lee et al., 2018; Shekhar et al., 2019; Pang and Wang, 2020b,a), the linguistic capabilities of these questioner models (Shukla et al., 2019a) and on improving guessing models (Pang and Wang, 2020a).

Most of the work on the questioner models was performed employing a simple oracle model to play the GuessWhat?! game. This oracle model was too simple and it struggled to answer questions that asked for anything beyond the available annotation information, thus pushing models to produce those type of questions (Mazuecos et al., 2020); a new SOTA for the oracle task (Testoni et al., 2020), based on LXMERT (Tan and Bansal, 2019) was proposed using this approach. It seems reasonable to investigate the impact on the question generation policy learned by questioner models and on task success. In this work we will focus on the latter[1]. We will show the impact of having access to better answers in the guessing task by evaluating a guesser model with questions from the human corpus that were answered by different oracle models.



Question	Answer
1. It is a person?	no
2. Is it something you sit on?	yes
3. Does it have pillows on it?	yes

Figure 1: Example of a game played by humans in the GuessWhat?! dataset.

In the next section we review previous work. Then we explain the GuessWhat?! task, the models, and the experiments. Finally we argue that having access to better answers could be the difference between success and failure in the guessing task.

[1] The code for reproducing this paper is available at github.com/mmazuecos/ReInAct2021-Impact-of-answers

Proceedings of the Conference on Reasoning and Interaction (ReInAct), pages 8–13
October 4-6, 2021. ©2021 Association for Computational Linguistics

2 Previous Work

GuessWhat?! (de Vries et al., 2017) is a cooperative guessing game in which two players hold a dialog intended to identify a secret object in a picture. We call this object the *target object*. The two players have different roles: the *Questioner* has to pose questions and guess the object at the end of the dialog, and the *Oracle* has to answer these questions. The corpus comprises more than 155K dialogs with more than 821K question-answer pairs made across 67K images extracted from the MSCOCO dataset (Lin et al., 2014).

GuessWhat?! is a simplification of the collaborative process of referring studied by Clark and Wilkes-Gibbs (1986). The process of multimodal reference resolution had received attention from the vision and computational linguistics communities (Pineda and Garza, 1997; Schlangen et al., 2009). The task requires both reference resolution capabilities and the ability to ground the language expressions to objects in the real world (Roy, 2005).

In this task, much work has been done on training question generation policies to perform the questioner role (Strub et al., 2017; Zhang et al., 2018; Abbasnejad et al., 2019; Shukla et al., 2019b; Shekhar et al., 2019; Pang and Wang, 2020b) with different levels of task success at guessing the target objects. Most of the approaches receive some sort of reward that weights to some extent the task success at the game. Being a two player game, the task success will be conditioned by both the questioner performance and the oracle performance. Most works use the same oracle model proposed with the GuessWhat?! dataset (de Vries et al., 2017)[2].

We previously showed that the baseline oracle proposed by de Vries et al. (2017) does not have the same performance for human and model generated questions, and that performance was linked to the type of question (Mazuecos et al., 2020). Most RL-based models would not ask for information other than the type of object and its location, exactly the two manually annotated features that the oracle receives. As a result the grammatical and lexical diversity of the generated questions is poor.

Following this line, Testoni et al. (2020) proposed a more complex oracle model based on the multimodal transformer LXMERT (Tan and Bansal, 2019). This model achieved SOTA for the Oracle task and proved to perform better across most question types except for object questions (due to not having access to the gold standard category label for the target).

The impact of answers has previously been noted for the VisDial task (Das et al., 2017). Guo et al. (2019) show that a visual dialog model with integration of better answers achieves better performance in the Visual Dialog Challenge 2018. We show the impact the answer has for the GuessWhat?! task.

3 GuessWhat?! task and Models

In the GuessWhat?! game there are two roles: the *Questioner*, that makes questions and guesses the target object at the end of the dialog, and the *Oracle* that answers those questions. At the beginning of a game, the oracle is assigned an target object in the image and the questioner has to pose yes/no questions in order to identify the target. An usual computational modeling for each player divides the Questioner role into two components: the *Question Generator* and the *Guesser*. In our experiments we make use of a guesser model and oracle models.

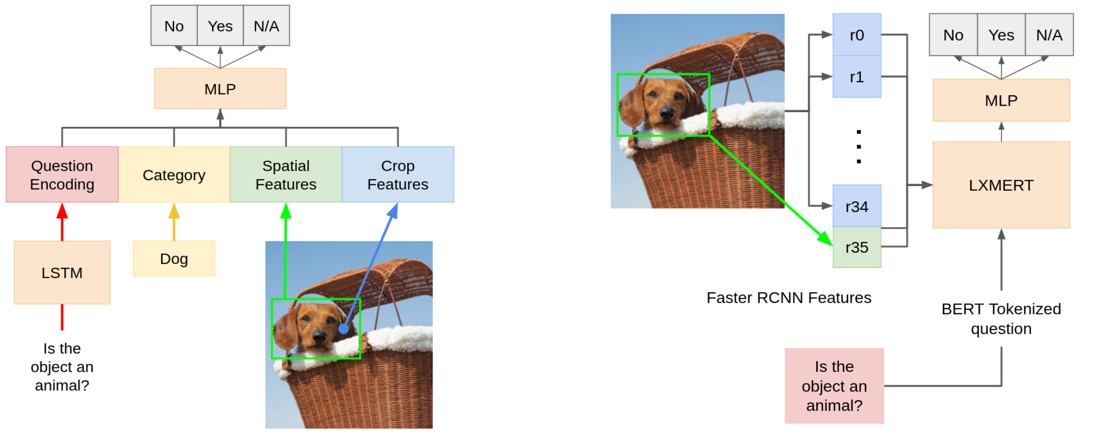
For the Oracle models we used the baseline model (de Vries et al., 2017) as well as the LXMERT based model (Testoni et al., 2020).

The baseline models (Figure 2a) use a set of features passed through a multilayer perceptron (MLP). We consider a subset of the features according to previous work (de Vries et al., 2017): The question (**Q**), the spatial information of the target (**Sp**), the target object's category extracted from the MSCOCO (**Ca**) and the visual features of crop of the target (**Cr**), extracted with a ResNet152 (He et al., 2016).

The LXMERT based model (Figure 2b) receives the question, the visual features of 36 regions of the image (the same as in (Anderson et al., 2018)) and the crop of the target inserted in the 36th position of the regions. Notice that this model has no access to the category's label for the target object.

We use the Guesser model (Figure 3) proposed by Shekhar et al. (2019). This guesser model adds an encoding of both vision (the image) and language modalities (the full history). A single MLP processes each object's spatial information and category and outputs a score for each object. These concatenated scores combined with the vision and language encoding output the probability of each object being the target used to make the final guess.

[2] We could not confirm nor deny that Abbasnejad et al. (2019) employed that Oracle mode. They stated that they kept the classical GuessWhat?! setup.



(a) Baseline oracle model (de Vries et al., 2017)

(b) LXMERT based oracle model (Testoni et al., 2020)

Figure 2: Oracle models.

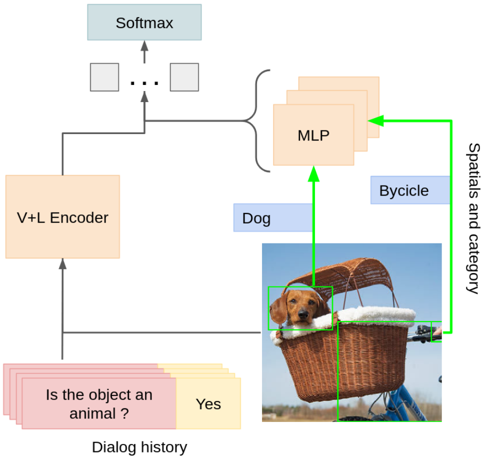


Figure 3: Guesser model (Shekhar et al., 2019).

4 Experiments

In this section we show the results of our experiments. Our experiments were performed using a single 1080ti GPU. We retrained all of our models following the procedure stated in their respective paper using the publicly available source code.

For our experiments we take the human dialogs from the test set of GuessWhat?!. We keep the human posed questions but change the answers with the ones given by the different oracle models we employed. We stick to the successful games in the test set, as failed and incomplete games tend to contain malformed questions, misunderstandings or are not finished and, thus, guessing is not possible with the information available.

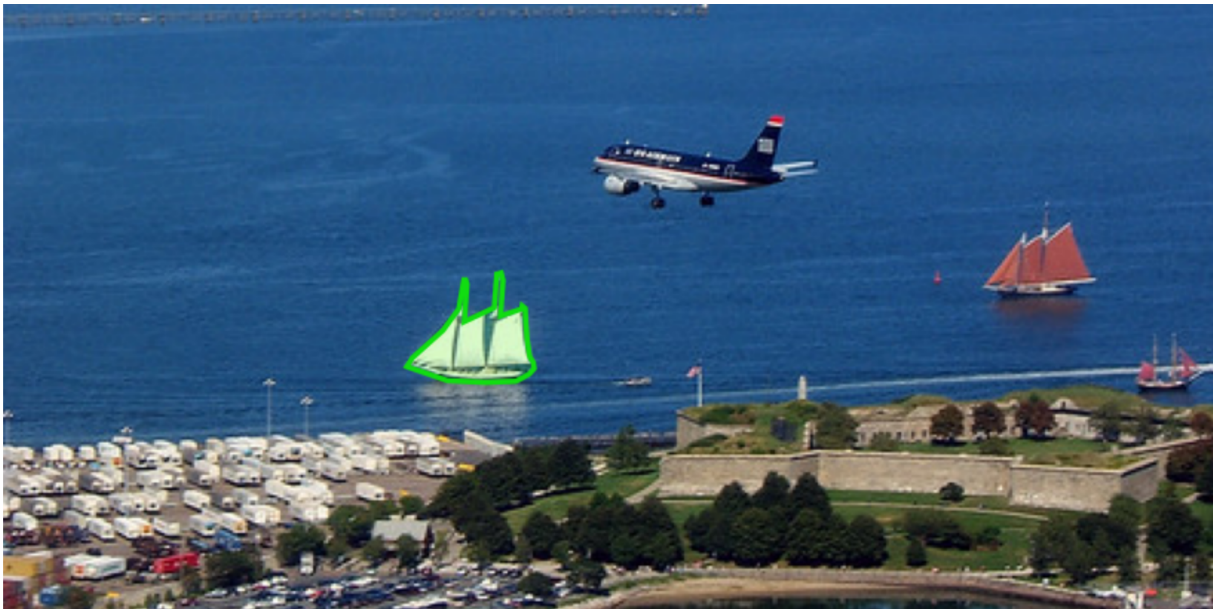
We measure the task success of the guesser at guessing the target object in these resulting dialogs. In Table 1 we see the tasks success for the guesser in each setup. We see that the **Q+Sp+Ca** answers have a strong performance, just points below the more complex model based on LXMERT. To test the impact of the target category (**Ca**) we discard it (**Q+Sp**) or change it for the crop of the target (**Q+Sp+Cr**). Our hypothesis was that the category was playing a heavy role on the performance of the oracle models. The first and second row of Table 1 show that leaving the category out reduces up to 13 points of task success on the guesser. Replacing the category with the crop improves the performance adding information about the target but still is almost 7 points below the **Q+Sp+Ca** baseline.

Answers from	Guesser Task Success
Q+Sp	46.9
Q+Sp+Cr	52.7
Q+Sp+Ca	59.4
LXMERT	59.7
Human	62.2

Table 1: Task success of the Guesser model in human generated questions from the GuessWhat?! test set with answers from different Oracles.

The LXMERT model, despite not having the gold stardard labels for the categories achieves a similar and even higher performance when paired with the guesser model. This shows that the LXMERT oracle can be used in settings with no annotation for the objects categories.

In Figure 4 we see an example of a dialog and the answers given by different oracles. In this example, models other than LXMERT miss at least one answers and fail at guessing. This shows that missing a single answer can end up in failure at guessing



Question	Human	Q+Sp	Q+Sp+Cr	Q+Sp+Ca	LXMERT
1. is it a ship?	yes	yes	no	yes	yes
2. is it white?	yes	no	no	yes	yes
3. is it under the plane slightly left?	yes	yes	no	no	yes
Status	Success	Failure	Failure	Failure	Success

Figure 4: Example different answers for the same human dialog given by the different Oracles. **Human** corresponds to the human answers given in the corpus. In this example, the guesser model correctly guesses the target for the human and LXMERT oracles, while the game resulted in failure for the rest of the Oracle models.

the target object, as seen for the **Q+Sp+Ca** and **Q+Sp** oracles.

Oracle	IAns	Avg IAns/failure
Q+Sp	31.73%	2.08
Q+Sp+Cr	25.60%	1.78
Q+Sp+Ca	22.03%	1.66
LXMERT	16.05%	1.27

Table 2: Percentage of incorrect answers (IAns) with respect ot the human answers and Average number of incorrect answers per failed game in the test set of the GuessWhat?! for the different oracle models evaluated.

Following this we compute the percentage of incorrect answers (**IAns**) and the average amount of them per failed game (**Avg IAns/failure**) for each model. In Table 2 we see the result of this analysis. There is a negative correlation (-0.93 Pearson's R) between task success and IAns as well as for Avg IAns/failure (-0.87 Pearson's R).This suggest that future evaluated models should take into account the oracle's performance when it comes to gameplay between agents. The oracle could be hindering the real potential of questioner and guesser models as they would learn to either exploit the oracle's annotation or risk failure.

5 Conclusions

In this paper we described work in progress on studying the impact of having access to better answers in GuessWhat?!. Dialogs with better answer quality had a higher task success when sent to an automatic guesser. The task success when using the widely used **Q+Sp+Ca** and the LXMERT oracles are comparable although LXMERT does not require the object manual annotations. Task success drops when the gold standard category label for the target is not a feature. This suggest that the MLP oracles rely strongly on the manual annotations.

The next step will be to investigate the impact of the answer quality on the quality of the generated questions. In order to do so we will perform a similar analysis for the different questioner and guesser models proposed in the literature. We hypothesize that this could lead the question generation policies to have richer linguistic capabilities and to learn better strategies for identifying the target object.

References

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, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

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

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

Dalu Guo, Chang Xu, and Dacheng Tao. 2019. Image-question-answer synergistic network for visual dialog. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10426–10435.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

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, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context. Cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list.

Mauricio Mazuecos, Alberto Testoni, Raffaella Bernardi, and Luciana Benotti. 2020. On the role of effective and referring questions in GuessWhat?! In *Proceedings of the First Workshop on Advances in Language and Vision Research*, pages 19–25, Online. Association for Computational Linguistics.

Wei Pang and Xiaojie Wang. 2020a. Guessing state tracking for visual dialogue. In *ECCV*.

Wei Pang and Xiaojie Wang. 2020b. Visual dialogue state tracking for question generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11831–11838. AAAI Press.

Luis. A. Pineda and E. Gabriela Garza. 1997. A model for multimodal reference resolution. In *Referring Phenomena in a Multimedia Context and their Computational Treatment*.

Deb Roy. 2005. Grounding words in perception and action: computational insights. *Trends in Cognitive Sciences*, 9(8):389–396.

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental reference resolution: The task, metrics for evaluation, and a Bayesian filtering model that is sensitive to disfluencies. In *Proceedings of the SIGDIAL 2009 Conference*, pages 30–37, London, UK. 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*, pages 2578–2587. ACL.

Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, Matthew Turk, and William Yang Wang. 2019a. 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.

Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, Matthew Turk, and William Yang Wang. 2019b. 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 Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *Proceedings of international joint conference on artificial intelligenc (IJCAI)*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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)*, Hong

Kong, China. Association for Computational Linguistics.

Alberto Testoni, Claudio Greco, Tobias Bianchi, Mauricio Mazuecos, Agata Marcante, Luciana Benotti, and Raffaella Bernardi. 2020. They are not all alike: Answering different spatial questions requires different grounding strategies. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 29–38, Online. Association for Computational Linguistics.

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4475. IEEE Computer Society.

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.