A Statistical Approach to Analyzing and Mitigating Question-Answering Artifacts in SQuAD

Hubert Luo [1]

Abstract

A statistical approach was used to analyze the behaviour of pre-trained question-answering models, identifying areas where it demonstrates poor textual and logical understanding. This was improved using an adversarial challenge, showing some improvement in questions that require a higher degree of logical understanding and extrapolation.

1. Introduction

Pre-trained question-answering models generally perform well on the datasets they are trained on. However, this may not reflect true understanding of the underlying text rather than simply learning and regurgitating patterns based on the data it was trained on.

Therefore, a statistical approach was used to inspect the behaviour of these pre-trained models, drawing inspiration from previous work (Gardner et al., 2021) to identify cases where the model demonstrates poor textual and logical understanding.

An adversarial challenge set was then used to improve on the areas identified by this statistical framework, demonstrating improvement in some areas where a higher degree of logical understanding and extrapolation was required from the underlying text.

2. Previous Work

The ELECTRA-small (Clark et al., 2020) model was used as the preliminary model given its flexibility and architecture. The SQuAD dataset (Rajpurkar et al., 2016) was also used to train and evaluate the model.

A "competency problems" framework to find spurious n-gram correlations with answers was previously found (Gardner et al., 2021) and this was the inspiration for using a statistical perspective to analyze model behaviour and identify trends in model performance.

Adversarial data augmentation to the SQuAD dataset was then leveraged using previous work by (Jia & Liang, 2017) to improve areas where the model underperformed.

Table 1. Most-Prevalant Question Types in SQuAD evaluation set.

QUESTION TYPE	QUESTION FREQUENCY	EXAMPLE
WHAT	0.449	"WHAT IS PPP?"
HOW	0.103	"HOW DID WAR START?"
WHO	0.100	"WHO WAS WARSZ?"
WHEN	0.066	"WHEN DID TEMUR RULE?"
WHICH	0.043	"WHICH COMPANY OWNS ABC?"

3. Method

A statistical approach leveraging permutation testing was used to identify patterns in correctness of pre-trained models. Specifically, the ELECTRA-small (Clark et al., 2020) model was first trained on the SquAD dataset (Rajpurkar et al., 2016), and its performance on the corresponding evaluation set was analyzed.

3.1. Model Analysis: Feature Engineering

Each question in the evaluation set was classified as a question type depending on the initial words in the question. For example, a question such as "Why are the small lakes in the parks emptied before winter?" was classified as a "Why" question. Most-prevalent question types, their frequencies, and an example question of each type are shown in table 1.

Another feature that was created depending on whether the answer was numeric. For example, the question "Super Bowl 50 decided the NFL champion for what season?" was a numerical question since the answer was the numeric value 2015.

Furthermore, features were created to account for the num-

*Equal contribution [1]Department of Computer Science, University of Texas at Austin.

ber of characters, or length, of not only the context, but also the question. Figure 1 shows context length was right skewed with a long right tail. For context lengths under 500 or over 1000 characters, there was slightly higher frequency of question contexts which the pre-trained model answered incorrectly. On the other hand for context lengths between 500-1000, there was slightly higher frequency of question contexts which the pre-trained model answered correctly.

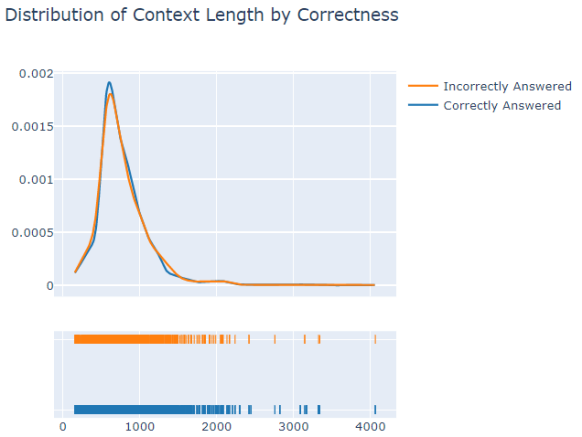

Figure 1. Distribution of Context Length by Correctness.

Figure 2 demonstrates a right-skewed distribution and correctness trend for question lengths. Specifically, question lengths under 45 and over 75 saw slightly higher frequency of questions which the pre-trained model answered incorrectly. However, the opposite was true for question lengths between 45-75 which had a slightly higher frequency of questions which the pre-trained model answered correctly.

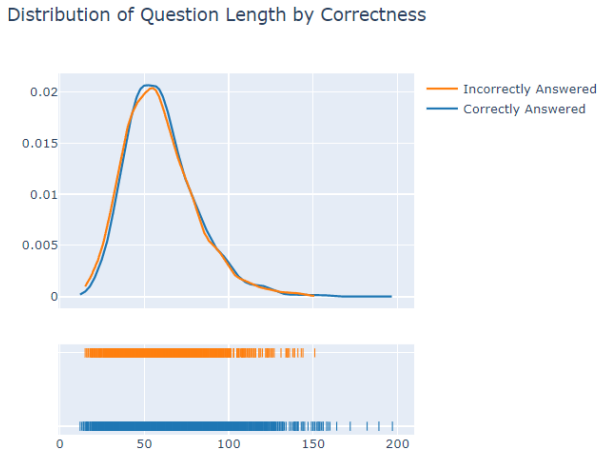

Figure 2. Distribution of Question Length by Correctness.

Both context and question lengths were therefore binned into three categories: under 500, 500-1000, and over 1000 for context lengths; under 45, 45-75, and over 75 for question lengths given the observations outlined and demonstrated earlier in figures 1 and 2.

Table 2. Permutation Tests for Categorical Variables.

VARIABLE	OBSERVED TVD	P-VALUE
QUESTION TYPE	1.402	0.008*
QUESTION LENGTH (BIN)	0.016	0.019
CONTEXT LENGTH (BIN)	0.011	0.276
(*): SIGNIFICANT		

3.2. Model Analysis: Permutation Testing for Categorical Variables

A two-sided permutation test was first conducted for categorical columns with more than two categories, specifically question types, context lengths (binned), and question lengths (binned).

The null hypothesis H_0 was the categorical variable, such as question type, has no impact on correctness - any observed differences in correctness between question types was due to random chance alone. The alternative hypothesis H_1 was there was some impact of question types on the distribution of correctness, something other than random chance alone. Given there are more than two possible question types, Total Variation Distance (TVD) was used as the test statistic:

$$N = \#\text{ possible question types} \tag{1}$$

$$\mu = \%\text{ correct for evaluation set} \tag{2}$$

$$X_i = \%\text{ correct for question type } i \tag{3}$$

$$TVD = \frac{1}{2}\sum_{i=1}^{N} |X_i - \mu| \tag{4}$$

The observed TVD was first calculated for each variable, such as for question type. Each simulation, labels of these variables were shuffled without replacement and the TVD was calculated on this simulated data. This was then repeated for a million simulations. The p-value was the percentage of simulated TVD's greater than or equal to the observed TVD, which supports the alternative hypothesis. Summary of permutation test results are in table 2.

An asterisk in table 2 denotes a p-value significant on a 5% threshold, adjusted using a Bonferroni correction to control family-wide error rate, which is the probability of Type I (false positive) errors when running multiple permutation tests. The Bonferroni correction is calculated below:

$$N = \text{\# permutation tests} \quad (5)$$

$$\alpha = \text{significance level} \quad (6)$$

$$\alpha_{\text{adjusted}} = \frac{\alpha}{N} \quad (7)$$

Given there were three permutation tests using this TVD statistic where $N = 3$, the adjusted p-value used here was $\frac{0.05}{3} \approx 0.0167$.

The p-value for question types is under this threshold. Therefore, we reject the null hypothesis and find evidence which supports our alternative hypothesis that there is some dependency between question types and correctness.

However, for both question length and context length, the p-value is greater than the threshold after applying a Bonferroni correction - therefore, we fail to reject the null hypothesis that the observed patterns between question/context length and correctness are due to random chance.

3.3. Model Analysis: Permutation Testing for Binary Variables

Given the earlier permutation test on categorical variables found there was some relationship between question type and correctness in the observed results of the pre-trained model beyond random chance alone, this next section focuses specifically on question types.

First, one-hot encoding was applied to all question types which occurred at least ten times. This was to better understand which specific question types are associated with less accurate results in the pre-trained model, without being mislead by false trends in question types which rarely occurred in the data.

Next, an one-sided permutation test was conducted on these binary variables. This was different from the two-sided permutation tests conducted on categorical variables earlier. The null hypothesis H_0 was the question type, such as "Why" questions, has no impact on correctness - difference in correctness in "Why" questions was due to random chance alone. The alternative hypothesis H_1 was correctness was more likely to decrease for "Why" questions and this was due to something other than random chance alone.

Given this was now a one-sided permutation test on a binary variable, the test statistic used was the delta: % Correct for non-"Why" questions subtracted by the % Correct of "Why" questions. High values of this test statistic supported the alternative hypothesis that "Why" questions have lower correctness due to something other than random chance. Note this was repeated for each question type that occurred at least ten times.

The observed delta was first calculated for each variable, such as for "Why" questions. Each simulation, labels of these variables were shuffled without replacement and the delta was calculated on this simulated data. This was then repeated for a million simulations. The p-value was the percentage of simulated deltas greater than or equal to the observed delta which supported the alternative hypothesis. Summary of permutation test results are in table 3

Table 3. Permutation Test for Selected Binary Variables.

VARIABLE	OBSERVED DELTA	P-VALUE
WHY	0.154	< 0.001*
WHAT	0.046	< 0.001*
NAME	0.186	0.101
HOW	0.008	0.265
WHICH	0.010	0.286
WHO	-0.062	1.000
WHEN	-0.109	1.000
(*): SIGNIFICANT		

An asterisk in table 3 denotes a p-value significant on a 5% threshold, adjusted using a Bonferroni correction to control family-wide error rate, as explained in the section above.

There were 38 permutation tests done on these binary variables where $N = 38$, the adjusted p-value used here was $\frac{0.05}{38} \approx 0.00132$.

The p-value for "Why" and "What" question types is under this threshold. Therefore, we reject the null hypothesis and find evidence which supports our alternative hypothesis that the lower correctness the pre-trained model has on these question types is due to something other than random chance alone.

However, for all other question types, the p-value is greater than the threshold after applying a Bonferroni correction - therefore, we fail to reject the null hypothesis that these have an inherently lower correctness and such patterns to that extent are due to random chance alone for non-Why/What questions.

3.4. Model Improvement: Adversarial Challenge Set

In earlier sections, it was found that the pre-trained model struggled with "Why" and "What" questions due to something other than random chance alone. These are question types which require a question-answering model to potentially infer facts and statements in the context, therefore necessitating a greater logical understanding of the underlying text.

Therefore, errors to these kind of questions could signify the pre-trained model was only learning a surface-level understanding of the text instead of a true understanding of its

Table 4. Selected delta changes after adversarial data augmentation.

QUESTION TYPE	ADVERSARIAL DELTA	ORIGINAL DELTA	DIFFERENCE
WHERE	0.033	0.007	0.041
WHY	0.175	0.154	0.014
WHAT	0.041	0.046	-0.005
WHAT'S	-0.008	0.032	-0.039
WHOSE	-0.121	-0.082	-0.039

content. This was reinforced as the pre-trained model far outperforms more straightforward questions such as "When" and "Who" questions relative to complex questions such as "Why" and "What".

Therefore, an adversarial SQuAD challenge set (Jia & Liang, 2017) was used to fine-tune the pre-trained model to make it more robust and teach it to better learn the true understanding of contexts rather than just surface-level representations.

This adversarial data augmentation consisted of automatically generating up to five candidate adversarial sentences which did not actually answer the question, however were potentially misleading for the model. (Jia & Liang, 2017) Results are outlined in the following section.

4. Results

4.1. Model Results

Results of the model after fine-tuning on the adversarial challenge set outperformed the pre-trained model on specific subsets of questions, however underperformed on the questions in the shared evaluation set as a whole.

Overall, the F_1 score of the model dropped from 0.864 to 0.830 when evaluating on the same SQuAD evaluation set. This may have been due to the fact that inserting adversarial candidate sentences confused the model as these sentences were purposefully crafted to be misleading and close to the correct sentence.

In Table 4, the delta was the same as the test statistic used in section 3.3 - it was the % Correct for all questions not of a specific question type subtracted by the % Correct of questions of that question type. High values of this delta demonstrated the model struggled with answering questions of this type relative to other question types.

Table 4 shows questions of type "What's" has a delta of 0.032 on the original model, demonstrating the pre-trained model struggles to answer questions of type "What's", more so than other question types.

However, after adversarial data augmentation, the delta is now -0.008, decreasing by 0.039. This demonstrates the model now performs better on questions of this type "What's" relative to other question types after leveraging this adversarial challenge set.

Similar results are seen for question types such as "What" and "Whose". These are especially noteworthy as section 3.3 previously identified that the pre-trained model underperformed on questions of type "What" and this was likely due to something other than random chance alone. Using this adversarial challenge set approach decreases underperformance on questions of "What" questions.

On the other hand, however, the model's performance decreased on questions of type "Where" in particular after applying an adversarial challenge set. This was likely due to the fact adversarial questions were often crafted in such a way that directly impacted the location, thus having a strong misleading effect on the model during training.

5. Conclusion

Overall, an adversarial challenge set was helpful in boosting performance on challenging subsets of the SQuAD data that the pre-trained question-answering model struggled with in particular, as identified by a statistical testing framework. These were often areas that required models to have a deeper logical and textual understanding, beyond the surface-level associations and patterns that the pre-trained model may have learned more quickly.

References

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. Electra: Pre-training text encoders as discriminators rather than generators, 2020. URL https://arxiv.org/abs/2003.10555.

Gardner, M., Merrill, W., Dodge, J., Peters, M. E., Ross, A., Singh, S., and Smith, N. A. Competency problems: On finding and removing artifacts in language data, 2021. URL https://arxiv.org/abs/2104.08646.

Jia, R. and Liang, P. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL https://www.aclweb.org/anthology/D17-1215.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text, 2016.