CosmosQA Dataset: A Multiway Attention based implementation

Jaykrushna Avatar
1217096660
Arizona State University
ajaykrus@asu.edu

Shah Samkit
1218404161
Arizona State University
skshah15@asu.edu

Vakil Yash
1217130746
Arizona State University
yvakil@asu.edu

Pushparajsinh Zala
1217568222
Arizona State University
pzala@asu.edu

Abstract

Natural Language Question Answering is one of the most challenging and open ended problems of AI in the current days. There has been unprecedented advancements in the creation of pre-trained models which exhibit a very remarkable performance on many datasets for reading comprehension. Having said this, we are still a long way from creating a model which can achieve human level performance in any given Question Answering dataset. The main reason behind this is that answering questions from context more often than not need some prior knowledge of the domain to which the context belongs to and the ability to infer answers that does not lie directly in the sentences of the given context but is present between the lines. In this paper, we are going to apply one such pre-trained model to a given reading comprehension dataset and analyze the areas where our model fails and how it can be improved by feeding some prior knowledge about the context.

1 Introduction

As individuals have gotten progressively versatile, the extensive availability and simplicity of internet usage only adds fuel to the ever growing user base and information. The size, scope, and type of information on the internet continuously grows as users engage with it on a daily basis. Subsequently, a pressing problem would be to have a framework that is able to cater to it's users' questions. The primary objective of search engines like Google, Bing, and Yahoo is to effectively answer the users' queries. For some predefined scenarios, they do return a direct answer to their question; However, for most of the time they end up returning a list of ranked pages for the user to manually read on and find their answer. QA (Question-Answering) systems not only help cut down the time for finding a suitable answer, but they can conceivably help computerize errands in different spaces. Today, the industry has seen a rise in the application of QA systems due to its ability to automate tasks. They can be seen as programs embedded in systems for but not limited to documentation, tutoring, management, and human-computer interaction.

Natural Language Question Answering involves building a system that is able to understand to given paragraph and answer the necessary questions related to the said passage. Due to their ability to objectify how well a system understands a sentence, they are also used as a measuring factor for NLU (Natural Language Understanding) systems as stated in (McCarthy), (Lehnert, 1977) and (Winograd, 1972). Be that as it may, parsing a sentence with human level comprehension is complex and involves many challenging tasks such as Recognition, Parsing, Linking, Information Extraction, and Synthesis. Current QA systems are held back mainly because of two reasons.

- To be able to answer a query precisely and completely it is fundamental for the system to have knowledge on every concept alluded by the question. It implies that there are times that the system might need some common sense or domain specific knowledge to be able to understand the passage.
- Language itself is complex with ambiguity at different levels such as Lexical, Syntactic, Semantic, Discourse, and Pragmatic. In this way, making it necessary for computers to be able to detect and resolve such types of ambiguities proficiently.

2 Dataset/Task Description

Reading comprehension based Question answering problems requires more understanding of context as well as choices. (Huang et al., 2019) have

	Train	Dev	Test	All
#Question (Paragraphs)	25888(13715)	3000(2460)	7000(5711)	35588(21866)
Ave/Max. #Tokens/Paragraphs	69.4/152	72.6/150	73.1/149	70.3/152
Ave/Max. #Tokens/Question	10.3/34	11.2/28	11.2/29	10.6/34
Ave/Max. #Tokens/Correct Answer	8.0/40	9.7/41	9.7/36	8.5/41
Ave/Max. #Tokens/Incorrect Answer	7.6/40	9.1/38	9.1/36	8.0/40
Percentage of Unanswerable Questions	5.9	8.7	8.4	6.7

Table 1: Statistics of Train,Dev and Test of CosmosQA adapted from (Huang et al., 2019)

posited a CommonsenseQA Dataset which includes 35588 total question in train, validation and testing files. The data itself is unique because it contains multiple choice questions on commonsense reasoning, where most of the answers are not present in the context which makes the tasks more arduous.

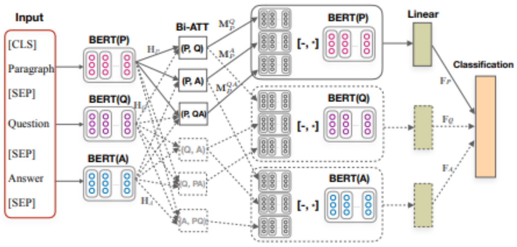

Figure 1: BERT Multiway Attention Model Architecture adapted from (Huang et al., 2019)

In Table 1, Ratio of Tokens per paragraph is really high while ratio of Tokens to questions and choices are very low that means this requires commonsense to answer the questions instead of context based inference. To create unanswerable question creation, authors have taken 70% of good questions and made random samples via fine tuning with BERT on next sentence prediction. (Huang et al., 2019) have included a significant amount of frequent trigram prefixes in their COSMOSQA dataset where the questions are based on the categories such as “why”, “what may happen”, “what will happen” etc. which seem to be absent from other renowned datasets making it more challenging to compare it to other datasets such as SQUAD created by Purkar et al. (Rajpurkar et al., 2016).It also contains commonsense inference based on Pre/Post condtions, Motivations, Reactions, Temporal Events, Situational Facts and CounterFactuals which contains 93.8% total questions.

Figure 1 demonstrates the model that we are implementing in this paper. It has been adapted from (Huang et al., 2019) and the main idea behind selecting this model as base was that it has proven to be effective in the Question Answering Dataset.

3 Methods/Implementation

The problem statement was self-explanatory. However, there were many approaches to realize the solution for the said issue. Before we began with creating our own model we figured it would be best for the team to get an insight into the present frameworks and solutions, which we could then build upon. The information gathering stage started by combing through the various preexistence solutions presented in the CosmosQA leaderboard.

After agreeing upon our Base model i.e the BERT Multiway Attention Model, we delved into getting a clear understanding of the BERT model itself. Moving forward, we chipped away taking a gander at the dataset samples and understanding their decent variety. It was clearly noticeable that our local system was not ideal to train such a large model. This was when we decided to work on a lower scale model i.e. DistilBERT proposed by (Sanh et al., 2019). The logic behind this is that DistilBERT has the same configuration as that of the BERT base model only that it has lesser trainable parameters. This entails that the accuracy of the DistilBERT would be lower to that of BERT base but it would also reduce our computation time and load. Moreover, if we build upon this DistilBERT model and are successful in increasing its accuracy, it would directly imply that the same method would increase BERT base’s accuracy as it has only architectural difference.

After that we split into groups of two implementing the Bert model on Google Colab and a DistilBERT for the local system, because at the end we would need to check our additions on the main Bert model itself. We implemented the DistilBERT model for this task. After that we were also able to successfully implement the BERT base model using GPU. This concluded the end of our progress until the first phase after which we dived into moving forward with a better base model and doing experiments on it to improve the accuracy.

4 Error Analysis

4.1 Analysis

We hand picked a few questions to analyze the behavior of the models and we were able to discern some traits of failure of the model. Authors of COSMOS QA(Huang et al., 2019) paper have set a few categories for the error analysis section that we have used to classify our errors. Both models have shown significant dissimilarities in learning and it can be inferred that with less parameters and same data, we are not able to achieve considerably good results in DistilBERT(Sanh et al., 2019) as we did in BERT-large and BERT-base model which performed remarkably better. We achieved 62.2% dev accuracy with fine tuning BERT-base on various parameters. It is evident from the results that BERT-base is not catching word or sequences which contains negation or negation keywords like *not, but and negative syntax structure with long dependencies in sequences*.

From our validation samples, we have taken 200 selected questions to analyze for errors and some questions lie in categories of Complex Context Understanding which can be exemplified using the below statement. For example for given paragraph: *"I watched the first McCain / Obama debate last night. It was full of moments I had to pause the DVR because I had to discuss what they were saying with my husband. I learned a lot about the Iraq war and Afghanistan , and I saw both McCain and Obama make some going points, and I saw them both make some blunders."* and question is: *How would this person be classified?*. Here, the answer should be *Independent* from the choices but it chose *None of the above*. Here it's hard to infer that person himself has liberal or independent as a part of classification.

Some questions can be classified as "Inconsistent Human commonsense" where there is inconsistency with human commonsense. For example for given paragraph: *Oh no ! I went to Starbucks tonight to meet with a friend. I went early, you know any way I can get out of the house without the children and spend some quiet time on my own, works for me!* and question is: *What may be your reason for going to Starbucks?* and model predicted *I invited someone there* , while real answer is: *I wanted a break from my family*. Here where inconsistency happened where from context it can be inferred that she want some alone time as human commonsense while model predicts something else. Another category, where there are multiple inferences are given in same question and model confused to choose between two and some questions are fell into "Unanswerable" category where even human cannot infer the answer directly.

5 Experimental Approaches

5.1 RoBERTa Multiway Attention Model

After installing Apex library, we were successfully able to implement the RoBERTa-large model which was a significant improvement compared to our previous base model whose results we will be discussing in the later parts of this paper. For improving the accuracy of our new base model, we explored various possibilities ranging from tweaking the architecture of the model itself to adding new pre-processing techniques to the data.

5.2 Generative Model

Firstly, we explored the idea of creating a Generator model approach because after error analysis we realized that there were some particular areas where the model was not able to catch the inferred knowledge in the context and failed. Our idea was to generate these "close call" adversarial samples where the possibility of 2 answers was really high thus confusing the model. But, the idea wasn't feasible given the time constraint and the computation restrictions because there was an issue of creating the other three options which were adversarial as well.

5.3 Text Fooler

This is a very brilliant method to generate malicious adversarial examples to fool Natural Language Based systems. It attacked the systems by precisely removing and substituting the most relevant information according to the model. This way, after training on these adversarial examples, the previously misdirected model would now focus on the actual relevant information and not the one it thinks to be most important. However, the main issue that we faced in this method was incorporating the adversarial examples in our training data. This was due to the fact that our training data not only consisted of the answer options but also the question and context. This made the task of creating adversarial samples more taxing and complicated

Model	Batch Size	Learning Rate	# Train Epochs	Dev Accuracy
DistilBERT	64	3e-5	5	50
BERT-base-SocialIQA	64	2e-5	10	58.8
BERT-base	64	1e-5	10	62.2
RoBerta-large	36	3e-5	3	76.08
RoBerta-large	32	2e-5	3	78.99
RoBerta-large + Text Similarity	32	3e-5	3	79.22

Table 2: The results for different models which we implemented, can be seen from Table 2. Here, we have used 220 as the maximum sequence length for all the models.

as we would need to take care that the model does not get distracted by the adversarial samples because we would still need a right answer to the question. So, generating a new right answer and a new adversarial option from the previous right answer seemed to be a redundant task.

5.4 Knowledge Infusion

Secondly, we explored the possibility of adding more knowledge in the model before training on our dataset hoping that it would lead to the model catching relationships more efficiently. We trained BERT-Large on SocialIQA dataset because we were facing some errors in training the RoBERTa-Large base model on the dataset. We were unable to solve the errors and start the initialization epoch so we downgraded to BERT-Large model for this task. It contained question answering task with some inference so knowledge infusion work this way and then trained same BERT model on CosmosQA which gives us almost same results compare to previous. This method did not prove to be very effective as well. We were only able to train the model for a few epochs due to computational limitations. Our inference for this method did not span out as we assumed, the reason could be attributed to the fact that the pre-trained model is not able to learn any extra knowledge from the given SocialIQA dataset. Another issue can be that the knowledge or relationships that the model is unable to catch in the SocialIQA dataset is the same kind of questions that BERT is getting wrong.

5.5 Text Similarity & Text Summarization

Considering that BERT and RoBerta have restrictions on the context size it takes as input, we believe that inputting only relevant context data into the model would help it to focus on the main context and not get wavered by irrelevant information.

One such process that would help with minimizing of context data is to use text similarity between the question and every sentence of the context. A threshold value then decided which sentences to include and which to exclude. After careful consideration we decided upon a threshold value of 70%. Using this method we got a slight better result compare with Roberta Multiway attention model which can be seem in Table 2.

Another method would be using text summarization to reduce the size of the context paragraph. However, it is not practically possible to apply a text summarization on COSMOSQA context as the size of context paragraph is extremely small, which in itself is a summarized version. It would result into a summarized version which would contain irrelevant information. Thus, we were not able to successfully execute this idea in our model.

6 Future Extensions

For the future work, adding new knowledge to the base model in some informative format can aid in improving the accuracy. From exhaustive analysis of the samples which the model predicted wrong, we were able to narrow down the specific areas and types of questions where the model failed. The idea is to not only feed direct background knowledge but to utilize inherent relationships of the text data and feed that information as well to the model. There are various Graph Based approaches which have proven to be effective in many studies similar to this paper which can be utilized in this work as well.

During our error analysis of the predictions by RoBerta-large we were able to conclude that the model failed to properly infer answers for contexts containing injunctions, conjunctions and negations. One way to help overcome this problem is to use predefined linguistic rules to re phrase the context. An example of such rule would be to replace negations with their positive counter parts.

7 Conclusion

After discerning the shortcomings of our model, presented in the error analysis section, we have categorized the areas where we can improve the performance of the model. We brainstormed various ideas to increase the accuracy of our base model which are discussed in Section 5.

Our implementation of the query based text similarity proved successful among the other tried approaches. The assumption that the model might get confused due to irrelevant information gave us an edge to address the problem in a way that would reduce the context size by discarding least helpful information from the context. However, limited by the computation we achieved only a trickle increase in the accuracy. From this experimental approach of ours we achieved a slightly better accuracy than our base RoBERTa model.

References

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning.

Wendy Lehnert. 1977. The process of question answering.

John McCarthy. An example for natural language understanding and the ai problems it raises.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1 – 191.