Determining Question-Answer Plausibility in Crowdsourced Datasets Using Multi-Task Learning

Rachel Gardner   Maya Varma   Clare Zhu   Ranjay Krishna

Department of Computer Science, Stanford University, CA

{rachel0, mvarma2, clarezhu, ranjaykrishna}@cs.stanford.edu

Abstract

Datasets extracted from social networks and online forums are often prone to the pitfalls of natural language, namely the presence of unstructured and noisy data. In this work, we seek to enable the collection of high-quality question-answer datasets from social media by proposing a novel task for automated quality analysis and data cleaning: *question-answer (QA) plausibility*. Given a machine or user-generated question and a crowd-sourced response from a social media user, we determine if the question and response are valid; if so, we identify the answer within the free-form response.

We design BERT-based models to perform the QA plausibility task, and we evaluate the ability of our models to generate a clean, usable question-answer dataset. Our highest-performing approach consists of a single-task model which determines the plausibility of the question, followed by a multi-task model which evaluates the plausibility of the response as well as extracts answers (Question Plausibility AUROC=0.75, Response Plausibility AUROC=0.78, Answer Extraction F1=0.665).

1 Introduction

Large, densely-labeled datasets are a critical requirement for the creation of effective supervised learning models. The pressing need for high quantities of labeled data has led many researchers to collect data from social media platforms and online forums (Abu-El-Haija et al., 2016; Thomee et al., 2016; Go et al., 2009). Due to the presence of noise and the lack of structure that exist in these data sources, manual quality analysis (usually performed by paid crowdworkers) is necessary to extract structured labels, filter irrelevant examples, standardize language, and perform other preprocessing tasks before the data can be used. However,

obtaining dataset annotations in this manner is a time-consuming and expensive process that is often prone to errors.

In this work, we develop automated data cleaning and verification mechanisms for extracting high-quality data from social media platforms[1]. We specifically focus on the creation of question-answer datasets, in which each data instance consists of a question about a topic and the corresponding answer. In order to filter noise and improve data quality, we propose the task of *question-answer (QA) plausibility*, which includes the following three steps:

- *Determine question plausibility:* Depending on the type of dataset being constructed, the question posed to respondents may be generated by a machine or a human. We determine the likelihood that the question is both relevant and answerable.

- *Determine response plausibility:* We predict whether the user's response contains a reasonable answer to the question.

- *Extract answer from free-form response:* If the response is deemed to be plausible, we identify and extract the segment of the response that directly answers the question.

Because we assume social media users generally answer questions in good faith (and are posed questions which they can answer), we can assume plausible answers are correct ones (Park et al., 2019). Necessarily, if this property were not satisfied, then any adequate solutions would require the very domain knowledge of interest. Therefore, we look to apply this approach toward data with this property.

In this study, we demonstrate an application of QA plausibility in the context of visual question

---

[1]All code is available at github.com/rachel-1/qa_plausibility.

answering (VQA), a well-studied problem in the field of computer vision (Antol et al., 2015). We assemble a large VQA dataset with images collected from an image-sharing social network, machine-generated questions related to the content of the image, and responses from social media users. We then train a multitask BERT-based model and evaluate the ability of the model to perform the three subtasks associated with QA plausibility. The methods presented in this work hold potential for reducing the need for manual quality analysis of crowd-sourced data as well as enabling the use of question-answer data from unstructured environments such as social media platforms.

2 Related Work

Prior studies on the automated labeling task for datasets derived from social media typically focus on the generation of noisy labels; models trained on such datasets often rely on weak supervision to learn relevant patterns. However, approaches for noisy label generation, such as Snorkel (Ratner et al., 2017) and CurriculumNet (Guo et al., 2018), often use functions or other heuristics to generate labels. One such example is the Sentiment140 dataset, which consists of 1.6 million tweets labeled with corresponding sentiments based on the emojis present in the tweet (Go et al., 2009). In this case, the presence of just three category labels (positive, neutral, negative) simplifies the labeling task and reduces the effects of incorrect labels on trained models; however, this problem becomes increasingly more complex and difficult to automate as the number of annotation categories increases.

Previous researchers have studied question relevance by reasoning explicitly about the information available to answer the question. Several VQA studies have explicitly extracted premises, or assumptions made by questions, to determine if the original question is relevant to the provided image (Mahendru et al., 2017; Prabhakar et al., 2018). A number of machine comprehension models have been devised to determine the answerability of a question given a passage of text (Rajpurkar et al., 2018; Back et al., 2020). In contrast, we are able to leverage the user's freeform response to determine if the original question was valid. Our model is also tasked with supporting machine-generated questions, which may be unanswerable and lead to noisy user-generated responses.

While the concept of answer plausibility in user responses has also been previously explored, existing approaches use hand-crafted rules and knowledge sources (Smith et al., 2005). By using a learned approach, we give our system the flexibility to adapt with the data and cover a wider variety of cases.

3 Dataset

The dataset consists of questions and responses collected from an image-sharing social media platform. We utilize an automated question-generation bot in order to access public image posts, generate a question based on image features, and record data from users that replied to the question, as shown in Figure 1 (Krishna et al., 2019). Because the question-generation bot was designed to maximize information gain, it generates questions across a wide variety of categories, including objects, attributes, spatial relationships, and activities (among others). For the sake of space, we refer readers to the original paper for more information on the method of question generation and diversity of the resulting questions asked. All users that contributed to the construction of this dataset were informed that they were participating in a research study, and IRB approval was obtained for this work. For the privacy of our users, the dataset will not be released at this time. Rather than focus on the specific dataset, we wish to instead present a general method for cleaning user-generated datasets and argue its generality even to tasks such as visual-question-answering.
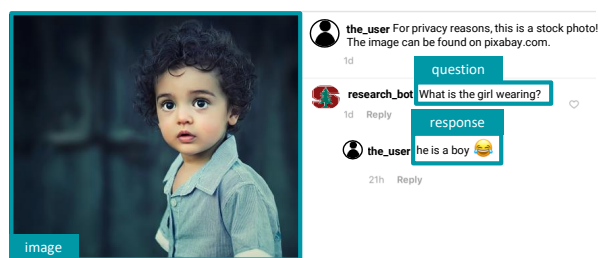

Figure 1: *An example question and response pair collected from social media.* Note that since the questions are generated by a bot, the question may not always be relevant to the image, as demonstrated here.

The dataset was labeled by crowdworkers on Amazon Mechanical Turk (AMT), who performed three annotation tasks, as shown in Table 1: (1) determine if the question was plausible, (2) deter-
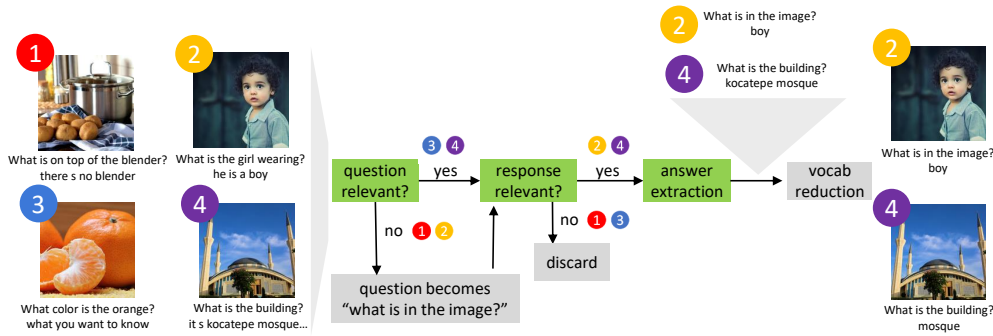
Figure 2: *Overview of the QA plausibility task, with representative examples.* Given a question and user response, we determine if the question and response are plausible given the image. If so, we then extract a structured answer label from the response.

#		Example Question/Response	Plausible?	%
1.	Q	What is on the table?	Y	50.6
	R	beet and carrot juice☺☺	Y	
2.	Q	What is the person doing?	Y	22.8
	R	not much lol	N	
3.	Q	What is on top of the cake?	N	11.4
	R	that is not cake that's chicken	Y	
4.	Q	What is the hamster doing?	N	15.3
	R	that is not a hamster	N	

Table 1: *Representative examples of cases present in the data, and the percentage of examples represented by each class in our dataset.* Examples (1) and (2) have valid questions that accurately refer to the corresponding images, while (3) and (4) do not correctly refer to objects in the image. However, in example (3), the user identifies the error made by the bot and correctly refers to the object in the image; as a result, this response is classified as valid.

mine if the response was plausible, and (3) if the response was deemed to be plausible, extract an answer span. Plausible questions and answers are defined as those that accurately refer to the content of the image.

It is important to note that since the question-generation process is automated, the question could be unrelated to the image due to bot errors; however, in such situations where the question is deemed to be implausible, the response may still be valid if it accurately refers to the content of the image. If the response is judged to be plausible, the AMT crowdworker must then extract the answer span from the user's response. In order to capture the level of detail we required (while discouraging AMT crowdworkers from simply copy/pasting the entire response), we set the maximum length of an answer span to be five words for the labeling step. However, the final model itself is not limited to answers of any particular length.

For cost reasons, each example was labeled by only one annotator. While we could have averaged labels across annotators, we found that the majority of the labeling errors were due to misunderstandings of the non-standard task, meaning that errors were localized to particular annotators rather than randomly spread across examples. This issue was mitigated by adding a qualifying task and manually reviewing a subset of labels per worker for the final data collection.

While one might expect images to be necessary (or at least helpful) for determining question and response plausibility, we found that human annotators were able to determine the validity of the inputs based solely on text without the need for the accompanying image. In our manual analysis of several hundred examples (approximately 5% of the dataset), we found that every example which required the image to label properly could be categorized as a "where" question. When the bot asked questions of the general form "where is the X" or "where was this taken," users assumed our bot had basic visual knowledge and was therefore asking a question not already answered by the image (such as "where is the dog now" or "what part of the world was this photo taken in"). This led to valid responses that did not pertain to image features and were therefore not helpful for training downstream models. Table 2 gives one such example. Once we removed these questions from the dataset, we could not find a single remaining example that required image data to label properly. As a result, we were able to explore the QA plausibility task in a VQA setting, despite not examining image features.

Our preprocessing steps and annotation procedure resulted in a total of 7200 question-response

	Example Question/Response		Valid?
	Q	Where is the dog?	Y
	R	sitting next to me on the sofa	N

Table 2: *Example requiring analysis of the original image (removed from dataset along with other "where" questions which often lead to confusion).*

pairs with answer labels. We use a standard split of 80% of the dataset for training, 10% for validation, and 10% for testing.

4 Models and Experiments

Model Architecture: As shown in Figure 3, we utilized a modified BERT model to perform the three sub-tasks associated with QA plausibility. The model accepts a concatenation of the machine-generated question and user response as input, with the [CLS] token inserted at the start of the sentence and the [SEP] token inserted to separate the question and response.
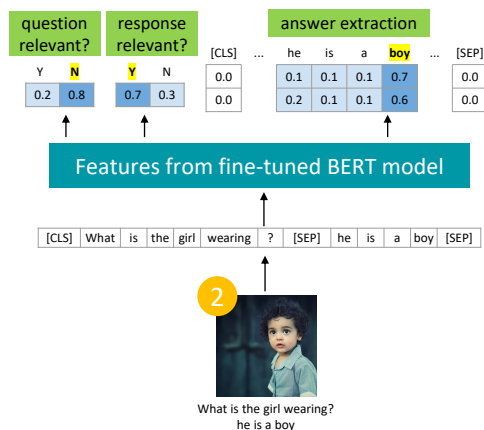

Figure 3: *Model architecture.* The question and user response serve as input to a modified BERT model, which will output question plausibility, response plausibility, and an answer label.

In order to perform the question plausibility classification task, the pooled transformer output is passed through a dropout layer (p=0.5), fully connected layer, and a softmax activation function. An identical approach is used for response plausibility classification. To extract the answer span, encoded hidden states corresponding to the last attention block are passed through a single fully connected layer and softmax activation; this yields two probability distributions over tokens, with the first representing the start token and the second representing the end token. The final model output includes the probability that the question and response are

plausible, with each expressed as a score between 0 and 1; if the response is deemed to be plausible, the model also provides the answer label, which is expressed as a substring of the user response.

Experiments: We utilized a pretrained BERT Base Uncased model, which has 12 layers, 110 million parameters, a hidden layer size of 768, and a vocabulary size of 30,522. We trained several single-task and multi-task variants of our model in order to measure performance on the three subtasks associated with QA plausibility. In the multi-task setting, loss values from the separate tasks are combined; however, an exception to this exists if the user's response is classified as implausible. In these cases, the answer span extraction loss is manually set to zero and the answer extraction head is not updated.

We evaluated performance on question and response plausibilities by computing accuracy and AUC-ROC scores. Performance on the answer span extraction task was evaluated with F1 scores, which measure overlap between the predicted answer label and the true answer (Rajpurkar et al., 2018).

5 Results

We investigated performance of our BERT model on the various subtasks associated with QA plausibility. Results are summarized in Table 3. Single-task models trained individually on the subtasks achieved an AUC-ROC score of 0.75 on the question plausibility task, an AUC-ROC score of 0.77 on the response plausibility task, and an F1 score of 0.568 on the answer extraction task. A multi-task model trained simultaneously on all three tasks demonstrated decreased performance on the question and response plausibility tasks when compared to the single-task models. We found that the highest performance was achieved when a single-task model trained on the question plausibility task was followed by a multi-task model trained on both the response plausibility and answer extraction tasks; this model achieved an AUC-ROC score of 0.75 on question plausibility, an AUC-ROC score of 0.79 on response plausibility, and an F1 score of 0.665 on answer extraction.

Our results suggest that multi-task learning is most effective when the tasks are closely related, such as with response plausibility and answer extraction. Since the BERT architecture is extremely quick for both training and evaluation, we found that the increase in performance afforded by using a single-task model and multi-task model in series

Combined Task	Question Plausibility		Response Plausibility		Answer Extraction
	Acc	AUROC	Acc	AUROC	F1
Question Plausibility (QP) only	**65.51%**	**0.7488**	-	-	-
Response Plausibility (RP) only	-	-	64.62%	0.7674	-
Answer Extraction only	-	-	-	-	0.568
RP and Answer Extraction	-	-	**70.13%**	**0.7870**	**0.665**
QP, RP and Answer Extraction	63.90%	0.6803	60.91%	0.6881	0.6160

Table 3: *Model Evaluation Metrics.* Performance metrics of our model are shown here. Multi-task learning helps improve performance when the model is simultaneously trained on the response plausibility and answer extraction subtasks, but decreases performance when the model is simultaneously trained on all three subtasks.

was worth the overhead of training two separate models. It is worth noting that a more complicated model architecture might have been able to better accommodate the loss terms from all three subtasks, but we leave such efforts to future work.

6 Discussion

Deep learning studies are often hindered by lack of access to large datasets with accurate labels. In this paper, we introduced the question-answer plausibility task in an effort to automate the data cleaning process for question-answer datasets collected from social media. We then presented a multi-task deep learning model based on BERT, which accurately identified the plausibility of machine-generated questions and user responses as well as extracted structured answer labels. Although we specifically focused on the visual question answering problem in this paper, we expect that our results will be useful for other question-answer scenarios, such as in settings where questions are user-generated or images are not available.

Overall, our approach can help improve the deep learning workflow by processing and cleaning the noisy and unstructured natural language text available on social media platforms. Ultimately, our work can enable the generation of large-scale, high-quality datasets for artificial intelligence models.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering.

Seohyun Back, Sai Chetan Chinthakindi, Akhil Kedia, Haejun Lee, and J. Choo. 2020. Neurquri: Neural question requirement inspector for answerability prediction in machine reading comprehension. In *ICLR*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision.

Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. 2018. Curriculumnet: Weakly supervised learning from large-scale web images. *CoRR*, abs/1808.01097.

Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. *CoRR*, abs/1903.11207.

Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra, and Stefan Lee. 2017. The promise of premise: Harnessing question premises in visual question answering. *EMNLP 2017*, abs/1705.00601.

Junwon Park, Ranjay Krishna, Pranav Khadpe, Li Fei-Fei, and Michael Bernstein. 2019. Ai-based request augmentation to increase crowdsourcing participation. *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*.

Prakruthi Prabhakar, Nitish Kulkarni, and Linghao Zhang. 2018. Question relevance in visual question answering. *arXiv preprint abs/1807.08435*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *Association for Computational Linguistics (ACL)*, abs/1806.03822.

Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason Alan Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *CoRR*, abs/1711.10160.

Troy Smith, Thomas M. Repede, and Steven L. Lytinen. 2005. Determining the plausibility of answers to questions. *American Association for Artificial Intelligence*.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m. *Communications of the ACM*, 59(2):64–73.