Weaknesses of Exact Match Accuracy in Open-Domain Question Answering

Caleb Kumar - 2022-04-04

Abstract

Open-Domain Question Answering benchmarks such as Natural Questions ideally serve as a gauge for showing how well a model is able to answer generic factoid questions. We have found that using Exact Match as a scoring mechanism for these benchmarks can be ineffective as it can falsely mark predictions as incorrect. This effect is likely further exacerbated when test-train overlap in a dataset is removed. Additionally, we show that removing test-train overlap from ODQA datasets such as Natural Questions can introduce bias.

1 Introduction

Open-Domain Question Answering (ODQA) systems produce answers to natural language factoid questions in any domain. For example, given the question, “Where is the world’s largest ice sheet located today?” we would expect the system to produce the answer “Antarctica". Building Question Answering systems can have many applications. From creating an AI assistant to solving search engine questions, being able to derive answers for a user’s natural language query can provide immense value.

Training ODQA models is challenging as they often need to be quite large in order to store the necessary amount of parameters to solve generic questions. Building benchmarks for these models is another difficult task, particularly because these datasets must maintain an even distribution of generic questions and answers. It has been shown that there is a significant amount of overlap between the test and train portions of these benchmarks as demonstrated in Natural Questions, WebQuestions, and

```
def normalize_answer(s):
    """Lower text and remove punctuation,
articles and extra whitespace."""

    def remove_articles(text):
        return re.sub(r"\b(a|an|the)\b", " ",
text)

    def white_space_fix(text):
        return " ".join(text.split())

    def remove_punc(text):
        exclude = set(string.punctuation)
        return "".join(ch for ch in text if ch
not in exclude)

    def lower(text):
        return text.lower()

    return
white_space_fix(remove_articles(remove_punc(low
er(s))))
```

Figure 1: This function is used in Exact Match scoring to normalize predicted and annotated answers before checking if they exactly match character by character. The text is lowercased, whitespace normalized, and punctuations and articles are removed.

TriviaQA (Lewis et. al, 2020). Models are ideally able to solve questions without memorization of training data. When this overlap is removed from the benchmarks, many of the current best performing models score much worse (Table 1). The scoring for these datasets is done through Exact Match scoring where the model’s answer must entirely match the annotated answer character by character after some light normalization (Figure 1). Our central hypothesis is that this Exact Match scoring is unfair when evaluating an ODQA model because it is unlikely that the predicted answer will match character by character to the annotated answer when the model has not previously seen the question or answer. In this paper we will show a non-trivial amount of cases where Exact Match incorrectly labels a prediction. However, we will not provide the exact metrics for how many incorrect labels there were since this would require human annotation efforts. Additionally,

we will show that removing the overlap from these datasets introduces bias. We will do so by training two models that, given a Natural Question's test dataset question or answer will predict whether it contains test-train overlap.

These two findings reveal some shortcomings of modern ODQA modeling. The first is that Exact Match scoring struggles to be a fair evaluation metric in some cases. We suggest that this effect is exacerbated when overlap is removed between training and test ODQA datasets such as Natural Questions. The second finding is that, removing overlap from ODQA datasets is not a defacto solution to avoiding a model only reproducing memorized answers from the training dataset since it introduces bias.

2 Related Work

The primary ODQA benchmark we will focus on is Natural Questions. It is the first publicly available dataset that pairs real user queries to high-quality annotations in a document. Human performance was able to achieve 87.2 and 75.7 Exact Match accuracy on long answer and short answer questions respectively (Kwiatkowski et. al, 2019). Its creation is undoubtedly a significant milestone in the ODQA community. However, there are still some shortcomings to the dataset. As is pointed out by Lewis, Stenetorp, and Riedel (2020), there is a large amount of overlap between test and train examples leading to systems potentially memorizing answers rather than generating the answers on their own (Table 1).

Dataset	% Answer overlap	% Question overlap
Natural Questions	63.6	32.5
TriviaQA	71.7	33.6
WebQuestions	57.9	27.5

Table 1: Shows the amount of test-train overlap in the popular ODQA benchmarks (Lewis et. al. 2020)

In recent history generative models such as Fusion-in-Decoder, GPT-3, Google's t5 models and others have shown promise in solving ODQA (Weng 2020). These models vary in whether they are open-book, where they actively look for answers in a knowledge base such as Wikipedia, or closed-book, where they do not. Typically models that are closed-book are of the largest size. In this paper we will solely be experimenting with generative models as these models help demonstrate the weaknesses of Exact Match scoring.

When a generative ODQA model is open-book such as Fusion-in-Decoder, it requires a set of relevant documents before generating an answer. DPR, a dense retrieval model, has been demonstrated to be effective in retrieving relevant documents for a given question (Karpukhin et. al, 2020). DPR's embeddings are learned for questions and passages with two encoders. It finds the passage that is most similar to the question using the dot-product as the measure for similarity. All passage embeddings are using FAISS indexing to speed up the search process. For our work we will use DPR as a retriever of relevant passages for our models GPT-3 and Fusion-in-Decoder.

3 Data.

In our analysis we solely use Open-Natural Questions. Our findings likely apply to the other common datasets: TriviaQA and Webquestions, however, due to time constraints we did not include them.

Open-Natural Questions examples consist of a question, long answer(s), and short answer(s). For our experiment we omitted using the long answer(s) for an example. Note that there can be multiple short answers since more than one annotator produced an answer for each question. Each question must be longer than 7 tokens and start with the tokens "who", "when", or "where". Furthermore, the questions are anonymized and retrieved from actual user searches. We use the standard open-domain version of the dataset consisting of 79,168 train, 8,757 development and 3,610 test question answer examples. We use the annotations done by Lewis et. al, 2020 to obtain the test-train overlap.

4 Models.

We conduct our experiments with the models GPT-3, t5-xl-ssm-nq, t5-large-ssm-nq, and Fusion-in-Decoder. The model version of Fusion-in-Decoder we settle on is FiD-large.

DPR is used as a retriever for GPT-3 and Fusion-in-Decoder.

GPT-3 is an autoregressive model that is the largest of our choices having 175 billion parameters. We relied on the Open AI API to obtain predictions for the model. We provided three question-answer pair examples per inference call and had DPR retrieve the five most relevant wikipedia passages for every given question. It should be mentioned that GPT-3 was used conservatively due to financial constraints. The model was only run on 730 test examples from Natural Questions.

t5-xl-ssm-nq and **t5-large-ssm-nq** are both generative t5 models fine-tuned on all train splits for Natural Questions for 10k steps. Similar to GPT-3, t5 models use an encoder-decoder architecture. They are pre-trained on a multi-task mixture of unsupervised and supervised tasks where each task is converted to a text-to-text format. Both models were downloaded using the Hugging Face repository.

FiD-large is a closed-book model that is based on t5-large. It uses DPR to retrieve relevant passages and then fuses them together in an encoder and then its decoder can search for the answer in the encoded passages. Like the other t5 based models it has been fine-tuned on the Natural Questions training data.

5 Experiments

We evaluated each model on the Natural Questions test dataset and produced three different scores to evaluate their performance—Exact Match, BERT Score, and Meteor.

BERT Score is an evaluation metric typically used for evaluating text generation. It is obtained by computing the similarity score for each token in the candidate sentence with each token in the reference sentence. We use the embeddings of the Roberta-large model.

Meteor is an evaluation metric typically used for evaluating machine translation. It is scored somewhat similarly to BLEU except that it provides a penalty function for incorrect word order and allows for usage of some synonyms. It does so by computing a weighted F-score and mapping unigrams through Porter stemming and WordNet synonymy.

Exact Match is checking whether the prediction exactly matches one of the annotated answers with some light normalization (Figure 1).

As discussed earlier in the paper, we used the annotations provided by Lewis et. al, 2020 to partition the results by the overlap categories—question overlap, answer overlap only, and no overlap (Table 2, 3, & 4).

Exact Match				
model	**total**	**question overlap**	**answer overlap only**	**no overlap**
t5-large-ssm-nq	28.89	70.68	13.33	2.24
t5-xl-ssm-nq	32.96	73.15	19.05	5.60
GPT-3	28.08	38.57	29.51	15.00
FiD-large	53.13	76.23	47.62	37.25

Table 2: The Exact Match scores of the listed models on the Natural Questions test dataset

BERT Score				
model	**total**	**question overlap**	**answer overlap only**	**no overlap**
t5-large-ssm-nq	50.52	79.20	46.35	24.55
t5-xl-ssm-nq	53.99	81.37	47.36	29.79
GPT-3	47.71	59.41	51.19	33.80
FiD-large	68.36	83.64	67.46	55.06

Table 3: The BERT Scores of the listed models on the Natural Questions test dataset

Meteor				
model	**total**	**question overlap**	**answer overlap only**	**no overlap**
t5-large-ssm-nq	28.72	64.81	16.45	5.73
t5-xl-ssm-nq	32.25	66.47	20.08	9.16
GPT-3	31.23	42.08	30.95	19.14
FiD-large	47.88	66.09	41.74	38.19

Table 4: The BERT Scores of the listed models on the Natural Questions test dataset

All the models performed the worst when there was no overlap in the question-answer pairs. It is notable that GPT-3 performed worse on non-overlap examples scoring 15.00 because overlap should not be a factor in its performance since it was not fine-tuned on the Natural Questions (Table 2).

Another interesting result is that FiD-large scored higher than shown in Lewis et. al, 2020 for exact match accuracy with 37.25 (previously scored 34.5). This was likely due to the recent improvements to the model (Izacard et. al, 2021).

The models' scores for the Meteor and Exact Match are highly similar. The correlation coefficient between their scores for FiD-Large are 0.990 whereas for Meteor and BERT Score their correlation score is only 0.946. One observation is that the BERT scores are generally larger and show less decline than Meteor and Exact Match accuracy when overlap is removed from the dataset.

6 Analysis

Exact Match scoring is one of the most common metrics to evaluate the performance of models for ODQA tasks such as Natural Questions. By using BERT Score and Meteor we were able to identify numerous cases where the model produced the correct answers but Exact Match scoring failed to identify them. We did so by examining examples without test-train overlap and with high BERT or Meteor scores (Table 6 & 7).

question	answers	prediction	exact match	bert score
who was assassinated during a visit to sarajevo in bosnia	['Archduke Franz Ferdinand of Austria']	Archduke Franz Ferdinand	FALSE	82.54
what is the main mineral in lithium batteries	['lithium', 'Lithium']	lithium ions	FALSE	76.35
what was the religion in persia before islam	['Zoroastrian', 'the Zoroastrian religion']	Zoroastrian ism	FALSE	74.50
which mirror is used in vehicles for rear view	['rear - view mirror']	rear-view mirror	FALSE	72.73
where did immigrants enter the us on the west coast	['Angel Island Immigration Station', 'San Francisco Bay']	San Francisco	FALSE	72.44
the cuban revolt against spain was led by	['José Martí', 'Antonio Maceo', 'Máximo Gomez']	José Mart	FALSE	69.22
the general term for software that is designed to damage disable or steal data is	['Malware']	Computer malware	FALSE	57.70

Table 5: Examples where Exact Match scoring was incorrect identified by high BERT Scores for t5-xl-ssm-nq.

question	answers	prediction	exact match	meteor score
a good that can be used in place of another good	['substitute good', 'A substitute good']	substitute goods	FALSE	93.75
which episode does gideon die in criminal minds	["`` Nelson 's Sparrow "", "Nelson 's Sparrow"]	"Nelson's Sparrow"	FALSE	92.01
one piece english dubbed episode 564 release date	['September 16 , 2012']	September 16, 2018	FALSE	73.61
what is the meaning of auv in cars	['action utility vehicles']	utility vehicle	FALSE	64.66
where did immigrants enter the us on the west coast	['Angel Island Immigration Station', 'San Francisco Bay']	San Francisco	FALSE	64.66
who was assassinated during a visit to sarajevo in bosnia	['Archduke Franz Ferdinand of Austria']	Archduke Franz Ferdinand	FALSE	61.34
what was the religion in persia before islam	['Zoroastrian', 'the Zoroastrian religion']	Zoroastria nism	FALSE	50.00
who are the only 2 families that have had 3 generations of oscar winners	['Farrow / Previn / Allens', 'the Coppolas', 'Hustons', 'The Hustons', 'Coppolas']	Coppola	FALSE	50.00
what happens when iron reacts with oxygen and water	['Rust']	rusting	FALSE	50.00

Table 6: Examples where Exact Match scoring was incorrect identified by high Meteor scores for t5-xl-ssm-nq.

While we did not compute the number of incorrect labels produced by Exact Match scoring since this would require significant annotation efforts, we have demonstrated that it

is a non-trivial amount. We suspect that this is due to the model's inability to generate an answer exactly matching the annotated answer without having seen the specific answer-question pair in the training dataset.

Recall that GPT-3 scored the lowest on examples with no overlap (Table 2, 3, & 4). This is a surprising result because GPT-3 was not fine-tuned on the Natural Questions dataset and hence its performance should not be affected by the removal of test-train overlap. This suggests that there might be inherent differences between the examples with overlap and without overlap. To test this theory we trained two classifiers to detect overlap for a given test example. The first would take a test dataset answer and predict whether or not it contained overlap to the training dataset without having ever seen the training dataset. Similarly the second classifier would be trained to predict whether a question from the test dataset contained overlap with the training dataset.

Task	Model	precision	recall	f1	accuracy
Answer Overlap	bert-base	74.04	73.68	73.84	73.68
Question Overlap	bert-base	68.2	65	65.98	65

Table 7: Given an answer or question from the Natural Question Test dataset, predict whether there is overlap to the training dataset. We used a standard 80/20 train-test split of the Natural Questions test dataset.

As is shown in Table 7, both classifiers, while not perfect, were able to predict overlap with statistical significance. This would suggest that there is some bias created in the test dataset when overlap is removed. This conceived bias is likely why GPT-3 scores lower on non-overlap examples. Overall, this bias suggests to take caution before testing a model on the non-overlap examples from Natural Questions because they are more challenging examples to solve with the effects of overlap removal held constant.

7 Conclusion

In this paper we have discovered a few findings related to the Natural Questions dataset. Exact Match contains some weaknesses when evaluating model performance on this dataset. By including Meteor and BERT Score we detected many cases where Exact Match incorrectly penalized the model. We suspect that this effect is only exacerbated when test examples that have overlap from the training data are omitted. Additionally, we showed that removing test examples that have test-train overlap can lead to a biased dataset where the examples are naturally more challenging with the effects of overlap removal held constant.

We believe these findings help identify improvements that can be made in the ODQA community. We suggest that there needs to be some investment made into improving the metrics surrounding evaluation of these models. Ideally predicted answers would be able to be scored against annotated answers in a more fuzzy manner. Perhaps a standard model could be trained to determine whether a generated prediction matches an annotated answer. We hope that these findings will be taken into account when ODQA models are evaluated and ODQA datasets are being constructed.

8 Acknowledgements

We would like to thank Professor Christopher Potts for prompting us to choose a topic of our liking and offering guidance on how to approach the ODQA domain during office hours. We are grateful to the Hugging Face and Open AI organizations for hosting powerful Transformer models which are not easily inferencable on basic computing hardware. Furthermore, we would like to thank Gautier Izacard for providing the FiD-large model test results for Natural Questions directly to us. Retrieving the results is not trivial even if the model is downloadable. We would like to thank Jazlyn Akaka for helping us proofread our works. Finally, the greatest acknowledgement is for a debt redeemed by Jesus Christ. By His grace and mercy can I strive to do all things to the glory of God.

9 Authorship

Caleb Kumar - This author is primarily responsible for the work completed on this project. Outside of guidance from Professor Potts, proofreading from Jazlyn Akaka, and the test results obtained from Gautier Izacard, Caleb was entirely responsible for the hypothesis, experiments, and findings found in this paper.

References

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. *BERTSCORE: EVALUATING TEXT GENERATION WITH BERT*. ICLR

Cunxiang Wang, Pai Liu, and Yue Zhang. 2021. *Can Generative Pre-trained Language Models Serve as Knowledge Bases for Closed-book QA?* Institute of Advanced Technology, Westlake Institute for Advanced Study, China.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. 2020. *Dense Passage Retrieval for Open-Domain Question Answering.* Association for Computational Linguistics

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. 2019. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* Journal of Machine Learning Research 21 (2020) 1-67

Lilian Weng. 2020. *How to Build an Open-Domain Question Answering System?* Blogpost

Desh Raj. 2017. *Publications Manual.* Metrics for NLG evaluation. Blogpost

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, Slav Petrov. 2019. *Natural Questions: A Benchmark for Question Answering Research.* Association for Computational Linguistics

Gautier Izacard, Edouard Grave. 2021. *Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering*, Association for Computational Linguistics

Patrick Lewis, Pontus Stenetorp, Sebastian Riedel. 2020. *Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets*. Association for Computational Linguistics

Jonathan Berantm Andrew Chou, Roy Frostig, Percy Liang. 2013. *Semantic Parsing on Freebase from Question-Answer Pairs.* Association for Computational Linguistics