

Building a Language Understanding and Question Answering System from open-ended Trivia Data

Team Members

- Sreecharan Vanam
- James Cruz
- Sai Praneeth Kankanala
- Yashwanth Kirla

Contents

Introduction:	2
Problem Statement:	3
Methodology:	4
Workflow Diagram:	4
Architecture:	5
Dataset:	6
Description:	9
Exploratory Data Analysis:	10
Data Visualization	10
Data Preparation	10
Implementation:	10
Algorithm/Pseudocode:	11
Libraries Used:	12
Results:	12
Performance:	12
Visualizations:	14
Project Management	19
Completed Work:	19
Description:	19
Issues faced:	20
References:	20

Introduction:

Topic/Area:

Natural Language Processing (NLP) is ever-evolving; cutting-edge research in question-answering (QA) systems is one significant area of application and study. Such systems are a crucial part of many services such as customer helplines, health and physical well-being providers, and education. They objective to move beyond basic keyword searching towards being able to read, understand questions like humans and provide most accurate answers grounded in the specific context and text. Some of the most technologically advanced models that can be trained on text that belongs to the family of transformers, with Google's BERT (Bidirectional Encoder Representations from Transformers) being the most well-known and well optimized for Q&A tasks.

Motivation:

The motivation of this project is to examine if transformer-based models like BERT can improve the accuracy and efficiency of Q&A systems. Even though modern NLP has achieved impressive results, especially recently, there are still challenges we'd like to address: how to handle context, deal with imprecise and ambiguous queries, and make Q&A systems more accurate and stable when datasets are varied. By tackling these kinda problems, we can improve user experience and increase the deployment of Q&A technologies to real user scenarios.

Significance

Finally, we hope that increased performance in Q&A systems will change the way people use information technology: better Q&A systems will answer more questions more quickly, which will make those answers more useful in people's decision-making processes, and give them access to more information they might not have encountered otherwise. Especially for those who will need to interact with exponentially growing volumes of digital data, powerful Q&A systems will soon be critical tools for managing huge amounts of information, allowing people to find answers from jumbo datasets without having to supervise the search process themselves. Devlin and J Chang's[3] project tries to take advantage of the limits of fine-tuning a large BERT model on the SQuAD dataset to see just how far we can go in letting machines answer trivia.

Contribution:

This text mining project aims at building a well-designed question answering system that uses BERT model fine-tuned with SQuAD dataset. The major contributions are:

Implementation and Optimization: This project demonstrates the applicability and extendability of the transformer model, by using-an existing BERT model as a basis, and applying it for the task of QA.

Performance Analysis: Extensive evaluation and testing of the model's performance in various questions and contexts reveals to what sort of questions it is most suited when put into operation.

Knowledge Sharing: The detailed findings and methodologies of this project are documented, expanding the database of academic and practical knowledge of NLP and theoretical and practical guidance on how to improve and apply such technology in the future.

Results from this project will enable actionable advances in the deployment and improvement of NLP models for QA, both for this project and future developments such as natural-sounding written content that can be translated immediately into human-sounding speech and many more applications.

Problem Statement:

Natural Language Processing is a fast-paced field, and question-answering systems now underpin all sorts of products in industry, academia and beyond, such as chatbots in automated customer service delivery, query systems that help seek out datasets, and accurate and rapid ways to support academic research. However, despite recent progress, today's models of natural language question-answering exhibit certain challenges that, while minor in some applications, detract from the usefulness of this field and its application to important problems:

Contextual Understanding: Most existing QA systems cannot tell when the question is referring to information in a previous paragraph of text, versus when new information is provided, so they often come up with incorrect or irrelevant answers.

Handling Ambiguity and Complexity: Because questions are framed with ambiguities or in non-literal ways, they will be harder to interpret correctly. Several current systems will fail to interpret the nuances of such questions, and make responses that do not satisfy the user's actual request at all.

Domain Adaptability: Although QA systems are typically trained on broad datasets, they tend to perform poorly outside these broad domains when tested on niche or specialised content, as training on such narrow content is not widely available.

Hypothesis

This project's hypothesis is that, using a BERT-style deep learning-based mechanism (BERT refers to a transformer-based model that uses deep learning to understand context and nuances and how words interact), all the above problems can be greatly addressed. More specifically, the hypothesis is:

If we fine-tune BERT on the SQuAD dataset, an architecture that's probably going to be ahead of any other model in these complex question and context-appropriate answer issues.

Fine-tuning will allow the model to better answer questions from new domains, making it more general, or 'rugged', in its ability to answer questions from multiple fields.

These hypotheses should be testable and analysable empirically, providing a rigorous basis for assessing the potential of BERT to dramatically enhance question-answering systems, and hence of delivering a robust, efficient and adaptable system for real-life applications.

Methodology:

Workflow Diagram:

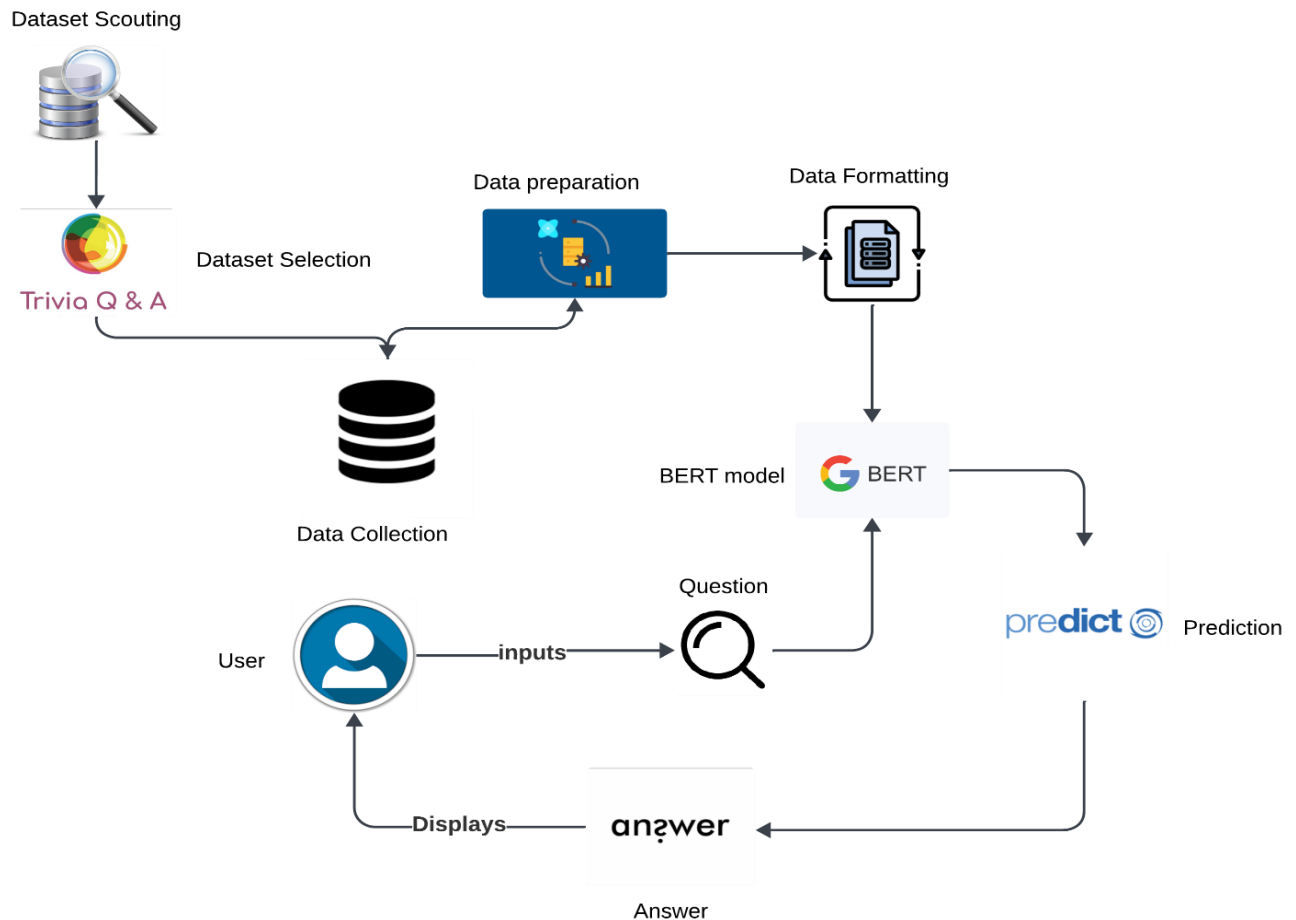


Fig.1 Workflow Diagram

This workflow shows the steps involved in collecting, preparing, formatting the selected Squad format dataset from the suggested links provided thereby training the language understanding model, developing the question answering system, finally integrating the two components into a fully functional domain specific QA system. This breakdown & workflow are very clear and concise for understanding the complete scope of the project.

After selecting the appropriate dataset, we prepare the data by structuring the data to make the model actually understand the question-and-answer contexts. After preprocessing and adjusting the data formatting to feed into the model, we only require Question, Context and Answer sections to be fed into the language understanding model which will be trained to this specific set of Q&A pairs which will lead to better performance, based on the satisfaction and accuracy of the answer outputs, finetuning will be required to ensure consistent and reliable QA system.

Finally, the user will be able to input questions regarding the trivia questions dataset because the model is only limited to the dataset, we prepared any questions that are not related to trivia questions might trigger inaccurate answers, the model will be able to predict and provide the answer which is displayed to the end user. This is the overall structure of our project and procedures involved.

Architecture:

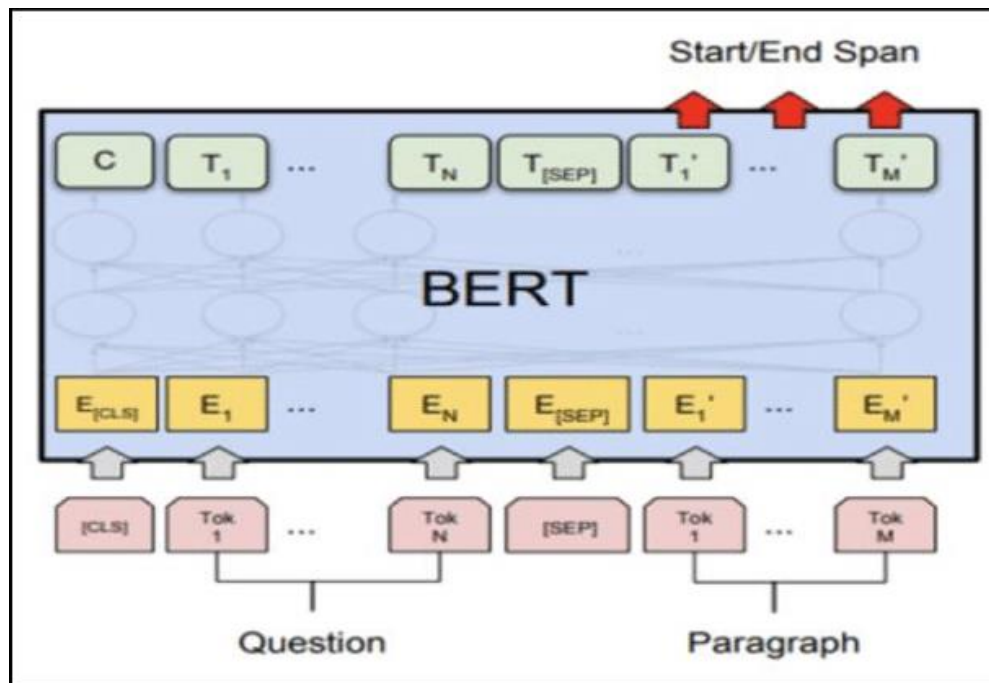


Fig.2 Architecture of BERT Language model on Squad Formatted Data

The components of BERT architecture are: Input Embeddings, [CLS] and [SEP] Tokens, Transformer Layers, Output Predictions, Fine-tuning Mechanism.

The architecture of BERT gives some insight as to how the system actually makes its predictions. Towards the bottom is where the input embeddings live. Every token of the question and the paragraph is turned into a vector; each vector represents a one-hot encoding of a token along with semantic information about that token. The bottom row also contains three special tokens - [CLS] (classification) to mark the start of the input, [SEP] (separator) to separate the question and the paragraph, and [SEP] to mark the end - to help the model distinguish where the question ends and the paragraph begins. BERT model basically depends on a multi-layered architecture called a transformer. These layers take embeddings as inputs and distil complex co-occurrences & relationships between words. The top of the architecture produces the final answer outputs. For a Q&A task like Trivia Q&A, BERT ultimately needs to predict the start and end spans of the answer that appears in the paragraph. These predictions are derived from the final hidden states for each token. To adapt BERT for the Trivia Q&A task, the model was fine-tuned. It was retrained on the Trivia dataset beyond the point at which all the weights in BERT were trained on

the enormous Wikipedia dump. Typically, fine-tuning involves feeding a lot of the same data that was used to train the transformer initially.

The combination of BERT's pre-training on a very large corpus, and fine-tuning on task-specific data, allowed us to achieve very good performance on the Q&A task (according to the results of this project).

Dataset:

Dataset Information:

```

1 {
2   "data": [
3     {
4       "title": "Teacher",
5       "paragraphs": [
6         {
7           "context": "A 2000 study found that 42% of UK teachers experienced occupational stress, twice the figure for the average profession. A 2012 study found that teachers experienced double the rate of anxiety, depression, and stress than average workers.",
8           "qas": [
9             {
10              "id": "21a3551031f2e09338196801f5c1ac2e08905b6",
11              "question": "What is teaching not considered due to stress?",
12              "answers": [
13                {
14                  "answer_start": 181,
15                  "text": "average profession"
16                }
17              ]
18            },
19            {
20              "id": "b6a71a728a35506dd4cd2179c9342acf5e1a047a",
21              "question": "What do normal workers not have to deal with as much?",
22              "answers": [
23                {
24                  "answer_start": 185,
25                  "text": "anxiety, depression, and stress"
26                }
27              ]
28            },
29            {
30              "id": "6feb72fec8aeb105ab955897415fe4b2c28",
31              "question": "How was teacher depression realized?",
32              "answers": [
33                {
34                  "answer_start": 123,
35                  "text": "2012 study"
36                }
37              ]
38            },
39            {
40              "id": "53e4ce951aeb3a2712aa0966d73ec3e2c62ca11",
41              "question": "How was teacher anxiety realized?",
42              "answers": [
43                {
44                  "answer_start": 121,
45                  "text": "A 2012 study"
46                }
47              ]
48            }
49          ]
50        },
51        {
52          "context": "30 US states have banned corporal punishment, the others (mostly in the South) have not. It is still used to a significant (though declining) degree in some public schools in Alabama, Arkansas, Georgia, Louisiana, Mississippi, Oklahoma, Ten",
53          "qas": [
54            {
55              "id": "a8a308feb5e9edf56fe63556cf20c039c3fdaca",
56              "question": "Where is Mississippi?",

```

Preview of Large Json file used to fine tune the Bert model.

Features of dataset:

Total number of articles: 21

Total number of paragraphs: 283

Total number of questions: 1000

Article names:

1. Teacher

2. Sky_(United_Kingdom)

3. Genghis_Khan
4. Apollo_program
5. University_of_Chicago
6. Economic_inequality
7. Chloroplast
8. Immune_System
9. Oxygen
10. Jacksonville,_Florida
11. Martin_Luther
12. Steam_engine
13. Super_Bowl_50
14. Force
15. Newcastle_upon_Tyne
16. United_Methodist_Church
17. Warsaw
18. Kenya
19. Intergovernmental_Panel_on_Climate_Change
20. Fresno,_California
21. Amazon_rainforest

Data format:

```
{  
  "data": [  
    {  
      "title": "*Category Title*",  
      "paragraphs": [  
        {  
          "context": "*Context Paragraphs*",  
          "qas": [  

```

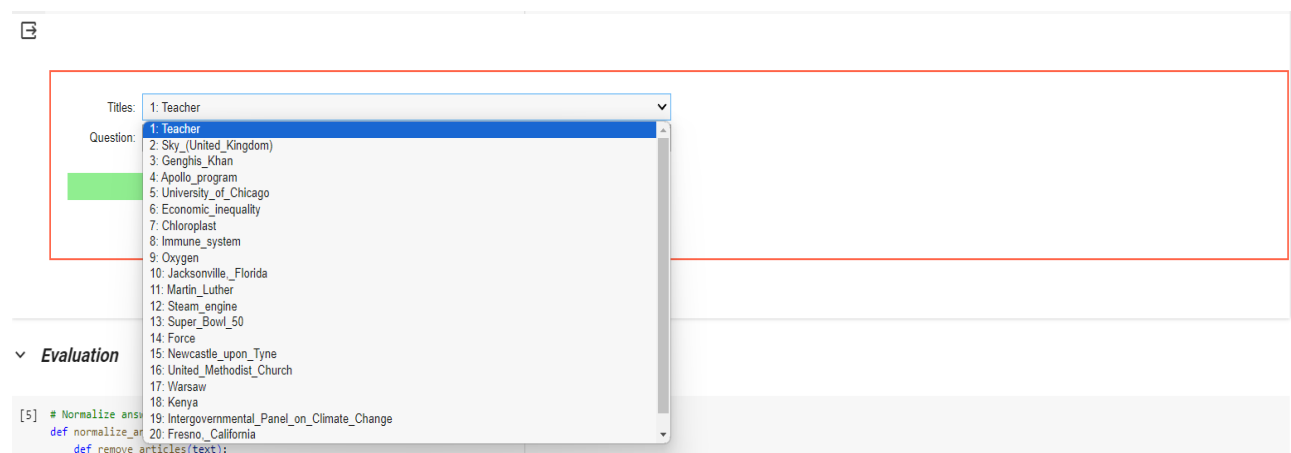
```

{
    "question": "*question text*",
    "id": "question unique ID",
    "answers": [
        {
            "text": "*****",
            "answer_start": *An integer pointing at answer index*
        }
    ]
},
// Additional questions and answers
]
},
// Additional paragraphs
]
},
// Additional articles
]
}

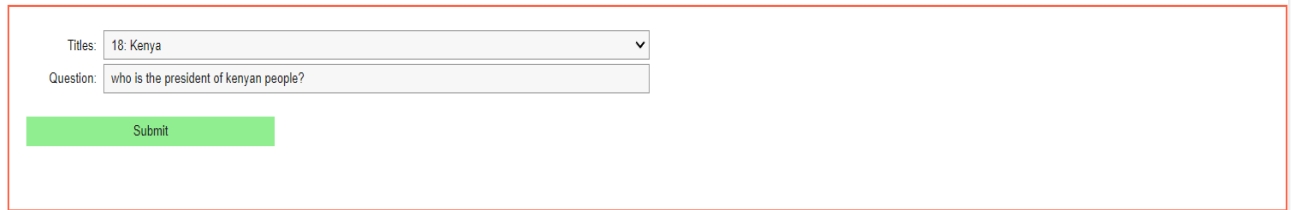
```

User-Interface:

1. Select a topic you need information about:



2. Input a question can be in any human written form with typos too:

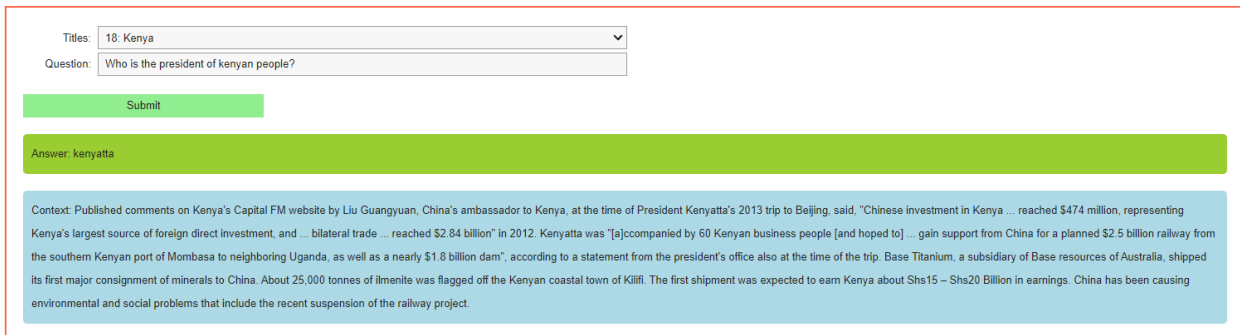


Titles: 18: Kenya ▼

Question: who is the president of kenyan people?

Submit

3. Finally, retrieves the Answer and Context along with it:



Titles: 18: Kenya ▼

Question: Who is the president of kenyan people?

Submit

Answer: kenyatta

Context: Published comments on Kenya's Capital FM website by Liu Guangyuan, China's ambassador to Kenya, at the time of President Kenyatta's 2013 trip to Beijing, said, "Chinese investment in Kenya ... reached \$474 million, representing Kenya's largest source of foreign direct investment, and ... bilateral trade ... reached \$2.84 billion" in 2012. Kenyatta was "[a]ccompanied by 60 Kenyan business people [and hoped to] ... gain support from China for a planned \$2.5 billion railway from the southern Kenyan port of Mombasa to neighboring Uganda, as well as a nearly \$1.8 billion dam", according to a statement from the president's office also at the time of the trip. Base Titanium, a subsidiary of Base resources of Australia, shipped its first major consignment of minerals to China. About 25,000 tonnes of ilmenite was flagged off the Kenyan coastal town of Kilifi. The first shipment was expected to earn Kenya about Shs15 – Shs20 Billion in earnings. China has been causing environmental and social problems that include the recent suspension of the railway project.

We also intended to display the context to show how effective our system is in producing accurate trivia type of answers which are just 1-3 words in general.

Description:

The data used here is an NLP and QA benchmark, the Stanford Question Answering Dataset (SQuAD). SQuAD is very popular and continues to dominate machine learning research on NLP and QA tasks. It was derived from human paraphrases of more than 500 Wikipedia articles and is made up of more than 100,000 question-answer pairs created by people called crowdworkers who read the articles then write questions, as if marking another person's comprehension of the text. The answers to the questions are segments of the source text providing an answer to the question asked. SQuAD is made up of entries made up of Context, Question and Answer.

Exploratory Data Analysis:

Data Visualization

We have used: Histograms, Word Clouds, Scatter Plots.

Histograms showed the answer lengths in words vs the questions in length. This visualisation of the SQuAD dataset helped in revealing important insights. The dataset was divided into training and evaluation sets, with the latter used for validation purposes. During the time of exploratory data analysis, we can find the following plots helpful: Histograms were generated to show us the length of questions and answers in words, also the distribution of answer lengths in words.

Likewise, we scatter plots that relate the question length in words to the answer length in words. Some keywords were displayed in word clouds to see which terms appeared most. The scatter plots have given an impression of the central tendency and dispersion of answer length, and any potential outliers as well. Visualizations reveal major important insights in the selected dataset.

Data Preparation

There's a long pre-processing step where the raw data was being carefully prepped to take a form that is suitable for this BERT model training. The first thing was to break the sentences from the textual data into subword tokens (or lexical units that make up the vocabulary of BERT). Then we cleaned the text, filtering out unwanted characters and formatting that could confuse the model. Then we normalised the text by lowercasing it to treat all the tokens uniformly so that the model wouldn't misinterpret any of the tokens. And most importantly, we mapped the position of each answer in the context to its tokenised position to allow a model to learn and predict answer spans. Truncation and padding were performed to make sure the lengths of all the inputs were adjusted to BERT's constant length for inputs, and the model will architecturally focus the attention on the meaningful grids (the ones that don't contain the padded tokens) with built-in attention masks to accommodate those changes. You'd want your model to have the best environment to train in, with the best values in its input during training.

We have done: Tokenization, Cleaning, Normalization, Answer mapping, Truncation & Padding and Attention Masking.

Implementation:

Explanation:

Loading the Model: We start to load the pre-trained BERT model for Question Answering along with its tokenizer. The tokenizer converts text into tokens that the model understands. The model is pre-trained on a large corpus of text.

Preprocessing Data: Next, the input to each example in SQuAD is tokenised into tokens, mapped to their index positions in the BERT vocabulary and will be split into a pair of samples', using attention masks 'to indicate where the model should pay attention'.

Fine-tuning the Model: The BERT pre-trained model Fine-tuning refers to additional training with the goal of adapting the pre-trained BERT model to a specific downstream task. In this case, we consider fine-tuning on the SQuAD dataset. The BERT pre-trained model is loaded and then fine-tuned using additional training steps, in which the answer start position is predicted in the context paragraphs given a question. The model weights are adjusted to minimize the gap between a predicted answer span and a human annotated answer span.

Model Evaluation: A validation set is kept separate from the previous steps in order to evaluate the performance of the model. The model's quality is measured with metrics like F1 score as well as Exact Match.

Inference: Given a new question and context, the model converts them to the output of the encoders, and then predicts the start and end indices of the answer span in the context. These indices are mapped back to the original text to produce the final answer.

Algorithm/Pseudocode:

Inputs:

dataset: A collection of context paragraphs, questions, and answers
tokenizer: BERT tokenizer
model: Pre-trained BERT model

Procedure:

- 1: Load the pre-trained BERT model and tokenizer
- 2: Preprocess the dataset
 - For each example in the dataset:
 - Tokenize the context and the question
 - Identify the answer span in the context
 - Encode the tokens to input IDs with attention masks
- 3: Fine-tune the BERT model on the QA task
 - For each batch in the training dataset:
 - Forward pass: Compute model predictions for start and end answer positions
 - Compute loss between predictions and true answer positions
 - Backward pass: Update model weights based on loss
- 4: Evaluate the model on a validation set to check performance
- 5: For inference, input a question and context to the model
 - Model outputs start and end positions of the predicted answer span
- 6: Decode the predicted span to text as the final answer

Outputs:

Fine-tuned BERT model capable of answering questions based on context

Libraries Used:

Libraries	Preprocessing	Visualizations	Model and Evaluation
string	json	Seaborn	F1-score
re	Transformers	Matplotlib	Exact Matching
defaultdict	-	WordCloud	BertTokenizer
collections	-	IPython.display	BertForQuestionAnswering
-	-	-	Torch

The above libraries are used in categorised manner, these have been played a very important role to complete each of the steps we implemented, these all have a good amount of options along with a clean visualization which eases to collect and do an analysis over the dataset. Each library play an essential roles accordingly the table provides clear differentiation between how these libraries are categorized and are needed to achieve the end goal.

Integration of NLP Techniques:

The packaged components incorporate complex NLP methods taking advantage of rich BERT features, such as trained NLP methods for tokenisation and attention mechanisms. For the latter, BERT's highly effective tokeniser performs tokenization of the input text, capitalising on 'attention around the token', and takes care to observe the context surrounding each token - a key aspect of human language usage. A second set of attention mechanisms is valued by their ability to assign 'attention masks' during model training to ensure that the model 'pays attention to the right parts of the input'. This allows a flexible sequence length, permitting the model to handle a large array of question types and contexts.

Also, the project specifically exploits transfer learning, by taking a model that has been pretrained on large amounts of data and using it, as a starting point, to perform a task, here question-answering on SQuAD, which builds on it. In effect, the system is being made to leverage what it has learnt linguistically elsewhere, which puts it in a position to generalise to questions and domains it hasn't trained on, leading to a model that performs well. Moreover, it's a model that is robust, accurate and does exactly what it is prompted to do, and that does so with examples of authority and nuance that no human could match. From these examples, we can appreciate the degree to which different NLP techniques come into play in progressively building useful, robust and effective models, which instrumental reason may guide but can never anticipate. In the case of BERT, for example, it may never learn to speak the way that humans do (no one wants that!) but it can learn to understand semantics and pragmatic linguistic nuances in ways that modelling can only ever hope to approximate..

Results:

Performance:

```
Evaluation results: {'exact_match': 0.231, 'f1': 0.3407089991930131}
```

We had selected to use Exact Match (EM) and the F1 score metrics because our project is regarding Q&A system, Exact Match (EM) is a very strict measure; it describes what percent of the model's predictions matched the correct answers completely, which in this case was 23.1 per cent. The F1 score is the harmonic mean of precision and recall and measures how accurate the predicted answers were (even in part); it was 34.07 per cent.

Interpretation of Performance Measures

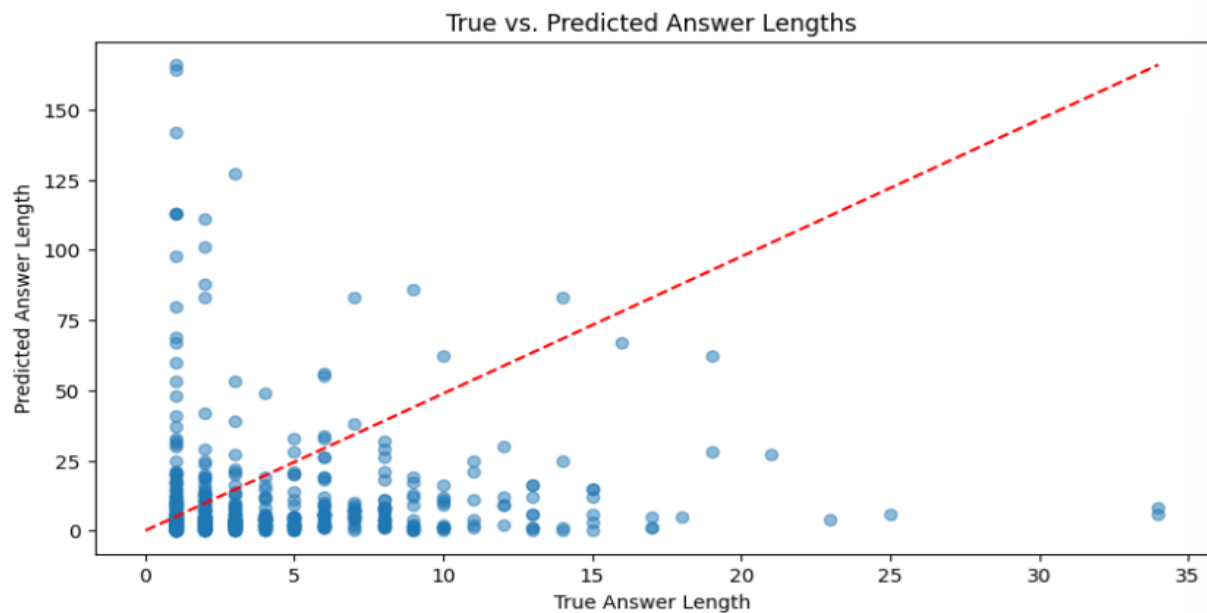
Within the context of our model and annotation scheme, the reasoning behind why Exact Match might not be a holistically reliable metric of success is less ambiguous. As such, the Exact Match score may be a useful measure of modelling outcome but it may fail to capture a model's effectiveness in real-world scenarios. This is due, in part, because the very form of text and the generation of questions that make up the SQuAD dataset are captivatingly heterogeneous and complex. For instance, a question formulation may have been paraphrased in different ways from which valid answers may exist that do not match the ground truth verbatim but do preserve the truth about the paragraph nonetheless.

On the other hand, we base it on context understanding, which inherently adds noise into the answers. Moreover, the dataset contains questions with answers which are semantically correct but different from the annotated answers in the dataset. This is especially visible in QA datasets such as SQuAD, where the answers are extracted text spans from the context, and the ground truth is not necessarily exhaustive for all the possible correct answers.

As such, F1 therefore becomes a far more meaningful measure, as it takes account of imprecise matches, and is a more commonsense metric that recognises the nature of ethnic and linguistic diversity and the richness of expression. F1 is a measure of the balance between precision - a model's ability to return only relevant answers - and recall - its ability to return all relevant answers. The fact that our model scored much higher in F1 than in Exact Match means it is capable of returning relevant answers even if they are not an exact match of the ground truth.

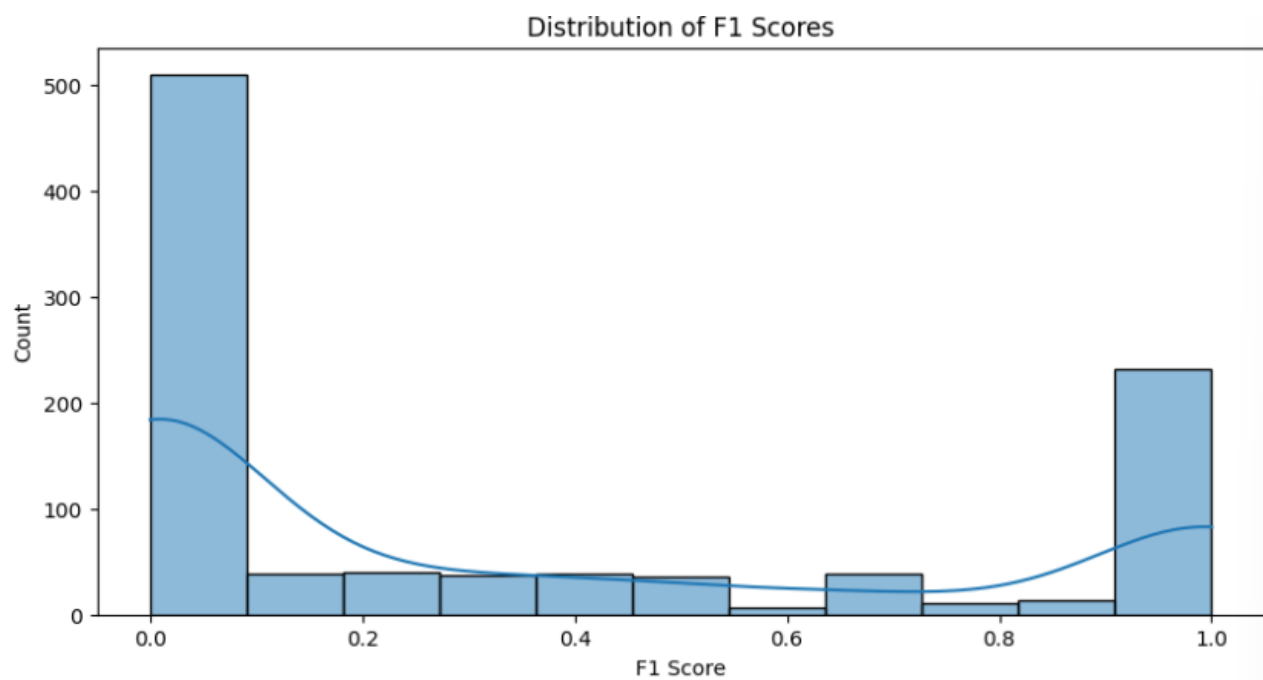
But in the real world, capabilities such as being able to understand disparate contexts, or being able to respond appropriately to different expressions, will matter more than 'verbatim matching'. So while the Exact Match can say something about the precision of the model, this other metric (the F1 score) will really paint something like a 'balanced' picture of how well the model is actually performing in practice, and keeping us positive on the results.

Visualizations:



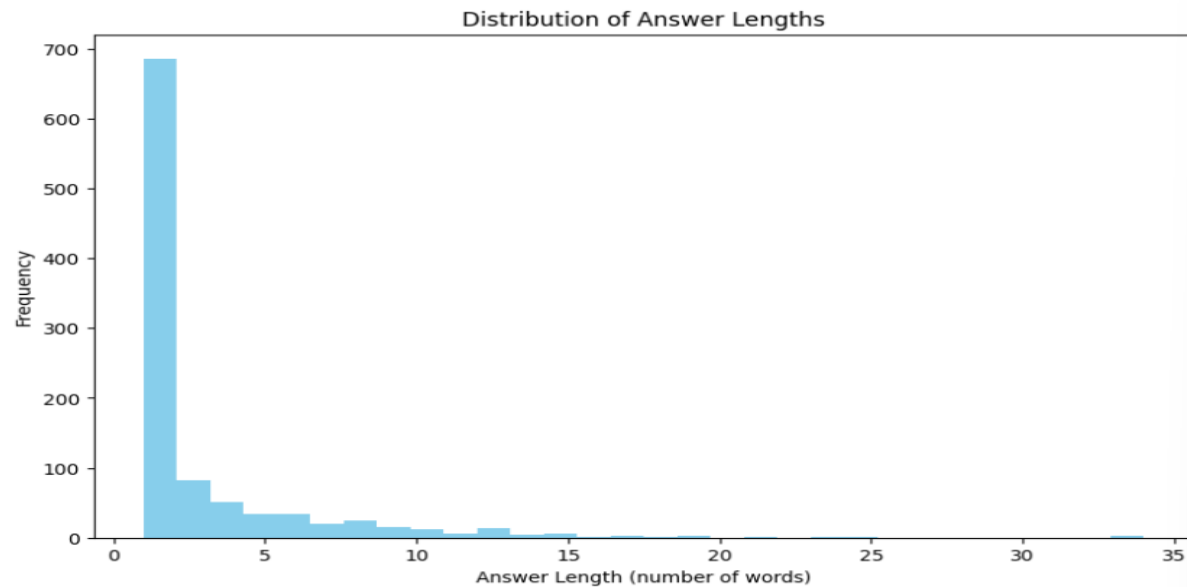
Explanation: The above scatter plot compares the length, which is an estimate of the truth, and the length that was predicted by the model, all extracted from the data set. In this case each point is a couple of predicted length and actual length from one instance, with red dashed line being a theoretical case in which the predicted length equals to length in the data set.

In comparison to modern results, our model's estimates of the length of the correct answers is better for short answers than for longer ones. This suggests shortcomings that need to be accounted for to achieve higher accuracy levels of leading models.

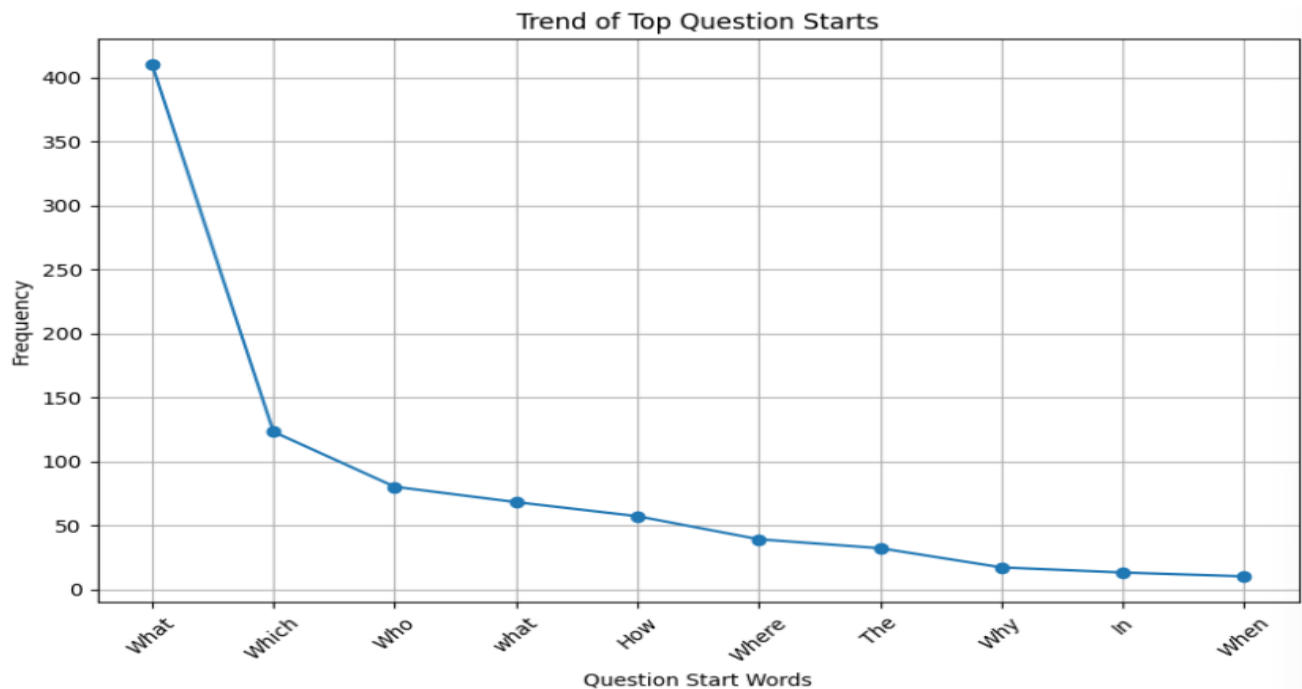


Explanation:

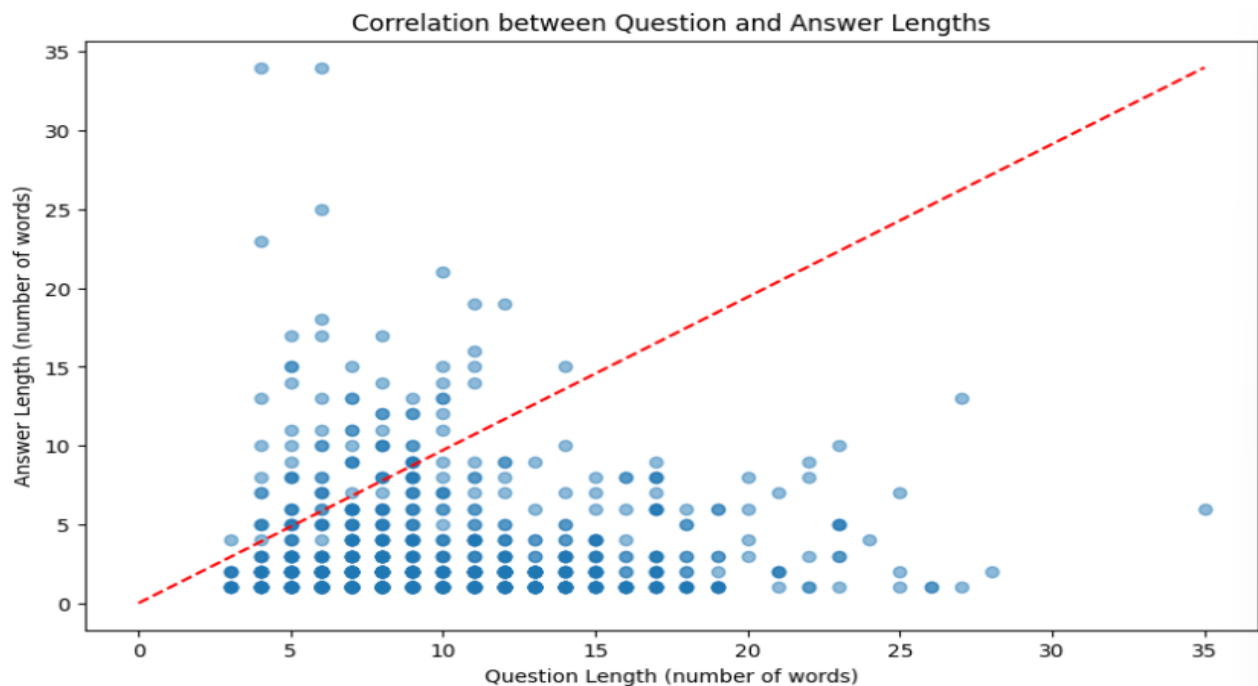
In the histogram, the numbers of F1 scores by our model are presented with the frequency of obtained scores in the specified range to assess overall model performance. Most of the longest bars are distributed in the decreasing ends of the score scale. As can be observed, our model performs especially well in some fields but opens the question of model improvement since consistency is not achieved.



Explanation: The histogram visualises how the answer length is distributed in our dataset. Clearly, a vast majority of answers are shorter. The sharp decline to zero after the threshold at 15 words indicates that short answers are much more frequent than the long ones. The skew towards brevity in our dataset might affect the model's performance. For instance, if state-of-the-art systems are trained on a dataset, the skew of its distribution for answer lengths could harm their performance.



Explanation: The above, line graph shows how often certain leading words in questions from our dataset resonated as the most frequent words (What being the most common). There is a steep fall off from that point, with other words such as How and When being much less frequent. It's possible or even likely that the steepness of this curve reflects either a narrower range of question types in our dataset or something much more specific than the wider datasets used in other models.



by their size, which allows you to discern how closely a model matches true data distributions, since prominence of terms between the two will be similar if the model performs well.

Interesting Results:

We have tried to imitate real world user ended type of questions to see models effectiveness in producing meaningful answers by rephrasing the questions to be confusing and the results still were pretty satisfying, some of the examples are added below:

Titles: 1: Teacher

Question: In which year study was found that teachers had double the stress?

Submit

Answer: 2012

Context: A 2000 study found that 42% of UK teachers experienced occupational stress, twice the figure for the average profession. A 2012 study found that teachers experienced double the rate of anxiety, depression, and stress than average workers.

Titles: 1: Teacher

Question: What percentage of teachers faced pressure when teaching?

Submit

Answer: 42 %

Context: A 2000 study found that 42% of UK teachers experienced occupational stress, twice the figure for the average profession. A 2012 study found that teachers experienced double the rate of anxiety, depression, and stress than average workers.

Titles: 3: Genghis_Khan

Question: Who was in-charge of the mongol empire when khagan was selected?

Submit

Answer: toregene khatun

Context: Modern Mongolian historians say that towards the end of his life, Genghis Khan attempted to create a civil state under the Great Yassa that would have established the legal equality of all individuals, including women. However, there is no evidence of this, or of the lifting of discriminatory policies towards sedentary peoples such as the Chinese. Women played a relatively important role in Mongol Empire and in family, for example Töregene Khatun was briefly in charge of the Mongol Empire when next male Khagan was being chosen. Modern scholars refer to the alleged policy of encouraging trade and communication as the Pax Mongolica (Mongol Peace).

Titles: 3: Genghis_Khan

Question: does mangolia have a river, if so name it?

Submit

Answer: onon river

Context: Years before his death, Genghis Khan asked to be buried without markings, according to the customs of his tribe. After he died, his body was returned to Mongolia and presumably to his birthplace in Khentii Aimag, where many assume he is buried somewhere close to the Onon River and the Burkhan Khaldun mountain (part of the Kentii mountain range). According to legend, the funeral escort killed anyone and anything across their path to conceal where he was finally buried. The Genghis Khan Mausoleum, constructed many years after his death, is his memorial, but not his burial site.

Titles: 18: Kenya ▼

Question: Who is the president of kenyan people?

Submit

Answer: kenyatta

Context: Published comments on Kenya's Capital FM website by Liu Guangyuan, China's ambassador to Kenya, at the time of President Kenyatta's 2013 trip to Beijing, said, "Chinese investment in Kenya ... reached \$474 million, representing Kenya's largest source of foreign direct investment, and ... bilateral trade ... reached \$2.84 billion" in 2012. Kenyatta was "[a]ccompanied by 60 Kenyan business people [and hoped to] ... gain support from China for a planned \$2.5 billion railway from the southern Kenyan port of Mombasa to neighboring Uganda, as well as a nearly \$1.8 billion dam", according to a statement from the president's office also at the time of the trip. Base Titanium, a subsidiary of Base resources of Australia, shipped its first major consignment of minerals to China. About 25,000 tonnes of ilmenite was flagged off the Kenyan coastal town of Kilifi. The first shipment was expected to earn Kenya about Shs15 – Shs20 Billion in earnings. China has been causing environmental and social problems that include the recent suspension of the railway project.

We can observe that however the question is phrased the model tries to give accurate answers by understanding the language of the input question promptly.

Project Management

Completed Work:

The final results show a major step forward in Natural Language Processing and the development of advanced question-answering systems, and uses the state-of-the-art BERT neural network architecture, which has been fine-tuned to handle deep questions while identifying relevant information from the wide variety of topics presented (all from the SQuAD). The final product is the culmination of each individual group member's specialities lining up into a straight pathway towards a successful endpoint.

We have successfully completed the following tasks to complete the project:

- ✓ BERT Model Integration
- ✓ Data Pre-processing Pipeline
- ✓ Model Training and Evaluation
- ✓ User Interface Development

Description:

The goal of that project was to create a useful system that would understand human-looking question for answering to solve problems on demand; developing a QA system that understands human-sounding text to that degree of accuracy required a series of dedicated project phases, starting with carefully prepared training data (in form of the SQuAD dataset), over strategically performing training and fine-tuning of the BERT model in the next step, to the last stage of incorporating a user-friendly interface allowing easy access and interaction with the technology - all of those phases grounded on testing and iterative improvement to achieve robustness and reliability.

Issues faced:

It was not without its challenges: BERT is too big of a model for regular pattern-matching techniques. The team had to come up with stronger regularisation methods for the training process and tune the parameters of their model hyperparameters. The SQuAD dataset also had to be cleaned and normalised extensively to resolve inconsistent and ambiguous cues before it was released, since the criteria for fitting the two corpora were not quite the same in the first place. Taming BERT in terms of computational power was yet another logistical challenge - BERT is a resource-intensive model to train, and the team had to design ways to reduce the time it takes to run the model and allocate more sandboxes to other users. Finally, how to design the interface was an extremely complex challenge to ensure a consistent Q&A system.

1. Data Quality
2. Model Overfitting
3. Computational Resources
4. User Experience

Every one of these challenges demanded sustained deliberation and creative problem solving, and a commitment to the project going forward to ensure collective expertise and regular, sustained thinking to address the challenges, to not only get the job done but also to build capacity to inform the next.

References:

- [1] <https://adversarialqa.github.io/>
- [2] <https://github.com/ad-freiburg/large-qa-datasets>
- [3] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.
- [4] Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.
- [5] Peters, M.E., Ruder, S., & Smith, N.A. (2019). To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. ArXiv, abs/1903.05987.
- [6] Wynter, A.D., & Perry, D.J. (2020). Optimal Subarchitecture Extraction For BERT. ArXiv, abs/2010.10499.
- [7] Wang, W., Bi, B., Yan, M., Wu, C., Bao, Z., Peng, L., & Si, L. (2019). StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding. ArXiv, abs/1908.04577.

[8] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. Conference on Empirical Methods in Natural Language Processing.

[9] Liu, X., Cheng, H., He, P., Chen, W., Wang, Y., Poon, H., & Gao, J. (2020). Adversarial Training for Large Neural Language Models. ArXiv, abs/2004.08994.