Human-Model Divergence in the Handling of Vagueness

Elias Stengel-Eskin **Jimena Guallar-Blasco** **Benjamin Van Durme**
Johns Hopkins University
{elias, jgualla1, vandurme}@jhu.edu

1 Introduction

Part of the power of language as a medium for communication is rooted in having a reliable mapping between language and the world: we typically expect language to be used in a consistent fashion, i.e. the word "dog" refers to a relatively invariant group of animals, and not to a different set of items each time we use it. This view of language dovetails with the supervised learning paradigm, where we assume that an approximation of such a mapping can be learned from labeled examples—often collected via manual annotation by crowdworkers. In natural language processing (NLP), this learning typically takes place by treating tasks as classification problems which optimize for log-likelihood. While this paradigm has been extensively and successfully applied in NLP, it is not without both practical and theoretical shortcomings. Guided by notions from the philosophy of language, we propose that borderline cases of vague terms, where the mapping between inputs and outputs is unclear, represent an edge case for the assumptions made by the supervised paradigm, and result in systematic divergences between human and model behavior.

To demonstrate this, we begin by identifying a set of canonically vague terms in the binary question subset of the Visual Question Answering (VQA) and GQA datasets (Antol et al., 2015; Goyal et al., 2017; Hudson and Manning, 2019) and isolating a subset of images, questions, and answers from these datasets centered around these terms. Using this subset, we show that although the accuracy of LXMERT (Tan and Bansal, 2019) on non-borderline cases is very high, its performance drops—sometimes dramatically—on borderline cases. We then compare the behavior of the model against that of human annotators, finding that while humans display behavior which aligns with theories of meaning for vague terms, model behavior is less predictable.

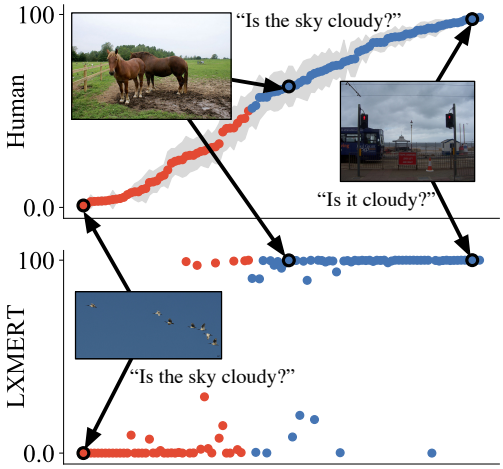

Figure 1: Given a binary question involving a vague term (in this case, *cloudy*) humans hedge between "yes" and "no," following a sigmoid curve with borderline examples falling in the middle. Standard error (grey band) shows that annotator agree even in borderline regions. In contrast, model predictions remain at extreme ends.

We extend our analysis of visually-grounded terms to a text-only case, re-framing the categorization of statements into true statements and false ones as a task involving vagueness. Controlling for world knowledge, we find that while probes over contextualized encoders can classify statements significantly better than random, their output distributions are strikingly similar to those observed in the visually-grounded case. When contrasted with scalar annotations collected from crowdworkers, these results support the notion that analytic truth itself admits of borderline cases and poses problems for supervised systems.

2 Motivation and Background

Vague terms, broadly speaking, are ones that admit of borderline cases; for example: *cloudy* is vague

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 390-393.
Held on-line February 14-19, 2021

because, while there are clearly cloudy and not cloudy days, there are also cases where the best response to the question "is it cloudy?" might be "somewhat" rather than a definitive "yes" or "no." Given this definition, we can see that a large portion of the predicates we use in every-day speech are vague. This even encompasses predicates such as *is true* and *is false*, as we might have statements that are true or false to varying degrees. Vague predicates in particular have been a focus of the philosophy of language, as they represent an interesting edge case for theories of meaning.

While unequivocal instances of vague terms fit well into the current paradigm of supervised learning with categorical labels, borderline instances present a problem. A key assumption made by supervised learning is that the ideal mapping between the input (in this case, questions and images) and the the label set (answers) is largely fixed. For example, given the question "Is this a dog?" we assume that the set of things in the world which we call "dog", also known as the *extension* of "dog", remains constant. In that case, the annotator's response to the question corresponds to whether what the image depicts could be plausibly considered as part of the extension of "dog." While we might easily be able to determine the set membership of poodles and terriers, we may have a harder time with Jack London's White Fang: half wolf, half dog. Thus it is clear that the borderline cases of vague terms demand a more nuanced account than merely a forced choice between two extremes.

3 Visually Grounded Vagueness

We focus here on binary questions about images, taking examples from VQA and GQA; this ensures that the vague term is the question's focus, excluding open-ended queries like "What is the old man doing?" which only implicitly involve vagueness.

We begin by isolating a number of vague descriptors (*sunny*, *cloudy*, *adult*, *young*, *new*, *old*) in the VQA and GQA datasets. We then use high-recall regular expressions to match questions from these descriptors in the development sets of both datasets, manually filtering the results to obtain high-precision examples.

While the VQA development data contains 10 annotations per example, GQA does not, and thus, in order to verify the quality of the VQA annotations and to collect annotations for GQA, we solicited 10-way redundant annotations from Mechanical Turk, presenting annotators with a question and its corresponding image from the vision-and-language dataset (e.g. "Is it sunny?").[1] Rather than providing categorical labels (e.g. "yes", "no") workers were asked to use a slider bar ranging from "no" to "yes", whose values range from 0 to 100. The results are rescaled per annotator.

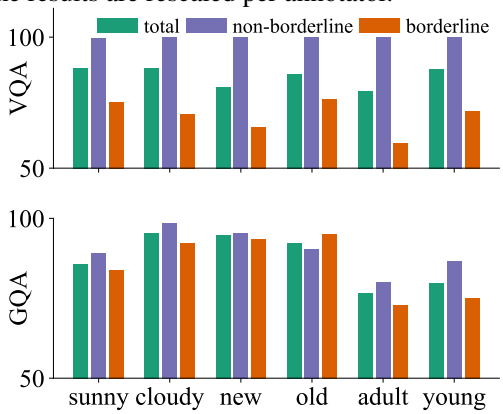

Figure 2: Accuracy of LXMERT on VQA and GQA Yes/No questions per predicate is highest for non-borderline examples, but drops in "borderline" regions.

We begin by demonstrating that vagueness is not merely a theoretical problem: Fig. 2 shows that while the total accuracy of LXMERT (Tan and Bansal, 2019) is fairly high, it drops on all descriptors (except for "old" for GQA) in the borderline regions. We argue that, given that these borderline examples account for roughly half of the data examined, the relatively high aggregate performance obtained by models on binary questions in VQA and GQA may be due to an absence of vague terms rather than to the strength of the model.

Having demonstrated that model performance is lower on borderline cases, we seek to further explore the divergence in model and human behavior. Fig. 1 plots the mean human scores in the top plot, with examples ordered by their mean human rating. The bottom plot shows LXMERT output scores for the same examples. The human scores display a sigmoid shape, while the model scores are saturated at either 0 or 1.

We posit a 2-parameter sigmoid response function given by $\left(1+\exp\left(-k*(x-x_0)\right)\right)^{-1}$ where k and x_0 are scale and shift parameters, respectively. This parameterization corresponds to the intuition that non-borderline examples will be annotated close to the spectrum ends (0 and 100) while the borderline examples will form a curve in the center of the spectrum. In some cases, this

[1] Since we were merely verifying the data quality for VQA, we only ran two descriptors: "sunny" and "cloudy".

curve is more stretched, nearing a line, while in others it is more pronounced.

We fit three separate logistic regressions: one to the mean of the annotator responses, one to the model response obtained from LXMERT, and a baseline fit against data drawn from a uniform distribution. The quality of the fit, measured by RMSE on 10% held-out data, repeated across 10 folds of cross-validation, is given in Fig. 3. For both datasets, sigmoid functions fit to model predictions have an RMSE comparable to those fit to uniformly random data, while the functions fit to human data have errors an order of magnitude lower. This indicates that the remaining GQA and VQA predicates follow a similar pattern to Fig. 1.

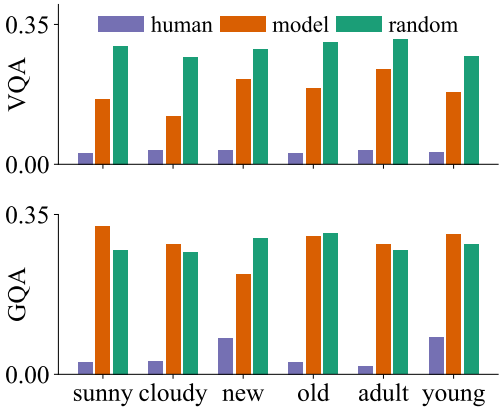

Figure 3: Mean RMSE from sigmoid fit to VQA and GQA data using 10-fold cross-validation. Human predictions result in a far better sigmoid fit, while model predictions have similar fit to data $\sim \mathcal{U}(0, 1)$.

These results suggest that the assumption that images can be identified as being in the extension of a descriptor or not (e.g. in the set of scenes described as "cloudy"), holds only at the ends of the example range, and is not warranted in the borderline region. In contrast, the training data which LXMERT sees makes the assumption that the descriptor either applies (examples with a "yes" label) or does not apply (examples labelled "no") in all regions; we see that this assumption may be too strong for capturing the nuances of vague terms.

4 Text-only Vagueness

§ 3 explored predicates grounded in another representation of the world, namely images. However, much of NLP deals with text in isolation, without grounding to some external modality. In an ungrounded setting, it is unproductive to evaluate models on external knowledge that they would not have access to—thus, we cannot evaluate a text-only model's performance on vague predicates the same way as a grounded model's performance. In other words, we need to develop a paradigm which does not rely on knowledge about a state of the world, but rather on linguistic knowledge. This is precisely the analytic-synthetic distinction, with analytic truths being truths *by virtue of meaning alone* (e.g. "a bachelor is an unmarried man") and synthetic truths being those which require verification against a state of affairs (e.g. "Garfield is a bachelor"). Furthermore, we can see the truth of a statement as being itself vague: there are statements which are only partially true or false.

Sentence	T/F	Mark
journalism is newspapers and magazines collectively	T	◆
T-shirt is an archaic term for clothing	F	▲
T-shirt is a close-fitting pullover shirt	T	●
a teammate is someone who is under suspicion	F	■

Table 1: Example sentences, with their label in the created dataset and corresponding color in Fig. 4.

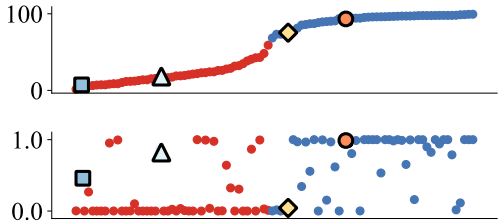

Figure 4: *Top:* mean truth score given by humans on 96 statements. False statements colored red, true blue; statements from Table 1 overlaid. *Bottom:* $P(\text{true})$ assigned by the best probing classifier (XLNet + `[CLS]`).

Following Ettinger et al. (2018), our prompts are created artificially to mitigate annotator bias. We create analytically true and false prompts by pairing a "trigger" word either with its definition or with that of a distractor term in a similar domain.

We probe 3 different encoder types: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019) with three different architectures: classifying based on a mean-pool of the representations, based on the CLS token, and based on a bilinear similarity function between the term and the definition. 96 sentences were sampled from the development set and annotated with 10-way redundancy by vetted crowdworkers on Mechanical Turk, using a similar interface as in § 3.

Annotators are able to perform the task with high reliability, achieving an accuracy of 88.54 with majority voting, with all models falling well this, but above the random baseline of 50%. Fig. 4 shows that certain sentences are easily classified as either true or false, while a smaller number of sentences are considered borderline. This demonstrates a similar trend to those seen in § 3, showing that the classification patterns of humans differ drastically from those of the best model, as illustrated by the overlaid examples. We also see the same overconfidence in the output distribution of the model, with predictions saturating at either end of the simplex. Fig. 5 further reinforces this; here, we perform the same analysis as in § 3. Across all models and all encoder types, we see that the RMSE of a sigmoid fit to the model predictions is close to or higher than the RMSE of a sigmoid fit to uniformly random data ($\text{RMSE}_{\text{random}} = 0.326$), as evidenced by the overlaid red horizontal line, while the sigmoid fit to human performance has a far lower RSME ($\text{RMSE}_{\text{human}} = 0.051$). This quantitatively reinforces the qualitative difference seen in Fig. 4.

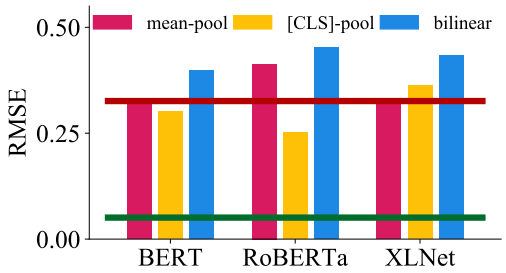

Figure 5: 10-fold cross-validated RMSE against model of 2-parameter sigmoid against model predictions from each encoder and model pairing. RMSE to human performance (green line, bottom) and against random data (red line, top) are overlaid. RMSE to model predictions is close to or worse than to random data.

5 Conclusion

We have identified clashes between the assumptions made under the current NLP paradigm and the realities of language use by focusing on the phenomenon of vagueness. By isolating a subset of examples from VQA and GQA involving vagueness, we were able to pinpoint some key divergences between model and human behavior which result in lower model performance. We then created an artificial text-only dataset, controlling for world knowledge, which we used to contrast multiple models building on multiple contextualized encoders, finding similar human-model contrasts. In closing, we would like to advocate for the broader use of concepts from the philosophy of language, such as vagueness, in challenging current models and providing additional insights beyond aggregate statistics and leaderboards.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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)*, pages 5103–5114.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.