Question Answering Model
DS5220

Praveen Kumar Sridhar, Kapil Sahu, Meghaana Tummapudi

15 March 2022

1 Abstract

Our goal is to create a question-answering model. The data set we use is the Stanford Question Answering Data Set (SQuAD), specifically SQuAD1.1. We intend to take two approaches to this problem: building an information retrieval model and experimenting with BERT model on the data set. We also intend to compare how the model works on the unanswerable questions from SQuAD2.0 data set.

2 Introduction

We will be using the SQuAD data set consisting of two parts: The SQuAD1.1 data set consists of 100,000+ question-answer pairs on 500+ articles where as the SQuAD2.0 data set combines the SQuAD1.1 questions with over 50,000 unanswerable questions written adversely by crowd workers to look similar to answerable ones.

Attribute Information:
1. id: An index
2. title: The title of the prose or the context paragraph
3. context: The prose for the reading comprehension
4. question: The question we need to find the answer for from the context paragraph
5. answers: A dictionary that contains the answer text and the answer start index
6. is_impossible: A flag to indicate if the question can be answered.

What is the problem?
Reading Comprehension (RC), or the ability to read text and then answer questions about it, is a challenging task for machines, requiring both understanding of natural language and knowledge about the world. (Refer figure 4.1)

Why do you think this problem is important?
Question answering has emerged as a fertile ground for end-to-end deep learning models and natural language processing technology advancements. It's an important problem since it has a great potential in the applications that require information synthesis and retrieval due to the extra complexity in interpreting text to generate meaningful conclusions.

What do you want to do?
We will be using two techniques which we have explained in later sections: 1.Information Retrieval(IR) similarity matching. 2.Machine learning (Implementing BERT)

3 Future Work

What do you anticipate?
We expect to face the following challenges with respect to our objective :

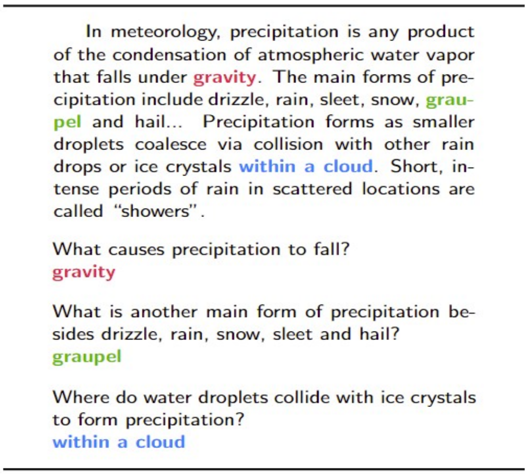
1. Ambiguity : The same phrase having different meanings; this can be structural and syntactic (like "flying planes") or lexical and semantic (like "bank").
2. Training task will need significantly high computational resources for which we will be using the Discovery Cluster.
3. Using methods like Word2Vec, GloVe, etc. for creating custom embeddings is a time consuming process.

What methods are you going to use?
The first thing we need to do is pre-process the data which involves cleaning the "context" and the "questions". Furthermore, we also intend to extract the whole sentences from the context paragraph where the answer phrases appear. Following pre-processing, we must begin modeling, and, as previously said, we will attempt to address this problem using two very different approaches. The first approach has its origin in information retrieval. So, for this, we are going to first convert the context and the question into vectors. To accomplish this, we'll employ a variety of embedding models, including GloVe and also custom-trained embedding models such as word2vec and Fastext. Here, we'll use the "context" to train the embedding model. After converting the context and questions to vectors, the next step is to use cosine similarity to find the sentence in the context that is the most similar to the question which will be the answer.

$$cos(sent_i, question) = \frac{sent_i.question}{|sent_i||question|}$$

The second approach has its origins in machine learning. Here we are going to attempt to solve this problem using BERT (Bidirectional Encoder Representations from Transformers). BERT is a masked language model (MLM). MLM enables/enforces bidirectional learning from text by masking (hiding) a word in a sentence and forcing BERT to bidirectionally use the words on either side of the covered word to predict the masked word. We first intend to use the bert-base-uncased pretrained model, observe its performance, and then intend to fine tune it with the context and questions from our training data set.

4 Figures



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 4.1 : Question-answer pairs for a sample passage in the SQuAD data set. Each of the answers is a segment of text from the passage.

5 References

1. https://arxiv.org/pdf/1606.05250.pdf
2. https://arxiv.org/pdf/1806.03822.pdf
3. https://rajpurkar.github.io/SQuAD-explorer
4. https://huggingface.co/blog/bert-101
5. https://huggingface.co/datasets/squad

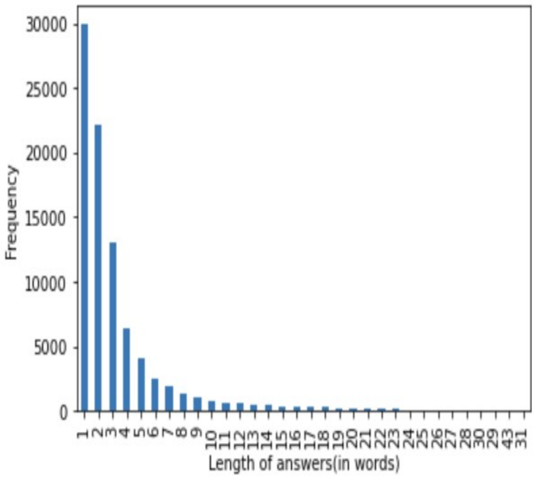


Figure 4.2 : Frequency plot of Length of answers(in words) shows that majority of answers consists of less than 10 words.