# SQuAD 2.0 Question Answering System using BERT-based Model

*CS6120 - Natural Language Processing - Final Group Project*

Morgan Levy, Keshav Bharadwaj Vaidyanathan, Bhanu Sai Simha Vanam, Courtney Wilkerson

*Khoury School of Computer Science, Northeastern University, 360 Huntington Avenue, Boston, MA, 02115, U.S.A.*

Submitted April 24th, 2022

## Abstract

Question Answering systems have become the base drivers of many intelligent technologies intended for human interaction and their prevalence is only increasing. One such type of these systems utilizes Reading Comprehension datasets to train a model aimed to answer a Question given associated context. The focus of the project was to create such a system, utilizing the SQuAD 2.0 dataset to train models to answer questions given context. The Stanford Question Answering Dataset (SQuAD 2.0) is a reading comprehension dataset that goes above and beyond the scope of the original iteration, SQuAD 1.0, by incorporating over 50,000 unanswerable questions for training.

Our group did not have much need to deliberate on initial model selection, as the SQuAD 2.0 is a widely utilized and recognized dataset in the field of question answering, and therefore many previous model implementations by various academic and private institutions have highlighted the substantial comparative performance of models that utilize Bidirectional Encoder Representations from Transformers (BERT). The group decided to use metrics from Long Short-Term Memory model implementations as well as from Human Performance to serve as a baseline, and focused on improving performance of  Light BERT implementations for our custom model.

## Introduction

In approaching the design of a Question Answering System, the team was fortunate to have a myriad of examples from which to learn what would best constitute a design for the most accurate model possible. One of the benefits of using SQuAD 2.0 was the provision of a leaderboard, where exact-match and f1-scores were provided for the best submissions of models over the past few years. This gave us additional benchmarks or baselines with which to compare our own model's performance as we iteratively made improvements to the design or modifications to the provided QA Reading Comprehension data. We started with a Baseline LSTM-based model and went on to the near state of the art transformer based BERT model.

## Datasets

*SQuAD 2.0 - Stanford Question Answering Dataset*

The reading comprehension dataset provided by Stanford features 150,000 instances with 4 features, one being questions written adversarially by crowdworkers. This includes over 50,000 questions written adversarially by crowdworkers, designed to be unanswerable given their provided context paragraph. The addition of these unanswerable questions is what constitutes the improvement of the SQuAD 2.0 dataset compared to earlier versions, and what makes the resulting model better, as the dataset gives the models the opportunity to define and recognize unanswerable questions, a very important feature if the system is going to be human-facing.

*Fields*

The training dataset contains four main features, a question, a context paragraph, an index of the answer within the paragraph, and the answer. All fields are represented as Strings other than the Answer Index which is represented as an Integer.

For example, the very first entry in the dataset has the following fields: Question - "What areas did Beyonce compete in when she was growing up", Context - Sample text from the Beyonce Wiki, Answer - "in the late 1990's", and Answer Index of 269, meaning the start of the answer text occurs at index 269 within the context paragraph.

*Preprocessing*

Preprocessing of the dataset ahead of training was limited, as the dataset is designed by Stanford's team for immediate utilization, and the group's decision to move forward with BERT-based models eliminated the need for the removal of such things as stopwords, given that Light BERT implementations link long-term dependencies between context inputs and model output, which highly benefit from the presence of context-adding sentence structures and syntax.

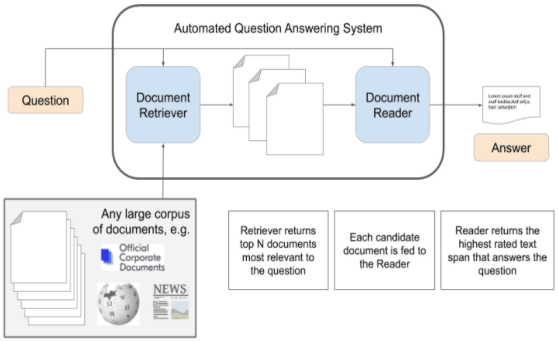
## Methods & Baselines

*Human Performance*

In addition to a baseline and the availability of accuracies and F1-Scores from previously submitted, high-performing QA implementations, Stanford's dataset website also provides a Human Performance accuracy, which is a measure of average human ability to accurately answer (or deem unanswerable) a standard reading comprehension question given a written passage. The exact-match and F1-scores for human-performance on the dataset were 92.83% and 95.45%, respectively.

*Model Diagram*

**Figure 1.** *Simplified diagram of an automated Question-Answering System training on Reading Comprehension Data*


Automated Question Answering System
Question
Document Retriever
Document Reader
Answer
Any large corpus of documents, e.g.
Official Corporate Documents
NEWS
Retriever returns top N documents most relevant to the question
Each candidate document is fed to the Reader
Reader returns the highest rated text span that answers the question

In Figure 1, you can see the flow process of our Question Answering System. First, a Question is posed to the model, secondly the Document Retriever returns top N Wikipedia documents most relevant to the given question, third each candidate document is fed to the Reader and scored, fourth and finally the Document Reader returns the highest rated text-span that answers the question from the chosen context.

*Baseline Model - Long Short-Term Memory*

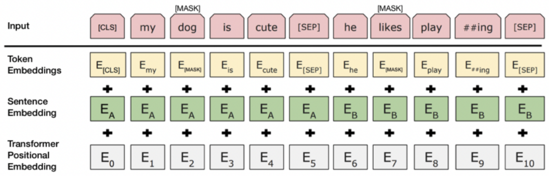
In addition to comparison to average human performance of context-based Question Answering, a baseline model utilizing the RNN architecture of Long Short Term Memory was used to compare the performances of our final model selection. The top performing LSTM-based model on the QA data in a study conducted by Stanford, combined with character embedding, exact match, and self attention, achieved an exact match accuracy of 63.57 and an F1-score of 66.74.

*Distilled BERT Implementation*

Distilled BERT is a quantized and pruned version of BERT which benefits from faster training, shipping, and deployment. The advantages of Distilled BERT in comparison to BERT-Base include having 40% fewer parameters, 60% faster training speed, ~4.5x faster inference time for standard NLP processing tasks, and a smaller relative size for easier deployment and tuning while retaining 95% of the performance.

Below, in Figure 2, you can see a diagram representing a Transformer-based Model ( BERT). You can see that the model utilizes a stack size of three, allowing it to break down the input into three embedding layers, the first for the word tokens, the second for sentence/clause separation, and the final for positional embedding within the whole corpus. The use of several multi-head attention blocks perform the effective fusion of context and question that Bidirectional Attention Flow/Transformer-based models  benefit so highly from.

**Figure 2.** *Diagram of a Transformer Based Model*



## Results

*Training the Model*

The training of our Distilled-BERT model was completed on an NVIDIA Tesla P100 GPU with Discovery Clusters. The following parameters were determined to be optimal for the model.

**Table 1.** *Optimal Parameters from Distilled BERT Training on Dataset*

| Optimal Distilled BERT Parameters | | | |
|---|---|---|---|
| **Name** | Distilbert-base-uncased | **Num Heads** | 12 |
| **Activation** | 'gelu' | **Num Layers** | 6 |
| **Architecture** | DistilBertForQuestionAnswering | **Pad Token ID** | 0 |
| **Attention Dropout** | 0.1 | **QA Dropout** | 0.1 |
| **Dimension** | 768 | **Sequential Classification Dropout** | 0.2 |
| **Dropout** | 0.1 | **Sinusoidal Position Embeddings** | False |
| **Hidden Dimension** | 3072 | **Tie Weights** | True |
| **Initializer Range** | 0.02 | **Torch D-Type** | float32 |
| **Max Position Embeddings** | 512 | **Vocabulary Size** | 30522 |
| **Model Type** | distilbert | | |

*Performance Metrics and Final Results*

Training and evaluation of the DistilBERT model yielded the following results and performance metrics contained in Table 2.

**Table 2.** *DistilBERT Performance Metrics & Results*

| Final Performance of DistilBERT Implementation | | | |
|---|---|---|---|
| **Train Loss** | 1.15 | **Epochs** | 5 |
| **Train** | 3807.09 | **Exact Match** | 77.32 |

| Runtime | | Score | |
|---|---|---|---|
| Train Samples | 88524 | Evaluation F1 Score | 85.42 |
| Train Samples/ second | 46.50 | Evaluation Samples | 10784 |
| Train Steps/ second | 3.87 | | |

*Example Model Run*

After training the model, it was imperative to test it on question/context pairs not provided in the training dataset. To test our model initially, we decided to write in our own Context Paragraph and related Question: Context - “*We the team consisting of bhanu, keshav, courtney, and morgan have built a model to answer questions. We are students at northeastern university and NLP is our favorite subject*”, Question - “*Who are the team members?*”. The group was delighted to find that our model was functioning as intended when the model responded correctly with: “*bhanu, keshav, courtney, morgan*”.

We wrote several other context-question pairs and the model performed as expected for those questions as well. We did not test an unanswerable question, but have plans to do so in further development stages following the semester’s end.

This Example along with several others are demonstrated in our inference notebook.

## Conclusion

*Final Model Performance*

Our Light BERT-based implementations performed well, as expected, with our final model testing accuracy for Question Answering set at 77.32 exact match accuracy and an 85.42 F1 Score over 10784 evaluation samples. These scores put our model in above the 15th percentile of model performance on the full testing set submitted to the official SQuAD website, which is a promising performance but also verifies the room for improved capabilities of other implementations of BERT-based models for Question Answering.

Our model outperformed the baseline LSTM implementation, but still has a ways to go ahead of beating the human-performance baseline scores.

## Acknowledgements

A big thank you to Professor Uzair and all Teaching Assistants for an informative class and a great semester!

## References

- *Zhang Rajpurkar and et al Lopyrev. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250, 2016, 2016*
- *Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable questions for squad. arXiv preprint arXiv:1806.03822, 2018.*
- *Sanh Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108, 2020.*
- *Yuan, Bithiah. (2021, January 7). Financial Question Answering with Jina and BERT - Part 1 [Blog Post]. Retrieved from https://towardsdatascience.com/ how-to-build-a-production-ready-financial-question-answering-system-with-jina-and-bert-48335103043f*