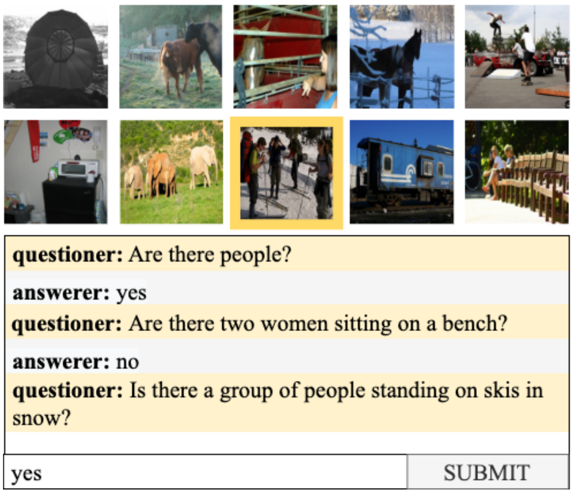
Generating informative, open-domain clarification questions without question examples

Julia White[1], Gabriel Poesia[2], Robert Hawkins[4], Dorsa Sadigh[1,2], Noah Goodman[2,3]
Departments of [1]Electrical Engineering, [2]Computer Science, and [3]Psychology at Stanford University
Department of [4]Psychology at Princeton University
jiwhite@stanford.edu

Abstract

Human-machine interaction relies on accurate transfer of knowledge from users. Rather than placing the full burden of providing high-quality input on the user, or the full burden of error-correction on the model, the ability to ask questions supports a more interactive form of cooperative repair. Drawing on recent work in cognitive science, we propose a visually-grounded agent that is able to ask context-sensitive clarification questions to resolve misunderstandings in dialogue. Our question generation framework uses an information gain objective to derive informative polar (yes-no) questions from an off-the-shelf image captioner without the need for question-answer data. We demonstrate the communicative efficacy of this strategy on a question-driven communication game task with both synthetic and human answerers.



questioner: Are there people?

answerer: yes

questioner: Are there two women sitting on a bench?

answerer: no

questioner: Is there a group of people standing on skis in snow?

yes

Figure 1: Our model takes the role of questioner in a question-driven communication game where it must guess which image is being described by the answerer. The interaction ends with the model returning a guess for which image the answerer is referring to.

1 Introduction

An overarching goal of natural language processing is to enable machines to communicate seamlessly with humans. However, natural language can be ambiguous or unclear. In cases of uncertainty, humans engage in an interactive process known as repair (Clark, 1996; Arkel et al., 2020): asking questions and seeking clarification until their uncertainty is resolved. In this work, we focus on the task of generating clarification questions aimed to reduce the uncertainty that can arise in human-machine interactions.

Generating useful natural language questions has posed a significant computational challenge. One popular approach is to use end-to-end machine learning to map visual and linguistic inputs directly to questions (Yao et al., 2018; Das et al., 2017). However, these approaches are heavily data-driven, requiring large annotated training sets for different goals and contexts. Another approach has drawn from work on active learning and Optimal Experiment Design (OED) in cognitive science to search for questions that are likely to maximize expected information gain from an imagined answerer (Rothe et al., 2017; Wang and Lake, 2019; Misra et al., 2018; Lee et al., 2018; Rao and Daumé III, 2019). Much of this recent work in question generation has relied on question-answer data like the GuessWhat!? dataset (de Vries et al., 2017), which contains question-answer pairs corresponding to the objects in particular images. We present a captioner-based approach that uses a similar expected information gain objective, but can instead be trained from more widely-available image-caption data.

In this paper, we demonstrate a framework to build a visually-grounded question asking model from a pretrained image captioner. Our key contribution is a goal-oriented approach to question generation that isn't limited to a specific domain or by a reliance on question-answer data. Our model gen-

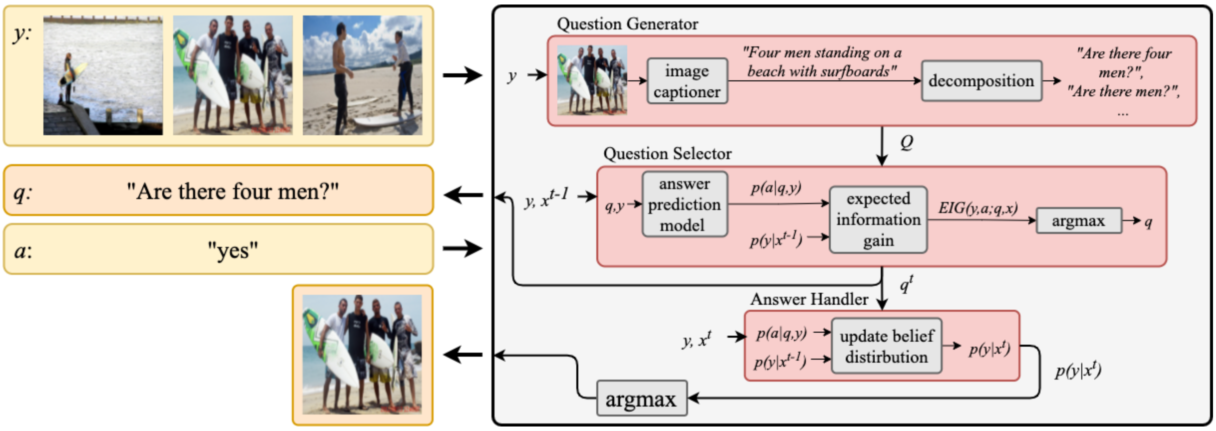


Figure 2: Illustration of our question generation framework. A set of candidate questions are produced by our question generator, and then ranked according to their expected utility in the question selector module. After posing the highest-ranked question and receiving an answer, the belief distribution over images is updated in the answer handler module and these updated beliefs are then either used to guess the target image or are fed back to the question selection module for the process to be repeated.

erates sets of polar candidate questions by applying rule-based linguistic transformations to the outputs from an image captioner. We then use a response model to predict the likelihood of different answers. Given these predictions, we estimate the expected information gain of each question and select the question with the highest utility. We demonstrate our method's ability to pose questions that resolve uncertainty in a question-driven communication game with synthetic and human answerers.

2 20 Questions Task

We study interaction within a Lewis-style signalling game (Lewis, 1969) between two agents, a questioner and an answerer (see Figure 1). Both agents are shown a set of images, and a target image is privately indicated to the answerer. The questioner's goal is then to produce questions that, when answered, would allow them to identify the target image. After a maximum of 20 questions the questioner attempts to guess the target image. Crucially, the most informative question changes depending on context and prior information. To approximate the setting of natural "clarification questions" we also consider games that begin with a, potentially ambiguous, description of the target. Crucially, the most informative question changes depending on context and prior information.

3 Model

Our model (Figure 2) maintains a belief distribution, $p(y|x^t)$, about the target in a set of images, $y \in Y$, given the history of the interaction, $x^t = (a_1, q_1, ..., a_t, q_t)$, which includes all questions, q, and answers, a, exchanged up to the current step, t. At each interaction step, our model selects a question based on expected information gain and updates its beliefs according to the answer.

Question Generator. We derive candidate questions from a pre-trained image captioning model which allows us to generate a set of contextually-relevant questions without observing any question data [1]. First, we use the captioner to produce a list of captions corresponding to each image in a given game. We then decompose each of these captions into multiple polar questions according to their constituency parse which we obtain using Berkeley Neural Parser (Kitaev et al., 2019). For each Noun Phrase (NP) subtree in the constituency tree, we generate a question of the form 'Is there <NP>?' or 'Are there <NP>?', depending on the appropriate plurality. We ignore subtrees containing personal pronouns (e.g. 'she'), since they usually refer to nouns introduced outside of the subtree. Finally, we replace definite by indefinite articles to better match the question template. Using these rules, we generate an average of 10 candidate questions from each caption which capture a rich variety of compositional information pertaining to features including the actions, locations, and objects observed in an image.

Question Selector. To determine the most informative question, q^t, at turn t, we estimate expected

[1] https://github.com/krasserm/fairseq-image-captioning

information gain, $EIG(y, a; q, x)$, for the set of candidate questions, Q. EIG is defined as the decrease in entropy of the distribution over images after observing answer, $a \in A(q)$, to question, q. Because the initial entropy is independent of the current question, maximizing the EIG is equivalent to minimizing the conditional entropy of the distribution over images after observing an answer: $q^t = \text{argmin}_{q_j \in Q} H(y|a; q_j, x^{t-1})$.

$$H(y|a; q_j, x^{t-1}) = -\sum_{y_i \in Y} \sum_{a_k \in A(q_j)} p(a_k|q_j, y_i) \\ p(y_i|x^{t-1}) \log(p(y_i|x^{t-1}, q_j, a_k)) \quad (1)$$

The posterior belief after the answer, $p(y|x^{t-1}, q_t, a) = p(y|x^t)$, is described below. The answer probability, $p(a|q, y)$, represents beliefs about how the answerer is likely to respond to a question given a particular image. We model this by pre-training a CNN classifier from positive pairs ('yes' answers) of an image and a question derived from a caption for this image, and negative pairs ('no' answers) derived by permuting these. This classifier is trained via cross-entropy loss on image and caption embeddings obtained from a CNN and RNN encoder respectively.

Answer Handler. The conditional probability $p(y_i|x^t)$ decomposes via Bayes rule:

$$p(y|x^t) = p(y|x^{t-1}, q^t, a^t) \propto p(a^t, q^t, y|x^{t-1}) \\ = p(a^t|x^{t-1}, q^t, y) p(q^t|x^{t-1}, y) p(y|x^{t-1}) \quad (2)$$

We make the simplifying assumption that the answer, a^t, depends only on the question q^t and the underlying target label y, and is independent of past interactions. This allows us to simplify $p(a^t|x^{t-1}, q^t, y_i)$ to the pre-trained response model, $p(a^t|q^t, y_i)$, described above. For our model, the selection of question q^t is deterministic given previous interactions x^{t-1}.

The initial belief distribution, $p(y|x^0)$, is uniform unless a target description, u, is provided. Then, the model begins with the belief distribution $p(y|x^0) \propto p(u|y)$, where $p(u|y)$ is the utterance likelihood according to the captioner.

4 Data

We ran experiments on two communication game datasets: Shapeworld and MS COCO which represent artificial and naturalistic settings, respectively.

Shapeworld (Kuhnle and Copestake, 2017) is an artificial dataset of images containing a single object of a random shape and color with a vocabulary of 15 words which describe the possible colors and shapes. Our model evaluation dataset contains a total of 1,000 games which contain 10 images each. A set of 1,000,000 images and their captions (which can include the shape, color, or both) was used to train a Shapeworld-specific image captioner and answerer model.

MS COCO (Lin et al., 2015) is a dataset of images of complex everyday scenes with corresponding human captions and a vocabulary of 9,808 words. The image captioner and answerer model were trained on the Karpathy splits which allocate 155,000 samples for training and 5,000 images for validation and testing each. Our model was evaluated on 1,000 games, each containing 10 images sampled from the unseen test split.

5 Experimental Results

We evaluated our question asking framework by observing communication accuracy on question-driven communication games played with both synthetic (Figure 3) and human (Figure 4) answerers. We judge the effectiveness of our question asking model in comparison to several baselines: a **full caption** model which generates candidate questions from full image captions without decomposition, such that this model is comparable to a linear search or checking one image at a time; a **random question** model which selects questions randomly instead of according to expected information gain; and, a **binary search** algorithm which serves as a stand-in for optimal question asking by halving the set of potential target images with each step of the interaction (but note that this is an upper bound, and there may be no natural language questions that could achieve such a split).

Synthetic Experiments. For the Shapeworld dataset we evaluate our methods with a player simulator constructed to provide ground-truth answers to generated questions (Figure 3, left). Our model outperforms the full caption model, which produces questions that are too specific to efficiently narrow the space of potential target images. Our model also achieves higher communication accuracy than randomly selected questions. This shows the utility of both having a question set of varying specificity and using expected information gain to adapt question selection to the model's current knowledge. Crucially, our model only slightly under-performs the upper bound, binary search al-

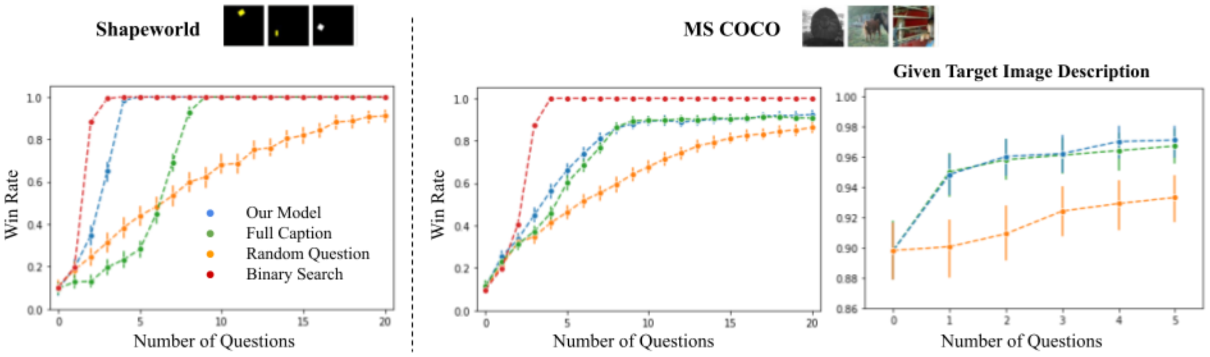


Figure 3: Win rate curves for games played with synthetic answerers. Error bars correspond to a 95% confidence interval across games.

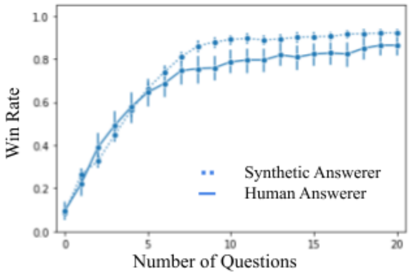


Figure 4: Win rate curves for games played with human answerers on the MS COCO dataset.

gorithm.

Ground-truth answers are not readily available for MS COCO, so we rely on answer heuristics – the answerer responds "yes" if a question is generated from the target image, and "no" otherwise (Figure 3, middle). On MS COCO we again see that our model outperforms questions derived from full captions, but to a significantly lesser extent, possibly because it is harder to find natural language questions that divide an image set in the natural image domain. The larger gap between our model and binary search, compared to Shapeworld, also indicates room for improved questions. And, information gain based selection is still shown to be important by the significantly weaker performance of randomly selected questions.

When models were given an initial description of the target image before asking any questions (Figure 3, right), we see that questions are still useful – improving accuracy by 6% from the caption alone.

Human Experiments. We evaluated our question generation model on the MS COCO dataset with human answerers (Figure 4). We recruited 20 participants from Amazon Mechanical Turk. Each participant played 10 rounds of our 20 questions game. Games were sampled from the 1,000 MS COCO games used for synthetic evaluation. We found that the model performed almost as well when paired with a human answerer as compared to a synthetic answerer. These results suggest that our model produces human-interpretable questions whose answers are effective for target image selection.

6 Conclusions

We introduce a question generation framework capable of producing open-domain clarification questions. Instead of relying on specialized question-answer training data or pre-specified question spaces, our model uses a pretrained image captioner in conjunction with expected information gain to produce informative questions for unseen images. We demonstrate the effectiveness of this method in a question-drive communication game with synthetic and human answerers. We found it important to generate questions varying in specificity by decomposing captioner utterances into component noun phrases. And, having generated this large set of potential questions, selecting based on estimated information gain yielded useful questions.

Without seeing question examples, our framework demonstrates a capacity for generating effective clarification questions. Future research should aim to generate wider question sets, with more types of answers, and improve computational efficiency. Integrating this clarification capacity more fully into collaborative, goal-directed dialog agents will allow them to share the referential burden with a user through interactive repair.

Acknowledgements

This research was supported in part by the Office of Naval Research grant ONR MURI N00014-16-1-2007 and the Stanford HAI Hoffman–Yee project 'Towards grounded, adaptive communication agents'.

References

Jacqueline Van Arkel, Marieke Woensdregt, Mark Dingemanse, and Mark Blokpoel. 2020. A simple repair mechanism can alleviate computational demands of pragmatic reasoning: simulations and complexity analysis. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 177–194.

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

Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *International Conference on Computer Vision (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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505.

Alexander Kuhnle and Ann Copestake. 2017. Shapeworld - a new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*.

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 *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, pages 2579–2589.

David K. Lewis. 1969. *Convention: A Philosophical Study*. Harvard University Press.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, C. Lawrence Zitnick Deva Ramanan, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755.

Ishan Misra, Ross Girshick, Rob Fergus, Martial Hebert, Abhinav Gupta, and Laurens Van Der Maaten. 2018. Learning by asking questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20.

Sudha Rao and Hal Daumé III. 2019. Answer-based adversarial training for generating clarification questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 143–155.

Anselm Rothe, Brenden M. Lake, and Todd Gureckis. 2017. Question asking as program generation. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, pages 1046–1055.

Ziyun Wang and Brenden M. Lake. 2019. Modeling question asking using neural program generation. *arXiv preprint arXiv:1704.04517.*

Kaichun Yao, Libo Zhang, Tiejian Luo, Lili Tao, and Yanjun Wu. 2018. Teaching machines to ask questions. In *International Joint Conferences on Artificial Intelligence*, pages 4546–4552.