Conversational Question Answering with BERT

Eshwar N Kumar
Prof. Rohini K Srihari
Computer Science & Engineering,
University at Buffalo, State
University of New York
eshwarna@buffalo.edu

ABSTRACT

Conversational Question Answering using machines requires understanding the context of the multi-turn dialogue. Using BERT (a language model), single-turn machine comprehension can be achieved. But achieving multi-turn question answering with BERT is a very tedious task because BERT has an upper limit on the number and length of input sequences. In this paper, we propose an effective way to utilize BERT for multi-turn conversational question answering. The method in this paper uses BERT to encode a paragraph and corresponding questions independently. Then, the model predicts an answer based on the paragraph representations encoded with BERT.

1. INTRODUCTION

Single-turn question answering is nothing but a question answering method containing only a single turn. However, conversational applications such as Google Assistant and Siri requires answering multi-turn questions in a dialogue in addition to single-turn questions.

In our day to day conversations, we ask a question to another person about a subject either to get knowledge or test knowledge. Depending on the first answer from the answerer, we ask him/her another question and their second answer builds on top of what they answered for the first question. Today's virtual assistants lack the ability to build a common ground on a particular topic when multiple conversational type of questions are asked to the them. To solve this problem, we propose a simple way to develop conversational question answering chatbot using BERT.

CoQA(Conversational Question Answering) is a chatbot that understands a text passage and provides natural form answers to questions asked to it in conversational format.

One of the main goals of the chatbot is that it has to understand the context of the multi-turn dialogue that consists of the question and answer history.

Learning machine comprehension models require a lot of question answering data. Therefore, we use pre-training language models like BERT which was developed by google and has been trained on a large-scale unlabeled corpus for improving our model accuracy. BERT is a very powerful language representation model used for producing word embeddings. It is an advanced version of word2vec. The main difference is that BERT takes the context of the word into consideration but word2vec doesn't take it into consideration. BERT can obtain language representation by unsupervised pre-training with a huge data corpus. And by supervised fine tuning of BERT, excellent results can be achieved for various NLP tasks like question answering, sentence tagging, sentence pair classification, etc.

When making use of BERT for conversational question answering, we use the question and the corresponding passage as input and fine-tune the pre-trained BERT model to extract an answer from the paragraph.

2. PROBLEM STATEMENT

A passage is given as input to our chatbot. The chatbot must analyze the passage and provide answers to questions asked to it in the form of a conversation. It must also provide the text string in the passage from which the text string can be inferred as a rationale.

In addition, there are three main objectives to be achieved by this chatbot. First, each question in the paragraph after the first must be dependent on the conversation history and the bot must be able to answer the question of such kind. The conversation will have two annotators in which the chatbot will act as a second annotator/answerer. Second, the bot must ensure the presence of naturalness of the conversation with coreferences. The answer must be free form text which is a modified version of the text span obtained from the passage(which is quoted as a proof for the answer). The conversation system must work across

domains like the field of literature, medicine, science, news, etc.

3. RELATED AND EXISTING WORK

Document Reader Question Answering(DrQA) uses bigram hashing and tf-idf to efficiently return subset of articles from the Wikipedia data store. Then it uses RNN to detect answer spans. However, it is applicable for Wikipedia articles only and there is no coreference and natural form in answers.

A simple but effective way to implement multi-turn context with BERT for conversational machine comprehension performs contextual encoding of previous questions, previous answers and current question and then applies softmax to find the start and end index of the answer span. However, this model lacks coreference and natural form in answers.

Standford Question Answering Dataset(SQuaD) have all the three necessary objectives of our chatbot except that free-form nature and coreferencing are absent in the output.

GraphFlow: Exploiting Conversation Flow with Graph Neural Networks for conversational machine comprehension makes use of graphical neural network mechanism to process the paragraph over the traditional Integration Flow(IF) mechanism. Handles the case of abstractive answers in the passage.

4. PROPOSED MODEL

To develop the Conversational Question Answering Chatbot, it is very much necessary to consider the question history in addition to the current question. The proposed model consists of three steps.

4.1 Contextual Encoding

We use BERT to encode the relationship between the paragraph, current question, previous questions, previous answers.

The paragraph is fed as input to BERT along with current question, previous questions, previous answers. BERT outputs the relation between the question and paragraph.

4.2 Answer Span Extraction

The output features of the previous step is concatenated and passed through a GRU(Gated Recurrent Unit) whose output is then passed through a softmax function to predict the answer types like Yes or No or Unknown.

4.3 Filter the span text and add naturalness

Split the span text into chunks starting from smallest size to largest size chunk. Apply POS(Part Of Speech) and NER(Named Entity Recognition) to these chunks and find out the exact answer. This can be done by extracting all the legal answers possible to the question in the span. Then use word2vec to find word embeddings. Finally apply cosine similarity to find the exact answer among all the legal possible answers.

To add naturalness, we can use tools like GPT2 which is an AI based text generator that can predict the next word in the sentence. We can also use it to provide answers in the form of sentences wherever needed.

The Fig 1 below shows the flow diagram of our proposed model. After predicting the answer type, we use POS, NER to find the exact answer and GPT2 to add naturalness to the conversation.

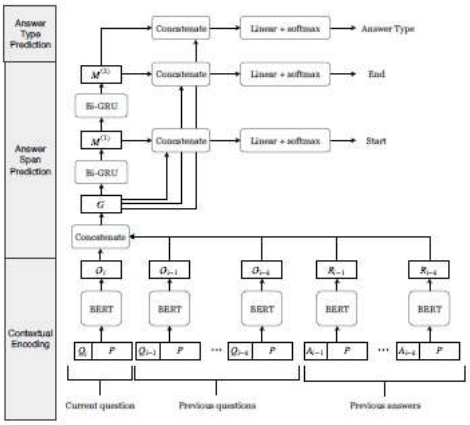

Fig 1. Proposed Model

5. DATASETS AND EVALUATION METRICS

Since the CoQA training and dev datasets contain the paragraphs from all the seven domains which are children stories, literature, middle and high school English exams, news articles, wikipedia articles, science articles and reddit articles. Hence, we use the official training and dev datasets to evaluate our model.

Similar to evaluation methodologies used by the submissions discussed in the literature survey, the evaluation metric used for this project will also be F1 score.

Data Set	In Domain					Out Domain		In Domain	Out Domain	Overall
	Children stories	Literature	mid high school	news	Wikipedia	Reddit	Science			
Dev Data Set	34.7	37.9	35.6	35.5	48.2	0.0	0.0	38.4	0.0	38.4
Train Data Set	36.6	38.2	37.4	36.9	48.7	32.4	31.2	39.6	31.8	35.7

Fig 2. The Model's F1 score on CoQA datasets

6. IMPLEMENTED BASELINE SYSTEM

Below are the tasks completed as a part of the implemented baseline system:

1. The Train and Dev datasets available in the official CoQA website have been preprocessed. This includes making the case of all the text in the datasets to lowercase, removing punctuation, extra white spaces, tokenizing the text, including offsets, etc.

2. We use the base version of BERT which is bert-base-uncased and feed our preprocessed text as input to the BERT model. This outputs the start and end index of the answer to the question.

3. The model is first trained with the preprocessed training data set. Then, our model is tested with the preprocessed dev dataset.

4. To evaluate our model, we find the F1 score by performing forward propagation and adjusting the weights during the back propagation. To do so, we perform L2 regularization using Adam optimizer during each epoch.

7. BASELINE RESULTS

Fig. 2 indicates the table of evaluation results for the implemented model using the datasets available in the official website of CoQA.

It shows the F1 score achieved by our implemented model under various categories like children stories, literature, mid-high school, news, wikipedia, reddit, science.

8. ERROR ANALYSIS

During the training phase of the model, we use Adam optimizer that performs L2 Regularization where the regularization is done during the gradient update step. We use a learning rate of 0.03 and we use 2 previous questions for learning about the previous context. We run this model for 10 epochs and compute the F1 score for each epoch.

Our model uses the paragraph representations independently conditioned with each question and each answer and this structure is suitable for the pre-trained BERT. This is the major reason why our model is able to capture the interaction between a paragraph and the corresponding dialogue history.

Model vs Humans The human performance during the training phase is 88.8 F1. On the other hand, our model performance is 38.4 F1. This indicates that our model has a lot of scope for improvements and upgrades in the future.

In the model developed currently, there are many issues and bugs. It is not able to predict the exact start and end index of the span in the passage for some questions which is the major issue. This drastically affects the performance of our model. Even if the number of questions to learn about the context is increased above 2, we don't see any major change in the F1 score. From the results table, we observe that the F1 score for the paragraphs from wikipedia is highest being 48.2 and the F1 score for the paragraphs from children stories is minimum being 34.7 for the currently implemented model. We also observe that the F1 score for out of domain questions is much smaller than the F1 score for in domain questions.

9. PROBABLE IMPROVEMENTS

Below are the drawbacks of the currently implemented baseline model in comparison with the proposed model and its goals:

1. The model does not possess naturalness in answers.

2. No rationale in output.

3. Absence of answer type prediction.

According to the current implementation, only the Step 1(4.1) of the proposed model has been completed.

During the upcoming days, we plan to complete the Step 2(4.2). During this phase of the project, the output of BERT is passed as input to GRU, and softmax functions to predict the answer types. We will also use the same to add rationale to the output of the chatbot.

After completing Step 2, we plan to complete Step 3(4.3) during which we fine-tune our model. During this step, we split the span text into chunks starting from smallest size to largest size chunk. Then apply POS(Part of speech) and NER(Named Entity Recognition). We will also use tools like GPT2 to add naturalness to the answer going forward.

10. CONCLUSION

In this paper, we propose an effective way to utilize BERT for conversational question answering based on the approach of fine-tuning. The model uses the paragraph, current question, previous question and answers and finds out the relationship between paragraph and each dialogue history independently. For future work, we will fine-tune our model and add naturalness to the conversation.

11. REFERENCES

[1] CoQA: A Conversational Question Answering Challenge. Siva Reddy, Danqi Chen, Christopher D. Manning (https://arxiv.org/abs/1808.07042)

[2] A simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension. Yasuhito Ohsungi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, Junji Tomita (https://arxiv.org/abs/1905.12848)

[3] Technical Report on Conversational Question Answering. Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, Yunfeng Liu.

[4] GraphFlow: Exploiting Conversation Flow with Graph Neural Networks for Conversational Machine Comprehension (https://arxiv.org/pdf/1908.00059.pdf)

[5] CoQA official webiste (https://stanfordnlp.github.io/coqa)

[6] Attention is all you need, Ashish Vaswani (https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)

[7] BERT: pre-training of deep bidirectional transformers for language understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. (https://arxiv.org/abs/1810.04805)

[8] The Illustrated Transformer, Jay Alammar (http://jalammar.github.io/illustrated-transformer/)