Final Project Report

**Project Name: Conversational Question-Answering using CoQA**

**Team Members: 1. Rutuja Saptarshi - 801100848**

**2. Sagar Talwar - 801100599**

* **Introduction**
* **We implemented a model that improved the ability of machines to participate in Question-Answer type conversations by analyzing the Conversation Question-Answer (CoQA) dataset.**
* **Project Domain – Natural Language Generation.**
* **Human beings communicate (verbally or textually) with each other and also with machines, in order to gain knowledge or to validate other participant’s knowledge about the subject.**
* **Generally, any conversation tends to be an incremental one, such that the answer to the current question by the questioner more or less depends and builds on the answer to the previous question by the answerer. This is the natural way of communication between humans.**
* **But in case of machines, the virtual assistants cannot keep up with this way of communication, and hence there is a wide window for improvement in the existing reading comprehension question-answer datasets and systems.**
* **The CoQA system has been developed with the following goals:**

**1. The nature of questions in natural human conversations, such as one word questions like “Which?”, “Who?” that rely on the lexical similarity and are difficult for existing systems to parse without the knowledge of the conversation history.**

**2. Ensuring the naturalness of the answers in the conversation by keeping the text spans for the answers open to multiple sentences in the rationale.**

**3. Keeping the system open to wide range of domains.**

* **We created a model in which a machine understands a reading comprehension passage and answered a series of questions based on that by maintaining the conversation history as its most important aspect in order to achieve the desired communication.**
* **This topic is of wide importance as there are many applications today that use the question-answer type conversations in order to communicate with the humans. To develop a model that exactly predicts what the answer for a particular question might be, given some context (a reading comprehension in our case) with a good accuracy is of great use today, since we are moving towards an era where machines have become the necessity for almost everything in human’s life. As conversation is the basis for gaining or giving knowledge, developing accurate models for machines to have conversations with humans has become very important in the computer science world.**

**Applications:**

* Chatbots
* Virtual Assistants
* Alexa/Siri/Cortana
* Information Retrieval based Question Answering
* Knowledge based Question and Answering (KBQA)

**Overview of Research Paper:**

* **Similar to CoQA dataset, we describe the existing related work under following criteria:**
* **Knowledge Source: The main source is are the reading comprehension passages based on which the questions are asked and answered. There are also other sources of data such as the databases which store the data in the form of tables, graphs and documents (Berant et al., 2013; Pasupat and Liang, 2015; Bordes et al., 2015; Saha et al., 2018; Talmor and Berant, 2018). But these sources require expertise which make it challenging to work on those without the predefined templates. Another useful source of data is in the form of multimedia such as images, videos and other graphics.**
* **Naturalness: Questions can be framed in various ways and not just in the form of interrogative sentences. The other ways are fill in the blanks type questions, paraphrasing artificial questions (Sahaetal., 2018; Talmor and Berant, 2018), using hand written grammar to question(Weston et al., 2016; Welbl et al., 2018), or just let humans ask questions. We try to achieve complete independence between the questions and the documents by collecting questions without seeing the source. But this is not possible as the questioner has access to the source while framing questions. Validation of the answers by the questioner is done on the spot in our setup which is an advantage.**
* **Conversational Modeling: The main focus of this model is question-answering. One solution is breaking down the complex questions into several simple questions. The CoQA answer set is free form text. There are some datasets which are regulatory texts requiring the answers in the form of “Yes” and “No” answers. For these QA systems, there are free form text answers in case the question needs clarification or cannot be answered. There are sequential sources of data in which multiple question-answer series is based on a particular content.**
* **Reasoning: In addition to our conversational datasets, there are other datasets requiring different types of reasoning such as algebraic reasoning (Clark, 2015), logical reasoning (Weston et al., 2016), common sense reasoning (Ostermannetal., 2018), and multi-fact reasoning (Welbl et al., 2018; Khashabi et al., 2018; Talmor and Berant, 2018)**

**- Input: Reading Comprehension Passage, Related Questions.**

**- Output: Answers to the Questions asked.**

**- The CoQA system takes into account the conversation history and thus the answer to question Qn depends on the answers to questions Q1, Q2 … Qn-1.**

**- Datasource: The parameters for collecting the dataset are as follows:**

**1. The collection interface is such that the questioners are not supposed to use the exact words from the passage in order to maintain the lexical diversity. If at all any word in the question appears from the passage, the question is paraphrased. On the other hand, the answerer is expected to use the exact words from the passage in the answer to maintain a limit on the possible answers.**

**2. The CoQA dataset has passages from seven domains which are children’s stories, literature, middle and high school English exams, news, Wikipedia, Reddit and science.**

**3. As the data is conversational, the questions influence answers which in turn influence the follow-up questions, the CoQA system collects three answer variations for each question in the development as well as test data.**

**- The following are the results from the CoQA dataset analysis:**

**1. CoQA has a richer variety of questions than SQuAD system because of its free form nature of answers. The distribution of CoQA is spread across multiple question types with prefixes such as “Did”, “Why”, “Was” and “Does”. Also, it is observed that CoQA is highly conversational due to the co-references as “he”, “she”, “it” appearing multiple times in the dataset. The questions are also short, many a times just one word.**

**2. For the linguistic phenomena, we observed that the majority of questions fall into the category of paraphrasing type of questions, which means that the questions were paraphrased from the rationale. A considerable amount of questions fall in the lexical match category where the exact words were used in the question as in rationale, while the others fell in no lexical cue category.**

**3. As the answers in CoQA dataset are free form in nature, they do not exactly overlap with the text from the passage. Majority of the answers are of the form “Yes” or “No”. The next majority are just one word edits with just insertion or deletion of a single word in the answer. The remaining are multiple word edits which are still challenging to evaluate using existing automatic metrics.**

**4. The conversations are in sync with the flow of data in the passage which establishes good coherent conversation.**

* **Background**
* Conversational Question-Answering models are complicated and challenging to implement because of the complex types of questions asked and the incremental nature of the conversations.
* A variety of factors must be taken into consideration while developing a model for this system which include whether to use short-term memory for storing previous answers, whether to use RNN or CNN for implementation, what F1 score is acceptable etc.
* We chose this topic for implementation given the tremendous use of conversations between machines and humans these days. Almost all web applications today use chatbots for virtual assistance and hence conversations have become very important in these applications.
* We developed a model for reading comprehension question-answering which can be improved further for normal day-to-day conversations and many other NLP applications.

**Other Models:**

* Conversational Modelling: There is an implementation of the conversational question-answering model where in the questioner is presented only with the title of the passage while the answerer can see the entire passage. This model puts limitations on the questioner to ask questions. The F1 score for this model is around 88.8.
* Regulatory datasets: There is another model that implemented conversational question-answering system for regulatory datasets. The answers to the questions in these datasets are of ‘yes’ and ‘no’ form and in rare cases text form in case the question demands answers in the form of clarifications.
* Seq2seq model implementation: This model gives the least performance because it predicts the same answers to multiple questions even though these answers do not appear in the passage.
* PGNet implementation: The shortcomings of the seq2seq model are improved in this model by considering only the vocabulary from the passage. This definitely improves the performance, but the disadvantage is that the passage has to be memorized by the model in order to map the answers.
* DrQA: This model improves the PGNet, but fails if the answer text overlaps the text from passage. The augmented DrQA model performs the best so far giving the accuracy of 65.4.
* BERT for Question Answering on SQuAD: This model has been implemented using LSTM Encoder, LSTM Decoder on top of the BERT base uncase model and achieved an F1 score of 77.96.

**Problems and shortcomings:**

* As mentioned above, the major problems with the models implemented for the conversational question-answering systems vary from model to model.
* The nature of the datasource creates problems to some models which have been implemented for machine-friendly databases, images and videos.
* Some models take into account exact text spans from the passages as answers to the questions which reduces the accuracy of the model for the questions whose answers are not found in the text directly. On the other hand, some models predict answers different from the passage text and reduce the accuracy if the expected answer text is directly derived from the passage.

**Our Model:**

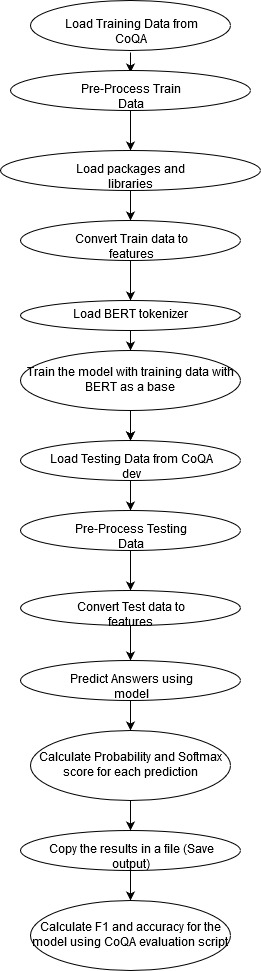
* For our model, we predict the answers based on the text spans from the passages. Since the predicted and expected answers are from the passages itself, the accuracy is more than many question-answering models implemented before.
* Also, BERT considers all the hidden states from the previous layers while making predictions. Hence while predicting the answer to the current questions the model considers previous context to some extent whenever required, even though this can be improved significantly.
* **Project**

**Approach**

We developed a model that predicted answers to the questions based on the context that was a passage given to it. The inputs to this model are the questions and the related passage. The output is the set of answers to these questions. The answers are nothing but the text spans from the passage that is the start and end index of the text in the passage that contains the answer. We used BERT to achieve this. The pre-trained BERT word embedding are highly contextualized and provide a good accuracy to the question-answering system models in natural language processing. Following are the steps that we performed to train our model and get the required solution to our problem statement:

1. Load the training data from the CoQA train dataset.
2. Pre-process the train data.
3. Load other packages and libraries that were needed.
4. Convert train data into features.
5. Load the BERT tokenizer and other BERT dependencies.
6. Train the model for using train dataset.
7. Load the testing data from CoQA dev dataset.
8. Pre-process the test data.
9. Convert test data into features.
10. Evaluate the model for test data.
11. Write predictions.
12. Calculate the probabilities and scores for each answer predicted by the model.
13. Copy the results in the prediction file.
14. Calculate F1 score and accuracy for the trained model.

**Flow Diagram:**



**Contributions:**

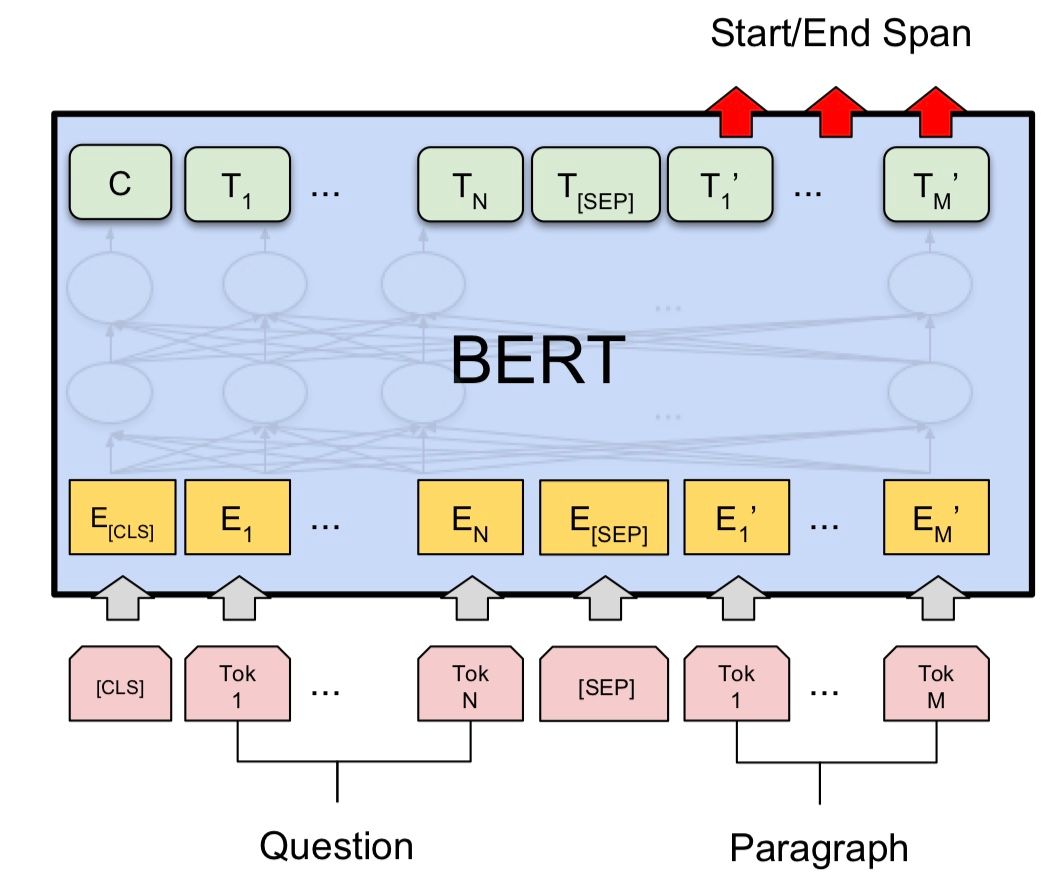
|  |  |  |
| --- | --- | --- |
|  | Team Member Name | Responsible For |
| 1 | Rutuja Saptarshi – 801100848 | 1. Data Preprocessing 2. BERT Implementation 3. Writing predictions 4. Documentation |
| 2 | Sagar Talwar - 801100599 | 1. Preliminary Research 2. Training the model 3. Writing predictions 4. Testing & Evaluation |

**What did/didn't work?**

* The model trained worked as expected, but the training data had to be reduced while training our model due to hardware limitations while training over the whole dataset.
* The model is based on BERT which provided many advantages over sequential RNNs. It worked as expected.
* The predictions were partially correct for all the passages and related questions.
* The accuracy for the model is average and has scope for improvement.

**Graphs and Diagrams:**

* BERT for our model:



**Results**

Predictions for 1 passage from the dataset:

"3dr23u6we5exclen4th8uq9rb42tel 4": [

{

"text": "in a barn near a farm house , there lived a little white kitten named Cotton",

"probability": 0.09644876605280879,

"socre": 3.958984375

},

{

"text": "Cotton",

"probability": 0.07358963103428157,

"socre": 3.6884765625

},

{

"text": ".",

"probability": 0.07090881516106831,

"socre": 3.6513671875

},

{

"text": "horses slept . But Cotton was n't alone in her little home above the barn , oh no . She shared her hay bed",

"probability": 0.06892882466570233,

"socre": 3.623046875

},

{

"text": "in a barn near a farm house , there lived a little white kitten",

"probability": 0.06368651744578366,

"socre": 3.5439453125

},

{

"text": "shared her hay bed with her mommy and 5 other sisters",

"probability": 0.05942036101259978,

"socre": 3.474609375

},

{

"text": "her hay bed with her mommy and 5 other sisters",

"probability": 0.0563130334370416,

"socre": 3.4208984375

},

{

"text": "white kitten named Cotton",

"probability": 0.05442079384493854,

"socre": 3.38671875

},

{

"text": "shared her hay bed",

"probability": 0.0534725370332517,

"socre": 3.369140625

},

{

"text": "her hay bed",

"probability": 0.050676244886469796,

"socre": 3.3154296875

},

{

"text": "hay bed with her mommy and 5 other sisters",

"probability": 0.04317670044160518,

"socre": 3.1552734375

},

{

"text": "in a barn near a farm house , there lived a little white",

"probability": 0.041239748909984386,

"socre": 3.109375

},

{

"text": "hay bed",

"probability": 0.03885482474345429,

"socre": 3.0498046875

},

{

"text": "She shared her hay bed with her mommy and 5 other sisters",

"probability": 0.03644730119315952,

"socre": 2.98583984375

},

{

"text": "white kitten",

"probability": 0.03593483855170733,

"socre": 2.9716796875

},

{

"text": "horses",

"probability": 0.035655191419489765,

"socre": 2.9638671875

},

{

"text": "She shared her hay bed",

"probability": 0.03279902090127054,

"socre": 2.88037109375

},

{

"text": "her mommy and 5 other sisters",

"probability": 0.03002468144116868,

"socre": 2.7919921875

},

{

"text": "kitten named Cotton",

"probability": 0.029616981114142432,

"socre": 2.7783203125

},

{

"text": "a barn near a farm house , there lived a little white kitten named Cotton",

"probability": 0.028385186710071977,

"socre": 2.73583984375

}

]

Accuracy:

{

"children\_stories": {

"em": 13.0,

"f1": 13.9,

"turns": 1425

},

"literature": {

"em": 12.1,

"f1": 12.7,

"turns": 1630

},

"mid-high\_school": {

"em": 13.3,

"f1": 13.8,

"turns": 1653

},

"news": {

"em": 12.2,

"f1": 12.5,

"turns": 1649

},

"wikipedia": {

"em": 10.3,

"f1": 10.4,

"turns": 1626

},

"reddit": {

"em": 0.0,

"f1": 0.0,

"turns": 0

},

"science": {

"em": 0.0,

"f1": 0.0,

"turns": 0

},

"in\_domain": {

"em": 12.2,

"f1": 12.6,

"turns": 7983

},

"out\_domain": {

"em": 0.0,

"f1": 0.0,

"turns": 0

},

"overall": {

"em": 12.2,

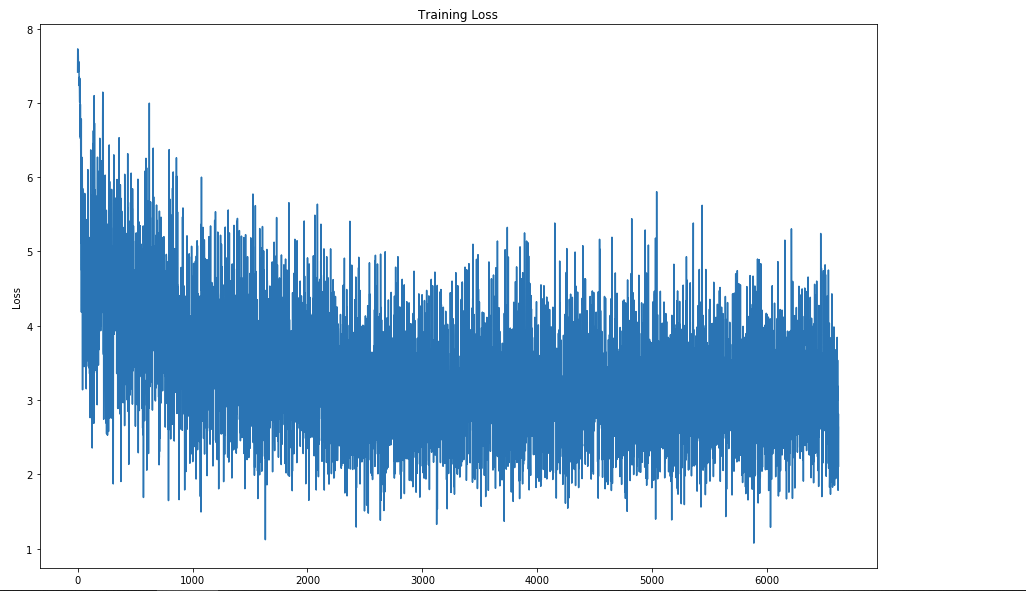
"f1": 12.6,

"turns": 7983

}

}

Graph: Batch Vs Training Loss:



* As we see in the results, the F1 score is calculated domain wise for the predictions.
* The model gives the best performance for the ‘children’s stories’ domain while the least for ‘science’. The overall of F1 score of the model is around 12.5 which is almost acceptable.
* **Summary**:
* Conversations are an important part of human lives as it helps people to gain and give knowledge to each other. Reading comprehensions give a context over which one can ask questions to another person and receives an answer for the same.
* The nature of this question-answering conversations for human beings is simple as we are able to process, store and give the answer given our cognitive abilities. However, when a computer has to be trained for such question-answering conversations, a high level of complexity is introduced as we need to train the computer to process, store and give the correct answers to the questions and also decide if the answer given by a human is correct or not.
* The nature of the conversation is incremental, the current question asked is more or less based on the answer of the previous questions. This brings more complexity to the model as the previous answers needs to be stored or the previous answers should be passed to the current question.
* We developed such a model using BERT as our base, where in a passage which is the context for the conversation is inputted to the model. With the passage, we also pass a set of questions to the model and the expected output is the set of answers to the questions asked.
* Our model finds out the text spans from the passage that contained the answers to the questions asked. Thus the output is the start and end index of the text span from the passage.
* This output is predicted by the model correctly with an average accuracy.

Important results

* The accuracy for the developed model is around 12.5.
* The accuracies are calculated domain wise for the predicted answers based on the passage and related questions. We observe that the accuracy is the highest for the domain ‘children’s stories’ which is 13.9 and least for the domain ‘science’ which 0.0.
* The graph is plotted for Batch vs Training Loss. We observe the nature of the graph to be consistently linear.
* **Conclusions**:

Accomplishments:

* We successfully developed a model with considerable accuracy that predicted the answers to the questions based on the given passage.
* This model was trained using BERT. The results were calculated on the development data set.
* We also calculated the softmax score for every prediction and it was found that we achieved an average accuracy for the model overall.
* We learnt the BERT model for CoQA dataset which was very useful since it helped us to increase the accuracy for our model.

Other possible approaches:

* Instead of using BERT for our model, we had the following options for our model implementation:

**1. Conversation Response Generation Problem Model: Sequence to Sequence Models are this type of models. In this, we append the conversation history and the current question to passage and store it in the short-term memory encoder. The answer is generated from the decoder.**

**2. Reading Comprehension Problem Model: These models focus on finding the extracts from the passage that closely match the questions. These models are effective learners since the answers are limited to spans, they cannot handle answers that do not overlap with the passage. The Document Reader model (DrQA) is a type of this model.**

**3. Combined Model: In this, we combine the above to models Sequence to Sequence and DrQA to develop a model in which we first find the extract evidence from the passage close to the original answer and then naturalize this evidence to the answer. We make the following changes to the two models to develop the combined one:**

**1. For DrQA, we need to model in such a way that it predicts directly if the answer is the substring of the rationale and vice versa.**

**2. For sequence to sequence, we input DrQA’s extract prediction to the encoder and the decoder predicts the final answer.**

Future work:

* We used pre-trained contextualized word embedding from BERT.
* Our model does not take into consideration the previous answers while predicting the answer for the current question.
* This reduces the accuracy of the model to a considerable extent.
* This could have been improved if we were to store the previous answers and use the same while predicting the answer for the current question.
* In addition to this, we can also implement a graphical user interface for the question-answer system.
* **References**:
* **Reference Paper: CoQA: A Conversational Question Answering Challenge**

<https://www.aclweb.org/anthology/Q19-1016>

* **Dataset**

**Train-**<http://downloads.cs.stanford.edu/nlp/data/coqa/coqa-train-v1.0.json>

**Dev -**<http://downloads.cs.stanford.edu/nlp/data/coqa/coqa-dev-v1.0.json>

* Technical report on Conversational Question Answering <https://arxiv.org/pdf/1905.12848.pdf>
* BERT for Question Answering on SQuAD 2.0 <https://web.stanford.edu/class/cs224n/reports/default/15848021.pdf>
* Understanding Text with Bert <https://blog.scaleway.com/2019/understanding-text-with-bert/>
* Natural Language Processing: the age of transformers <https://blog.scaleway.com/2019/building-a-machine-reading-comprehension-system-using-the-latest-advances-in-deep-learning-for-nlp/>
* Hugging Face

<https://huggingface.co/transformers/>

<https://huggingface.co/transformers/model_doc/bert.html>

# Bert Explained: State of the art language model for NLP

# <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>

# <https://www.lyrn.ai/2018/11/07/explained-bert-state-of-the-art-language-model-for-nlp/>

# CoQA Official Evaluation script: <https://nlp.stanford.edu/data/coqa/evaluate-v1.0.py>

|  |  |  |
| --- | --- | --- |
|  | |  |
|  |