Comparing methods to improve performance on why questions in QA tasks

1 Introduction

This project examines the performance of the RoBERTa model on a Question Answering (QA) task using the SQuAD dataset. It highlights the model's limitations in addressing ”why” questions and those requiring precise answer selection from contexts containing multiple potentially relevant pieces of information. To address these challenges, the study explores data augmentation techniques and leverages a specialized dataset, Quoref, designed to enhance model performance on complex QA tasks. The results provide insights into the constraints of current data augmentation methods in QA tasks and emphasize the role of dataset design and augmentation quality in achieving robust, generalizable model performance.

2 Analysis of the performance of RoBERTa on SQUAD

2.1 General results

I started by assessing the performance of RoBERTa on the Stanford Question Answering Dataset (SQuAD). This dataset contains 107,785 question-answer pairs on 536 articles [Rajpurkar et al, 2016]. Questions are accompanied by a context, an answer, and an answer start that indicates the string within the context where the answer begins. Ideally, the answer to the question should be inferable from the context provided. Running out-of-the-box (that is, untrained) RoBERTa on the validation set resulted in unsatisfactory performance, indicating that although the model has been previously trained, it is not fine-tuned to the specific question-answering (QA) task required by the SQuAD dataset. I trained the model for three epochs on the training set and assessed the performance on the validation set. The results are summarized in Table 1.

As Table 1 shows, the results after training are markedly higher and already similar to the 86.8 F1 obtained by human annotators on the dataset [Rajpurkar et

Model	Exact Match (%)	F1 Score (%)
Out-of-the-box	0.95	6.44
Trained	78.57	86.22

Table 1: Out-of-the-box and trained RoBERTa performance in QA task on the validation set.

al, 2016]. However, while the model performs well on this dataset, it may have problems generalizing to different QA tasks. An analysis of the errors made by the model can potentially help identify areas where the model is struggling or where it may not generalize well to other datasets.

2.2 Error Analysis

I started by conducting a manual error assessment. I took a random sample of 50 observations that the model potentially got incorrect (low similarity between the answer generated by the model and one of the possible answer options labeled in the dataset). I categorized the observations into one of five groups, as outlined in Table 2.

Category	Count
Not an error	16
Several dates are referenced in the passage and the model picks up the wrong one or several entities/events are referenced and the model picks the wrong one. Entity makes several claims and the model picks the wrong one.	12
Model identified a close answer but not the critical element of the passage.	9
Complex passage: The correct answer implies discarding some options due to context or identifying several elements from the passage.	9
Model selected other information similar in format but not relevant to the answer.	4

Table 2: Analysis of error categories and their respective counts.

The not an error category is the largest and includes responses where the similarity may be low due to the model paraphrasing the answer or providing more or less context, but that I identified as an acceptable answer (for example, the answer was once a year and the model's response was annually). The remaining categories indicate that the model may tend to struggle in questions that require discerning between actions taken by one of several possible actors, dates in passages including multiple events, or where the answer requires discarding among several potential options.

To understand further if there were some clear characteristics of either the question or the passage that correlated with lower model performance, I adjusted a logistic regression to the outputs of the model. In this case, I classified a response as correct based on an exact match with one of the potential answers. For the covariates I utilized indicators of whether the question included one of the following key words: what, where, when, why, who, which or none (other). Additionally, I included the length of the context, a dummy of whether the number of distinct entities in the context was greater than eleven, and a dummy of whether the number of distinct dates in the context was greater than three and some interaction terms. Results are presented in Table 3.

Variable	**Coefficient**	**Std. Err.**	**z**	**P**	**[0.025**	**0.975]**
const	1.2390	0.085	14.501	0.000	1.072	1.406
length	-0.0010	0.000	-2.204	0.028	-0.002	-0.000
entity_dummy	-0.0296	0.097	-0.304	0.761	-0.220	0.161
date_dummy	0.1117	0.107	1.040	0.298	-0.099	0.322
broad_category_what	-0.1981	0.077	-2.584	0.010	-0.348	-0.048
broad_category_when	0.9845	0.139	7.078	0.000	0.712	1.257
broad_category_where	-0.3548	0.119	-2.970	0.003	-0.589	-0.121
broad_category_which	0.0025	0.112	0.022	0.982	-0.217	0.222
broad_category_who	0.4056	0.099	4.085	0.000	0.211	0.600
broad_category_why	-1.2471	0.207	-6.011	0.000	-1.654	-0.840
entities_dates	-0.1183	0.137	-0.862	0.389	-0.388	0.151
entities_category_what	0.2911	0.120	2.428	0.015	0.056	0.526
dates_category_what	0.0455	0.128	0.356	0.722	-0.205	0.296
entities_category_why	0.3152	0.471	0.669	0.504	-0.609	1.239
dates_category_why	0.3535	0.411	0.861	0.389	-0.451	1.158

Table 3: Logit Regression Analysis

From this exercise, it can be inferred that the model tends to struggle in questions with a longer context, especially in questions containing the word why. In fact, the accuracy on questions why on the validation set is .51. The number of distinct dates and entities in the passage are not statistically significantly correlated with model accuracy. This could be because the complexity of the questions does not depend on the number of entities or dates, but on the order they appear in the passage, or the information the question is asking about. That is, the model may have learned some rules for how the information about an entity or a date should be present in the context (for example, information about an entity should come close to where it is named in a passage), and the number of entities or dates is not a good reflection of instances where these sort of rules do not generalize adequately. On the other hand, questions on why tend to require a careful parsing of the context to identify the correct answer. Indeed, I manually reviewed 50 instances of questions why and identified all the errors from the model were due to it not identifying the relevant elements of the passage (i.e. category complex passage in the categorization from Table 2). This analysis indicates that the model will likely not generalize well to the addition of irrelevant passages that break the structure of the context.

2.3 Contrast Sets and Adversarial SQuAD

To test this hypothesis I constructed ten contrast sets by conducting minor modifications to the structure or adding irrelevant information to examples that the model had scored correctly on the test set. Out of the ten examples, the model answered eight correctly. This further supports the theory that the model might struggle when the context structure is not in line with that seen during training.

To test this further I evaluated the model on the validation set of Adversarial SQuAD (Jia and Liang, 2017). This dataset was created by adding distracting sentences to SQuAD. As expected, the model performance was much lower with an F1 of 20.68. The accuracy on why questions is only .06.

3 Methods to improve performance on why questions

To improve the generalizability of the model, especially on why questions, I tried three approaches. The first was to use data augmentation. For this, I leveraged the TextAttack library (Morris et al, 2020) on the why questions, to augment the training data. Questions were augmented by either replacing some of the words with synonyms or by using back-translation. This allowed me to produce an additional 3,747 why questions (two for each dataset question by replacing for a synonym and one by back-translation). Then, I fine-tuned the model on the augmented data and evaluated its performance on both the SQuAD and Adversarial SQuAD datasets.

The second approach was equal to the first, but additionally, I perturbed the context with the CLARE augmenter from TextAttack. Finally, instead of augmenting the why questions, I fine-tuned the model on the entire training data from

Quoref (Dasigi et al, 2019). Unlike SQuAD, this dataset features longer questions and is designed to include questions that involve coreference resolution. In that sense, fine-tuning on the Quoref dataset may help with the limitations identified manually in questions involving multiple entities or on why questions, which typically require careful context parsing.

Statistics on the questions available in SQuAD (Validation set), Adversarial SQuAD and Quoref are presented in table 4.

Dataset	**Total Entries**	**Why**	**Mean Distinct Entities**	**Mean Distinct Dates**
SQuAD	10,570	150	7.13	1.87
Adversarial SQuAD	30,000	795	7.29	1.92
Quoref	19,399	17	18.18	5.52

Table 4: Summary of Dataset Statistics

While Quoref has a more limited number of questions Why, the context typically involves many more entities and dates.

3.1 Results

The results of the experiments are presented in Table 5. The training column indicates whether the model was trained only on the relevant dataset or if it was first trained on SQuAD and later fine-tuned on the dataset.

Training	**Model**	**SQuAD**	**Adversarial SQuAD**	**Why on SQuAD**
Full	SQuAD	86.22	20.67	0.51
Fine-Tune	Augmented Questions	73.89	17.76	0.50
Fine-Tune	Augmented Context	77.55	18.81	0.43
Fine-Tune	Quoref	73.17	21.30	0.40
Full	Quoref	54.73	15.29	0.11
Full	Augmented Context	13.97	7.23	0.31

Table 5: Performance of models across datasets.

From the results, it can be seen that the original model trained on SQuAD archives the highest performance on the SQuAD dataset. This is somewhat expected as it is logical that the model trained on SQuAD should perform best on it. What is more relevant is that none of the models managed to outperform the original SQuAD model on the why questions. Other important results are that the model fine-tuned on Quoref performed slightly better than all the other models on

Adversarial SQuAD. This indicates that adding questions that involve coreference resolution can help the model become slightly more robust to adversarial contexts, but the performance increase is not large enough to indicate that the model will generalize adequately to adversarial context. Finally, the model fine-tuned on augmented why questions had the best performance on why questions relative to its overall performance on the dataset. This can indicate that fine-tuning the model on questions augmented by simple modifications (back-translation and substitution by synonyms) can potentially help it obtain better performance on generalized contexts. Unfortunately, in this case, it was not possible to fully establish this conclusion as the Adversarial SQuAD dataset contains a limited number of why questions and given the low performance of all the models on Adversarial SQuAD, the accuracy on why questions was not significantly different in any of the models.

4 Conclusions

In this project, I identified a specific limitation of the RoBERTa model for answering why questions on the SQuAD dataset. I tried a few potential solutions such as fine-tuning or training the model on a dataset involving coreferencing (Quoref) and augmenting the SQuAD why questions and context. The results from the experiment show that while agumenting with Quoref can help the model perform better in adversarial contexts, the performance gains were too small to indicate that this is a plausible solution. Additionally, while augmenting why questions may help the model perform relatively better, the overall performance and generalizability to adversarial contexts did not improve. Augmenting the context also did not result in better performance. When reviewing the augmented questions and context manually, it seemed to me that the quality of the questions was not high. Some of the changes did not make sense in the context (for example, a synonym may be used out of context). I can conclude that automatic augmentation of datasets can be quite complex in a QA context and a better potential solution may be found in augmenting questions manually (although this would be time-consuming and more resources would be needed to conduct that exercise).

References

[Dasigi et al., 2019] Dasigi, P.; Liu, N. F.; Marasovic, A.; Smith, N. A.; and Gardner, M. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. 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 5925–5932, Hong Kong, China. Association for Computational Linguistics.

[Jia and Liang, 2017] Jia, R.; and Liang, P. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Morris et al., 2020] Morris, J. X.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP.

[Rajpurkar et al., 2016] Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.