Using Non-Domain Contrastive Data to Improve Medical Question Answering

https://github.com/mloet/Contrast-PubMed

Abstract

My goal was to investigate data augmentation strategies to improve the performance of question answering models on yes/no question answering in the medical domain. To achieve this, I focused on improving model performance on the PubMedQA dataset, a yes/no/maybe expert-labelled dataset using context passages from PubMed, a large dataset consisting of biomedical literature. I focused on improving the performance of a baseline ELECTRA-small model initially trained on the SQuAD 2.0 dataset, and finetuned on BoolQ for yes/no question answering. By using perturbed BoolQ questions to augment the training data, I was able to improve exact match score on PubMedQA by 7.9 points and F1 score by 13.5 points.

1 Introduction

In recent years, we have seen vast improvements in the performance of large language models in a variety of tasks and domains. As such, there have emerged opportunities to develop and train models to assist professionals in critical areas that suffer from a scarcity of expert analysis and reasoning. One such area is that of healthcare, which represents one of the greatest necessities for people, and yet plagued by staffing shortages in many areas due to the high levels of education and sometimes grueling work conditions and hours required for many different healthcare jobs. Fortunately, there have been several promising areas in which machine learning tools can be developed to assist in the daily lives of doctors and nurses, one of which being the use of language models in question answering to alleviate the necessity for some of the memorization of diseases, disorders, treatments, and other such information that doctors need to be aware of. As such, I wanted to explore strategies that could be used to improve question answering model construction and training specifically for the medical domain.

There are a number of challenges and roadblocks that face the development of model architecture and training pipelines in biomedical question answering, which I became aware of through other work and my own exploration of problematic training examples. Medical question/context pairings used in training tend to be lengthy and full of salient medical terms. While it is important for both questions and context to be specific and accurate when considering health-related inquiries, the length and abundance of terms can prove problematic for the model's semantic understanding of the question. I wanted to explore if we can achieve greater performance on biomedical question answering tasks by training the question answering model with not only domain specific question/answer examples, but also contrastive examples that help capture the semantic construction of the prompt.

2 Similar Work

2.1 Learning with Instance Bundles for Reading Comprehension

Machine learning models are typically trained with the assumption that the training instances sampled from some data distribution are independent, leading to lower performance on minimally different questions. Dua et al. introduced a method of training on *instance bundles*, or sets of closely contrasting examples. They were able to achieve 9% absolute gains on HotpotQA when using instance bundles from ROPES.

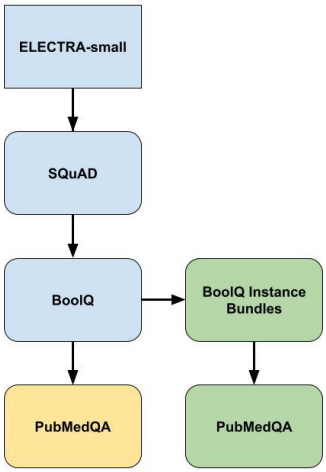
2.2 Evaluating Models' Local Decision Boundaries via Contrast Sets

Progress in NLP tasks is measured in large part by standard benchmark datasets; however, these

datasets often have systematic gaps that allow simple decision rules to perform well on test data. Gardner et al. proposes that constructing minimally contrastive examples can characterize the correct decision boundary, leading models to perform significantly better on small input perturbations.

3 Methodology

3.1 Experiment Overview



I propose that pre-training a model on contrastive bundles of perturbed BoolQ questions can help dissuade learning of rules unrelated to the semantic meaning of questions and context and improve performance on the yes/no classification questions of PubMedQA. Above is a simplified figure illustrating the experiment's training pipeline.

3.2 Data Sources

I used multiple different datasets to train my models and explore the effects of contrastive data augmentation on biomedical question answering, introduced below.

The **S**tanford **Qu**estion **A**nswering **D**ataset (**SQuAD**) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable. This was used as the initial training base for my models.

BoolQ is a question answering dataset for yes/no questions containing 15,942 examples. These questions are *naturally occurring* – they are generated in unprompted and unconstrained settings. This was used to finetune the model head for yes/no classification, as well as being the base for the contrastive dataset I used.

PubMedQA is a biomedical question answering dataset of expertly labeled question/answer pairs using context from the PubMed dataset. The objective of which is to answer research questions with yes/no/maybe using the corresponding abstracts. I sought to improve model performance on this dataset.

3.3 Experiment Setup

For the experiment, I chose to use an ELECTRA-small model. ELECTRA can achieve state of the art results on question answering datasets like SQuAD 2.0, but requires relatively few computational resources when compared to other SOTA models. As such, I decided to pre-train the model on SQuAD 2.0, which achieved an exact match score of 78.38 and an F1 score of 86.08.

While SQuAD provides a base for question answering, PubMedQA contains only yes/no/maybe answers. As such, it is necessary to add a classification head to the model and fine tune it on a relevant dataset. In my experiment I chose to focus on yes/no question answering, due to the relative scarcity of labelled data that includes all three classes compared to True/False or yes/no data. In this case, I chose to fine-tune the model on BoolQ, a True/False dataset with 15,942 examples.

3.4 Initial Analysis

Now, the model can infer on the expert labelled examples from PubMedQA. While the model has not yet been trained on domain specific data, we can still expect to see accuracy better than random due to the examples from SQuAD and BoolQ. When evaluating on the test set, we get the following metrics:

Before Training on PubMedQA

	Total	Yes	No
EM	0.603		
F1	0.489	0.731	0.247
Precision	0.537	0.631	0.443
Recall	0.560	0.868	0.176

As we can see, without domain specific training, our model does not perform notably well on the PubMedQA dataset. However, we can infer certain insights based on the metrics and the questions the model got incorrect.

First, the model performs notably better on questions that have 'yes' as their answer. This could be for a number of reasons. First of all, there are slightly more 'yes' answer questions in the dataset (62.31%). While this could contribute to the discrepancy, considering the size of the dataset it is unlikely to be the sole cause.

We can also see that the model performs notably poorly on 'no' recall, meaning that we are only correctly identifying 17.6% of questions that have 'no' as their answers. When looking at some of the false negatives we can see that many of them contain positive modifiers like "Does increased...", "Is intensive...", or "Does short-term...". An example of this can be seen below:

Question: Is there an increase in the incidence of gbs carrier rates among pregnant women in northern Israel?
Label: no

This sort of language is pervasive throughout the dataset and could be a large influence on why the model has a bias towards predicting 'yes'.

3.5 Contrastive Training

Before I trained the model on PubMed to improve its domain specific understanding, I wanted to train a separate model on a contrastive dataset to try to improve the model's semantic understanding of the question. To achieve this, I used a perturbed BoolQ dataset constructed by the Allen AI Institute. This dataset contains slightly altered versions of select BoolQ questions, examples of which are demonstrated below:

Original Question: Is pain experienced in a missing body part or paralyzed area?
Label: yes

Perturbed Question: Is pain experienced in organs that are not physically part of the body?
Label: yes

Perturbed Question: Is pain experienced only with spinal cord injury?
Label: no

As can be seen in the examples above, the dataset contains 410 perturbed questions related to various BoolQ questions. These questions are perturbed in a number of ways while staying similar to the original scope, such as rewording the question or restructuring a question to its negation. Ideally, this will encourage the model to learn semantic relationships in the inputted questions as well as domain specific information, thus improving its ability to attend to the appropriate span of text.

3.6 Domain Specific Training

Finally, I trained my two models, one finetuned on BoolQ only and the other on both BoolQ and the contrast set, on the PubMedQA dataset. I used the expert labeled subsection of the data, and only considered questions that had an answer of 'yes' or 'no'. This resulted in 890 distinct examples, such as the following:

Question: Do mitochondria play a role in remodelling lace plant leaves during programmed cell death?
Label: yes

4 Results

4.1 Model Performance Metrics

Now that we have trained on PubMedQA, we can see the two model's evaluation metrics below:

Control (not trained on contrast set)

	Total	Yes	No
EM	0.713		
F1	0.626	0.807	0.444
Precision	0.768	0.693	0.843
Recall	0.634	0.966	0.302

Contrast (trained on BoolQ contrast set)

	Total	Yes	No
EM	0.792		
F1	0.761	0.847	0.674
Precision	0.806	0.778	0.834
Recall	0.748	0.931	0.567

Difference (contrast - control)

	Total	Yes	No
EM	+ 0.079		
F1	+ 0.135	+ 0.040	+ 0.230
Precision	+ 0.038	+ 0.085	- 0.009
Recall	+ 0.114	- 0.035	+ 0.265

4.2 Analysis

As we can see, pre-training the model on the BoolQ contrast set has caused a notable improvement in nearly every metric when training on the PubMedQA dataset. Aside from large increases in total EM and F1 scores across the dataset, we see especially large increases in the F1 and recall scores for 'no' when compared to the model that wasn't trained on the contrast set. This indicates that pre-training on the contrast set has a large effect on decreasing the number of false negatives. Considering that the contrast dataset has a nearly 50-50 split of yes/no labels, it can be reasonably inferred that this decrease in false negatives for no is not due to the model learning the distribution of class labels. Instead, it seems that the contrast set has encouraged the learning of certain semantic relationships in the question and/or context which greatly assists the accuracy of the question answering on PubMedQA.

To further infer on the effect the contrast set has on the model, we can analyze the differences in which examples each model correctly classified. When analyzing which examples were correctly classified by the contrast model but misclassified by the control model, we can see that, as we hypothesized above, the contrast model performed much better on examples with modifiers that suggest a positive response, such as the following:

> **Question**: Does successful completion of the Perinatal Education Programme result in improved obstetric practice?
> **Label**: no

Another area that stood out from the examples was the handling of certain scientific terms like "significant difference" and "statistically significant". The contrast model performed better at identifying portions of context like "there was no correlation found" as negative, as is shown in the following example which the contrast model got correct while the control did not.

> **Question**: Is the clinically positive axilla in breast cancer really a contraindication to sentinel lymph node biopsy?
> **Context**: Clinically positive axillary nodes are widely considered a contraindication to sentinel lymph node (SLN) biopsy in breast cancer, yet no data support this mandate.
> **Label**: no

One final area that I noticed the contrast model performing better at was questions relating to study results, especially when portions of the study exhibited positive results but the portion in question did not. This can be illustrated in the example below.

> **Question**: "Does cognitive function predict frequency compressed speech recognition in listeners with normal hearing and normal cognition?"
> **Context**: "Speech-in-noise recognition was measured using Institute of Electrical and Electronic Engineers sentences presented over earphones at 65 dB SPL and a range of signal-to-noise ratios… …There was a statistically significant reduction in mean speech recognition from around 80% when unprocessed to 40% for 2:1 compression and 30% for 3:1 compression. There was a statistically significant relationship between speech recognition and cognition for the unprocessed condition but not for the frequency-compressed conditions."
> **Label**: no

Overall, when considering the evaluation metrics and looking at the examples that the two models classified, it seems that pre-training the model on the contrast dataset allowed it to reason better in a variety of different ways and made a significant difference in the accuracy of the model.

5 Conclusion

Over the past few years, there have been significant advances in the development of language models by leveraging the scale of training data to improve model performance. However, there are certain critical areas that could greatly benefit from improved model performance that have limited availability of high-quality domain-specific data, such as the healthcare field. Fortunately, there are a number of strategies that can be used to leverage non-domain specific data to improve technical language related tasks.

I set out to see if the use of an out-of-domain contrastive BoolQ dataset could improve model performance on the medical question answering dataset PubMedQA. To train the models, I made use of the 150,000 example SQuAD dataset, the 15,942 example BoolQ dataset, and an 890 example yes/no subset of the PubMedQA dataset.

I found that the use of a 410 example contrast set for BoolQ questions can lead to a substantial increase in accuracy on the PubMedQA dataset, increasing exact match score by 7.9 points and F1 score by 13.5 points. Despite being a fraction of the examples of the other datasets used in training, the addition of the contrast set in pre-training the model appeared to allow the model to reason better in a number of different ways and on a number of different examples, including improved semantic understanding of modifiers in the question, improved understanding of scientific descriptors in context, and improved extraction of relevant portions of study results. This helps to demonstrate that there are several ways to improve technical, domain focused language models without the use of in-domain examples.

References

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Aguera y Arcas, B., ... Natarajan, V. (2023). *Towards Expert-Level Medical Question Answering with Large Language Models*. arXiv preprint arXiv:2305.09617.

Zhang, M., Dou, S., Wang, Z., & Wu, Y. (2023). *Focus-Driven Contrastive Learning for Medical Question Summarization*. arXiv preprint arXiv:2209.00484.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). *PubMedQA: A Dataset for Biomedical Research Question Answering*. arXiv preprint arXiv:1909.06146.

Dua, D., Dasigi, P., Singh, S., & Gardner, M. (2021). Learning with Instance Bundles for Reading Comprehension. arXiv preprint arXiv:2104.08735. https://arxiv.org/abs/2104.08735

Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., Gupta, N., Hajishirzi, H., Ilharco, G., Khashabi, D., Lin, K., Liu, J., Liu, N. F., Mulcaire, P., Ning, Q., ... Zhou, B. (2020). Evaluating Models' Local Decision Boundaries via Contrast Sets. arXiv preprint arXiv:2004.02709. https://arxiv.org/abs/2004.02709

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., & Toutanova, K. (2019). BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. arXiv preprint arXiv:1905.10044. https://arxiv.org/abs/1905.10044