Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-021-05810-5

RESEARCH ARTICLE - SPECIAL ISSUE - AI BASED HEALTH-RELATED COMPUTING FOR COVID-19 (AIHRC)

COBERT: COVID-19 Question Answering System Using BERT

Jafar A. Alzubi[1] · Rachna Jain[2] · Anubhav Singh[2] · Pritee Parwekar[3] · Meenu Gupta[4]

Received: 7 October 2020 / Accepted: 2 June 2021

Abstract

In the current situation of worldwide pandemic COVID-19, which has infected 62.5 Million people and caused nearly 1.46 Million deaths worldwide as of Nov 2020. The profoundly powerful and quickly advancing circumstance with COVID-19 has made it hard to get precise, on-request latest data with respect to the virus. Especially, the frontline workers of the battle medical services experts, policymakers, clinical scientists, and so on will require expert specific methods to stay aware of this literature for getting scientific knowledge of the latest research findings. The risks are most certainly not trivial, as decisions made on fallacious, answers may endanger trust or general well being and security of the public. But, with thousands of research papers being dispensed on the topic, making it more difficult to keep track of the latest research. Taking these challenges into account we have proposed COBERT: a retriever-reader dual algorithmic system that answers the complex queries by searching a document of 59K corona virus-related literature made accessible through the Coronavirus Open Research Dataset Challenge (CORD-19). The retriever is composed of a TF-IDF vectorizer capturing the top 500 documents with optimal scores. The reader which is pre-trained Bidirectional Encoder Representations from Transformers (BERT) on SQuAD 1.1 dev dataset built on top of the HuggingFace BERT transformers, refines the sentences from the filtered documents, which are then passed into ranker which compares the logits scores to produce a short answer, title of the paper and source article of extraction. The proposed DistilBERT version has outperformed previous pre-trained models obtaining an Exact Match(EM)/F1 score of 80.6/87.3 respectively.

Keywords COVID-19 · CDQA · Question answering · TF-IDF · Cosine-similarity · BERT · DistilBERT · SQuAD · HuggingFace · CORD-19

1 Introduction

COVID-19 infections are characterized as a disease caused by a novel virus called Coronavirus now called extreme acute respiratory condition (SARS-CoV-2; previously known as 2019-nCoV) [1,2]. The worldwide reaction to COVID-19 has yielded a fast development of scientific publications - expanding at a pace of thousands every week - about COVID-19, SARS-CoV-2, different Covids, and related themes [3]. The people on the cutting edges of the battle - medical services experts, strategy makers, clinical scientists, and so on - will require expert specific methods to stay aware of this literature for getting proper knowledge of the latest research findings [4]. COVID-19 particular question-answering abilities are vital to making this abundance of information both helpful and significant. The risks are most certainly not trivial, as choices made on returned, fallacious, or counterfeit answers may endanger trust or general well being and secu-

✉ Jafar A. Alzubi
j.zubi@bau.edu.jo

Rachna Jain
rachna.jain@bharatividyapeeth.edu

Anubhav Singh
anubhav.singh2359@gmail.com

Pritee Parwekar
priteep@srmist.edu.in

Meenu Gupta
gupta.meenu5@gmail.com

[1] Al-Balqa Applied University, Salt, Jordan

[2] Bharati Vidyapeeth's College of Engineering, New Delhi, India

[3] SRM Institute of Science and Technology, NCR Campus, Ghaziabad, India

[4] Chandigarh University, Ajitgarh, Punjab, India

Published online: 23 June 2021

rity of the public. The author has recognized these dangers, however, accept the fact that our research technique benefits the more extensive COVID-19 research thus exceeding such dangers [5]. Taking these challenges into thought, our COBERT search system has the option to enormously help the endeavors of scientists to viably help to battle the current pandemic. The proposed system is a retriever-reader dual algorithmic approach that takes queries from the user as input and generates the most accurate responses consisting of a single line answer to the query, title of the literature, and a whole paragraph from the scientific literature [6–8]. The COBERT system searches a document of 59K corona virus-related literature made accessible through the Coronavirus Open Research Dataset Challenge (CORD-19) [9]. As of late, QA models have made huge advancements as far as both execution and throughput [10–12]. Such upgrades can be credited to the presentation of huge scope QA datasets and deep learning models, and the recent focus of the research towards the proficiency of such models. Lately, most of the QA models are extractors and re-rankers since their attention is on a generally small-sized dataset of text such as articles, sentences, or on the other hand entries, and so on [13,14]. Whereas, our framework works straightforwardly on a huge corpus of literature. Our proposed model coordinates best practices from the information retrieval with BERT to create a framework focusing on an end-to-end closed domain question answering system and analysis on a standard benchmark showing improvement over the past work. Our model fine-tunes pre-trained versions of BERT with SQaUAD adequately accomplishing high scores in recognizing answers [15]

The lucidity of this configuration is one significant element of its architecture. Additionally, our strategy is explicitly customized for BERT-based models to use this architecture in the enormous scope of pre-prepared language modeling. [16–18] Our proposed model exhibits that the pretrained model diminishes the requirement for some intensely designed task-specific models. BERT is the first fine tuning based model that accomplishes state-of-the-art results on a huge suite of sentence-level and token-level tasks, surpassing many task explicit models. Our proposed model holds the following components:

Firstly, retrieval is performed throughout the whole corpus which is divided into paragraphs thus creating features based on tf-idf focusing on bigrams and unigrams. The embedding is then used to calculate the cosine similarity with the query and sequential comparison scores are obtained which are then used to Retrieve the top 500 documents with the optimal scores. Then, comparisons are performed batch-wise to get the document that is most probable to contain the answer.

The reader then does the extraction of the documents by further splitting them into sentences and these sentences are then Fine-tuned with Bidirectional Encoder Representations from Transformers (BERT) for refining the automatically generated answers based on the similarity between the query. We have also used DistilBERT (i.e., Distilled-BERT) for the task of fine-tuning.

Then comes our Ranker which ranks the most accurate answers using weighted scores threshold score obtained from the Retrieval and Reader. Ranker compares the scores of each candidate answer from each paragraph and compares the logits of the best answer for each paragraph. Then the best answer from that document is produced as output to the user.

This paper is organized in the 9 folds. Firstly, we will discuss the literature review thus analyzing the various past works in the domain in Sect. 2. Section 3 then focuses on our problem description briefly. Then, Sect. 4 explains our proposed methodology where each component of our architecture is described. In Sect. 5 we have explained our Dataset, Preprocessing, Training,etc. Section 6 reported the results and analysis of the comparison of various techniques and our model. Whereas, in Sect. 7 we described some Limitations of the study .We have concluded in 8 and discussed Future Works in Sect. 9.

2 Literature Review

Kricka L.J. et al. [8] concentrated towards shedding light on how AI is reshaping medical field and helping in COVID-19's diagnostic research through three prominent initiatives: COVID-19 focused datasets (e.g., CORD-19); AI-Boosted inspection means (e.g., WellAI, SciSight); and get in touch with tracing supported mobile communication technology. WellAI uses NLP neural networks to learn from CORD-19 so as to summarize the existing knowledge through unsupervised learning. SciSight is an AI-powered visualizing tool for scrutinizing associations between concepts appearing with CORD-19/ Contact Tracing is an extensive auditing process for clashing of a communicable disease through 3 steps: Identifying, Listing and Follow-up. Apps such as Aarogya Setu were released in Indian government to invigilate erstwhile COVID Survivors and mapping them to hatch contamination plans accordingly.

Broth Andreas et al. [10] proposed a Question- Answering (Q/A) system over Knowledge Bases(KB), which overwhelms impediments which were endured—namely, porting KBs language-agonistic systems(Eg. SemGraph QA), structural gap (Eg. AskNow, DEANNA), Scalability (SINA) . This technique is established on the perception that a lot of questions are often presumed from the connotations of the words within the question while the syntax of the question has subordinated consequence. A loss of credit that authors admit of their implementation is that the identification of relations relies on a dictionary. They have applied their exemplary algorithm and adapted a group of existing services in order that end-users can query, using multiple languages, multiple KBs at an equivalent time, employing a unified interface .

Yu Wenhao et al. [11] developed a TransTQA, which is a novel system that gives automatic responses by retrieving proper answers supported correctly answered similar questions within the past. TransTQA is configured with the ALBERT model consistent with the required arguments that defines the model architecture, producing transformer encoders. MLM technique is employed for fine tuning source corpus. The author has developed TransTQA, which may be a novel system that gives automatic response by retrieving proper answers supported correctly answered similar questions. TransTQA is made upon a Siamese ALBERT network, which enables it to reply to questions quickly and accurately. Furthermore, TransTQA adopted transfer learning to enhance its performance on multiple tech domain QA.

Abacha Ben Asma et al. [12] presented a Medical Visual Question Answering task (VQA-Med) focusing primarily on 4 categories of clinical questions : Modality, Plane, Organ System, and Abnormalities. The first three tasks are classified as : classification task and the fourth can be put with answer generation. All of these questions are solved through images only, and do not require any additional domain based context for generating answers, and were evaluated on Accuracy and BLEU score.

Nogueira Rodrigo et al. [13] re-implemented the query based passage re-ranking. Trained on state of the art TREC-CAR dataset, it beats out the previous top entry by 27% at MS MARCO passage retrieval task. The researcher's pipeline can be divided into 3 stages: first being retrieving all conceivable documents pertinent to the given query, using archetype operations such as BM25. The second stage implicates passage re-ranking by computationally-accelerated technique. The final stage being top ten or fifty of these documents being responses to the user's question. Using the BERT LARGE Model as binary classification model and [CLS] vector as input to a single-layer neural network to obtain the probability of being germane.

Gao Luyu et al. [14], had studied if and the way of knowledge for search within BERT itself, can be re stationed to a smaller ranker through distillation, producing up to ninefold speedup, while perpetuating state-of the art performance. The facility of a BERT ranker is from two main origins : 1) general purpose language modeling knowledge learned in pre-training, and 2) search-specific relevance modeling knowledge learned in fine-tuning.The researcher minutiates three enunciated methods that coalesce fine-tuning and distillation to reach at a smaller rank from an incipiently full-sized BERT model: 1. Ranker Distill: distillation is used for search knowledge, 2. LM Distill + Fine-tuning: distillation is serviced for LM knowledge, and 3. LM Distill + Ranker Distill: distillation is employed for both LM and search knowledge. They found a higher degree of compression herald more predicaments in model optimization due to loss of knowledge.

Esteva et al. [19] implemented a retriever-ranker semantic search engine devised to handle complex queries regarding COVID-19. The retriever is composed of a Siamese-BERT (SBERT) as an encoder, with a TF-IDF vectorizer and reciprocal-rank fused with a BM25 vectorizer, as a ranking function to estimate the relevance of documents in a given search query. A query is given for which each document is provided with a retrieval score. The ranker comprises a multi-hop question- answering (QA) module that, alongside a multi-paragraph abstractive summarizer, adjusts retriever scores. The retrieved set of documents are run through a question answering module and an abstractive summarizer in the ranking. Based on this ranking, the documents are ranked by a weighted combination of their answer match, retrieval scores, and summarization match.

In [20] Abacha et al. approached one of the challenges in the large-scale information retrieval (IR) process that is developing accurate and domain-specific ways to answer natural language questions, especially when answering questions related to the medical domain. Dedicated Question Answering (QA) systems do not rely on any external information about the users, and hence, it makes them one of the viable solution approaches. Thus, the Recognizing Question Entailment (RQE) based QA system was developed to answer new medical questions using the existing answers to previously asked questions. A collection of 47K medical question answer-pairs was built and shared. This approach works well for both open-domain and specific-domain. Regardless, deep learning models achieved state-of-the-art results on an open-domain and clinical dataset, which showed low performance on consumer health questions.

Tang et al. [21] have created a dataset named COVID QA, containing 124 pairs of questions-article related to COVID-19. The architecture begins with keyword-based re-retrieval to determine a set of pertinent candidate documents that are then re-ranked by machine-learned models to assign higher relevancy documents to higher ranks. In the final stage, the model would input the query and the document as input and determine the most relevant passages. However, with limited data entries in the CovidQA dataset, it is impossible to train the QA model in a supervised manner. The dataset also lacks "no answer" examples, which is an unrealistic conjecture in real life.

In this paper, Lee et al. [22] proposed a system comprised of three folds: first, COVIDASK, a real-time QA on COVID-19; second, incorporation of traditional biomedical text mining tools into QA model to enhance usability (e.g., NEL); third, evaluation of COVIDASK on COVID-19 dataset created by the authors. Although COVIDASK is not an open-domain QA model, it has been made with many techniques from open-domain QA models since it has to handle a large amount of text. COVIDASK mainly focuses on reducing the latency in putting a query and receiving an answer. Latency can be reduced by decreasing the number of documents retrieved. Alternatively, all appropriate answer phrases are pre-indexed into dense and sparse vectors. Then, Maximum Inner Product Search is performed between query vectors and the phrase vectors to reduce the number of times a model passes through a document to only one.

The architecture proposed for the question-answering (QA) module, CAiRE-COVID by Su et al. [23], consists of three different modules: 1) Document Retriever: It preprocesses the user's query the most relevant n number of publications. It paraphrases long, complex queries to simple queries that are easier to comprehend by the system. These queries are run through the IR module, which returns paragraphs from the highest matching score. 2) Relevant Snipper Selector: The most relevant parts of paragraphs are highlighted and re-ranked based on the scores. 3) Multi-Document Summarizer: It returns an abstractive summary by summarising the top relevant paragraphs. This QA module is the perfect example of the extractor and re-ranked since their attention is on a small-sized dataset such as sentences or articles.

Many such models have been proposed to solve the problem of question-answering from a large corpus many of such have been focused on medical-question answering but the final capturing capabilities of models are proving ineffective when comes to a large corpus whereas our proposed model phase by phase increases the answering capability thus increasing the chances of the final answer being closer to ground answer as our model employs the retriever-reader dual algorithmic system. Whereas, coming to the deep learning models their performance is limited when capturing the embedding representations of question-answer pairs which generalize the model with one neural network which captures one-side features. Therefore it's necessary for the model to capture complex all-side features of QA-pair. Existing approaches deploys the heuristics on coming up with the possible answer whereas the proposed model is assessing its answering capabilities against the ground truth sentence in the ranker phase. The proposed architectures thus try to deal with such shortcomings which are discussed in the further sections.

3 Problem Description

In COVID-19 pandemic situation, many research analyses have been published in the form of a research paper, which is open access for all to do further analysis. Simultaneously, there were many solicitations that came from social orders (i.e., medical research community and Broader Society) for finding the response to the questions identified with Covid-19. Finding such sorts of answers is quite difficult with any system or can say no such system is capable of answering such kinds of questions. During the current situation of pandemic emergency, thousands of articles, blogs are being published, just some of which are empirical, factual, and peer-inspected. This may prompt the consideration of deceptive and the potential quick spread of scientifically reputable information or in any case fallacious exploration and information. Individuals on the forefronts - clinical Specialists, workers, and so on - are time-obliged in their capacity to parse this huge corpus, which could hinder their capacity to move toward the returned query items with the fitting degrees of distrust and queries accessible in less critical incidents [23]. To address this issue, we propose an end-to-end closed domain question answering system on COVID-19, which is present on the top of the HuggingFace BERT [24] transformers library. Since the system is an end-to-end cdQA system [17] that helps in answering a question related to the COVID-19 situation, mainly the questions proposed in the CORD-19 Kaggle. This proposed system based upon query, the user system gives the user three outputs, each having three components: the answer, Title of the paper, and Text described in the paper.

Datasest Link: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

4 Methodology

4.1 Closed Domain Question Answer (CDQA) Search

The CDQA based COVID-19 Search engine implemented here is based on a retriever-reader dual algorithmic approach, as shown in Fig. 1. The COBERT is model inspired by the DrQA Open Domain Question Answer model, developed by Chen et al. [25]. The main challenge with Question and Answering as a Natural Language Understanding task is that QA models often fail to perform when asked to produce an answer for a question from a large input text. To address the above-discussed challenge, the COBERT model was broken into two steps. Firstly, narrow down the input text to the top articles where the answer might be present (The Retriever) using search (e.g., TF-IDF, BM25), and secondly, it out of the narrowed down input text, find the best potential answer (the reader) using a QA model [4]. For this reason, ODQA and its CDQA are considered as an approach to providing ”Machine Reading at Scale.

In other words, it can conclude that a closed-domain system [26] deals with questions under an exact domain, i.e., medicine or automotive maintenance. Table 1 shows a comparison of the ODQA and CDQA model. The domain-specific knowledge achieved by using this model fitted to a

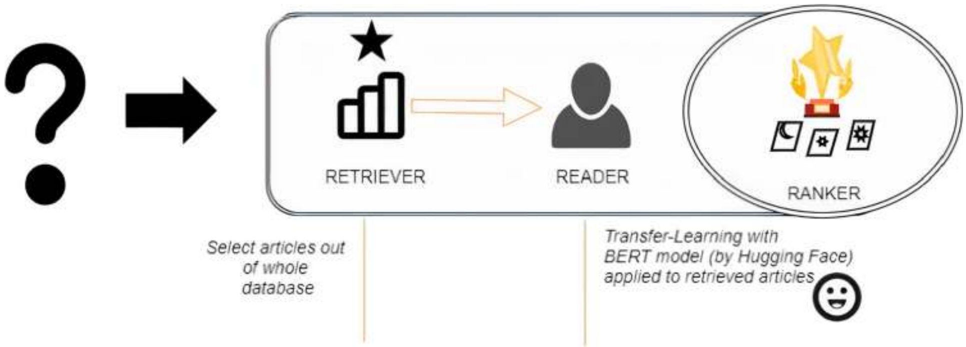

RETRIEVER
READER
RANKER
Select articles out of whole database
Transfer-Learning with BERT model (by Hugging Face) applied to retrieved articles

Fig. 1 COBERT pipeline Architecture

Table 1 Comparison Between Open-Domain QA an Closed Domain QA

Open-Domain QA	Closed Domain QA
Ability to answer about anything	Able to answer question regarding the specific domain.
Mainly rely on general ontology's and world knowledge	Able to exploit knowledge mainly domain-specific
Eg: DrQA(Facebook Research)	Example: cdQA

unique-domain database. The cdQA-suite built to facilitate the one who wants to model a closed-domain QA system.

4.2 Retriever

4.3 Architecture

The COBERT pipeline architecture shown in Fig. 1 is made on the top of the Hugging Face transformers library. This system is classified into two different parts, such as Retriever and Reader, as shown in Figs. 2 and 3 respectively. Whenever a query is asked to the COBERT, as we can in Fig. 1 it is passed into the Retriever which selects k documents or articles from the corpus that are well on the way to contain the appropriate answer. Basically, it converts these documents into Tf-IDF vectorizer along with an input query and evaluates the cosine similarity of the document and input query.

After selecting the most probable document, it passes these documents into the Reader. The reader is a pre-trained BERT model on our dataset. Now, the reader again calculates the best answers from the input documents. Finally, Ranker list down these candidate answer according to the logits score and threshold, and the final output is produced at the end.

The COBERT Retriever Pipeline schema shown in Fig. 2. The retriever schema chooses a couple of documents from a database to answer those questions that are set to a system. The COBERT system is working the same as a retriever of DrQA [27], which creates TF-IDF features based on uni grams and bi-grams. Further, this schema calculates the cosine similarity in-between the question sentence and every document of the database. TF-IDF Retriever segregated into two terms TF (Term Frequency) and IDF (Inverse Document Frequency). The TF is defined as $tf(t, d)$, the least challenging decision is to exploit the term in the report using its raw count, i.e, the occasions that term t happens in record d shown in Equation 1 .

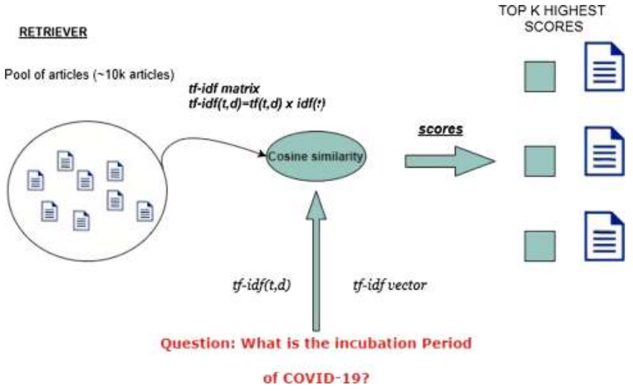

RETRIEVER
TOP K HIGHEST SCORES
Pool of articles (~10k articles)
tf-idf matrix
tf-idf(t,d)=tf(t,d) x idf(t)
Cosine similarity
scores
tf-idf(t,d)
tf-idf vector
Question: What is the incubation Period of COVID-19?

Fig. 2 COBERT Retriever Pipeline

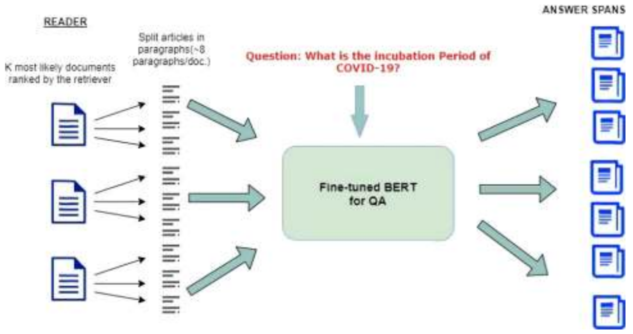

READER
ANSWER SPANS
Split articles in paragraphs(~8 paragraphs/doc.)
K most likely documents ranked by the retriever
Question: What is the incubation Period of COVID-19?
Fine-tuned BERT for QA

Fig. 3 COBERT Reader pipeline

We have represented ft,d as raw count whereas tf scheme is

$$tf = f_{t,d} \Big/ \sum_{t' \in d} f_{t',d} \tag{1}$$

The IDF, shown in Equation 2 is a proportion of the data the world gives, i.e., if it is normal or uncommon overall reports. It is the scaled logarithmically other division of the archives that comprise the word (acquired by separating, the all outnumber of reports by the number of records having the term, and afterward-logarithmic operation applied using the quotient):

$$idf(t, D) = \text{loglog}\left(\frac{N}{\{d \in D : t \in d\}}\right) \tag{2}$$

Where, N: total count of reports in the corpus.

It is consequently normal to modify the denominator to 1+ d D: t d—. Furthermore, —d D: t d— is the number of reports when the term t shows up (i.e., $tf(t, d)$ 0). If the term is not present inside the corpus, this will prompt a division-by-zero. Then tf-idf calculated as shown in Equation 3.

$$TF - IDF(t, d, D) = tf(t, d) * idf(t, d) \tag{3}$$

Further, the system will divide the document into paragraphs by selecting the document. Next, it provides direction to the document reader with the question, which is a pre-trained Deep Learning model. The proposed model used the BERT method (i.e., a Pytorch version of an NLP (Natural Language Processing) model) presented by HuggingFace. For our closed domain, we included the 59,000 articles as our corpus. As we have used the whole text of the paragraphs, this means we need to divide the data into chunks of 10,000 for pre-processing to avoid system crashes. The top articles identified based on a cosine similarity between the question string and the abstract text.

4.4 Reader

Figure 3 shows that the reader used the most likely document, which is ranked by the retriever. To answer the question, the reader splits the document into a paragraph. Further, in the reader, a final layer is present in the architecture that matches with the help of an internal score function and the most frequent one based upon the scores is the final output. The reader is based upon Bert Processor using the 'distilbert-base-uncased' model. Section 4.6 explains the Distil BERT transformer model.

4.5 BERT

BERT, stands for Bidirectional Encoder Representation from Transformer the most recent refinement of a series of neural models that make substantial use of pretraining, and has prompted noteworthy gains in numerous natural language processing tasks, going from a text classification to do tasks like question answering from the corpus.

BERT eases the unidirectional requirement by utilizing a "masked language model" (MLM) pre-training objective. The MLM model arbitrarily masks a portion of the tokens from the information, and the goal is to foresee the masked word dependent on its neighbors (left and right of the word). Unlike directional models, which scans the text info successively (left-to-right or option-to-left), the MLM objective empowers the representation to utilize both the left and the correct setting, which permits to pre-train a deep Bidirectional Transformer. BERT advanced for several NLP problems like QA. It was introduced in 2019, generating language. That is why it does not require decoding. Figure 4 shows the original architecture of QA tasks using BERT, where the input sequence consists of two parts: the question asked and a special [SEP] token and then followed by the context [17,24]. It learns the context of a solitary word from the two words, which encompass it from both directions using its capability of Bidirectional mechanism [28]. BERTBASE and BERTLARGE were the two distinct architectures introduced in the BERT model. BERTBASE has a hidden size of 768, having a transformer block having size 12 and self-attention heads of 12 sizes. In the end, we obtain 110 million parameters.

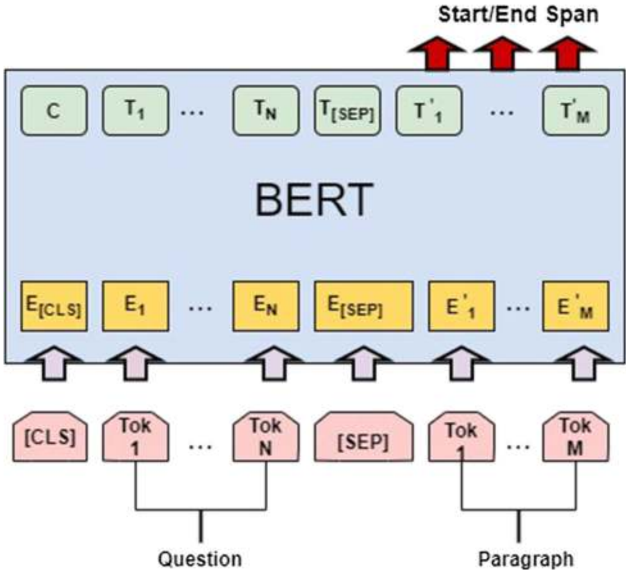


Fig. 4 BERT for question answering

Whereas we have 24 numbers layers, a hidden size as 1024, and 16 self-attention heads constitute around 340 million parameters in BERTLarge, making it a gigantic model.

4.6 Distil BERT

Distill* is a class of compressed models that started with DistilBERT (i.e., Distilled-BERT). It is a BERT based cheap, small, fast, and light Transformer model. Its performance is similar to BERT but runs faster by 60% but has 40% fewer parameters but still obtaining 97% performance if we compare it to BERT. These results were obtained using the GLUE language understanding benchmark. While using DistilBERT, it carries many similarities in comparison to the original BERT model. Thus, it (i.e., DistilBERT) is an interesting option to put a large-scaled trained Transformer model into production. The system has used a distilbert-BERT English language model, pre-trained on the same data used to pre-trained BERT (concatenation of the Toronto Book Corpus and full English Wikipedia) using distillation with the supervision of the Bert-base-uncased version of Bert. The model has 6 layers, 768 dimensions, and 12 heads, totalizing 66M parameters.

4.7 Ranker

Once the retriever and reader models are used, our solution presents the top 3 answers based on a weighted score between the retriever score (based on TF-IDF cosine similarity explained in Sect. 4.2) and reader score (based on DistilBERT QA Q-A pair probability). Cosine similarity ($\cos(\vec{q}, \vec{d})$), as shown in Equation 4, helps to find the similarity among two vectors of inner product space and conclude whether these vectors are indicating in the same direction. We frequently calculate the similarity of documents in tasks such as text analysis.

$$\cos(\vec{q}, \vec{d}) = \frac{\frac{\vec{q}}{|q|} * \vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|v|} \frac{\vec{q_i}}{\sqrt{\sum_{i=1}^{|v|}}} * d_i}{\sqrt{\sum_{i=1}^{|v|} * d_{i^2}}} \tag{4}$$

Where,
$[q_i]^{\text{th}}$ term tf-idf weight in query
$[d_i]^{\text{th}}$ term tf-idf weight in document
$|\vec{q}|$and$|\vec{d}|$ are the lengths of $\vec{q}$ and $\vec{d}$
$\frac{\vec{q}}{|\vec{q}|}$ and $\frac{\vec{d}}{|\vec{d}|}$ both are length-1 vectors which are normalized.

5 Experimental Setup

5.1 Dataset

COBERT system uses the dataset of CORD-19: COVID Open Research Data set collected by The White House, with the help of leading research groups like Allen Institute for AI and Kaggle [26]. The dataset consists of a collection of work of several Researchers and its analysis. It consists of around 59 thousand papers and around 41 thousand full texts [27] incorporating papers distributed in more than 3200 journals. Many texts are related to institutions situated in the United States (over 16 thousand papers) followed by the United Kingdom (over 3 thousand papers) and the European Union and then Asian countries. Chinese organizations have seen a brilliant ascent this year (over 5K papers) because of China's status as the principal focal point of the COVID-19 episode, thus, making the source of data very much diverse.

5.2 Pre-processing

The files stored in the JSON format and processed in the form of chunks of size 10,000 one by one and the final output was generated in the form of .csv format having attributes like 'title,' 'abstract,' 'paragraphs,' where 'paragraphs' specifies the complete text of the specific research paper. The data then translated using Google Translator into English, as our main aim is to provide a QA system in the English language. Now the paragraphs column further split into sentences to optimize the training of the model. We have used the DistilBERT model for the training. DistilBERT model trained on SQuAD 1.1 using Knowledge Distillation and Bert-base-uncased-whole-word-masking-finetuned-squad as a teacher. A separate copy is prepared using json format according to the SQuAD1.1 dataset for evaluation of the model.

5.3 Hardware and Training

We have used an Nvidia Tesla P100 GPU for training running on CUDA 9.2. 148 and cuDNN 7.4.1, having 16.4 GB RAM. The model trained using a pre-trained model of DistilBERT Tuning batches of size 400 in a chunk of 9000 rows at a time having a maximum sequence length of 384. Trained on GPU for 10,000 steps.

Table 2 shows the hyper parameters of both of the pre-trained weights namely BERT Base Uncased and Distil BERT with Knowledge Distillation fine tuned on CORD-19 dataset using Reader.

We have set the total batch size for training to be 128 and 256 in both the models. 'Max_query_length' are the size of the input tokens of the question, if found longer than this would be truncated. 'Doc_stride' is the size of stride when splitting of documents into chunks is performed.

Table 2 Hyperparameters Comparison of Model

Model	Bert-Base-Uncased	DistilBERT
max_seq_length	384	384
doc_stride	128	128
max_query_length	64	64
train_batch_size	128	256
max_answer_length	30	30
Learning Rate	1e-8	1e-8

'Max_seq_length' is the size of the input document sequence length which is set to 384. Whereas, 'max_answer_length' denotes the maximum size of the answer that can be generated which is set to 30. Adam weight decay was used as an optimization having a learning rate of 1e-8 finally. All the other hyper parameters are set to default

In addition, fine-tuning performed on the reader using an annotated dataset having the same format as the SQuAD dataset. The model needed two attributes, namely the 'title' and 'paragraphs.'

5.4 Prediction and Evaluation

A list of queries was prepared for the initial testing phase consisting of question like:

- "What is the optimal quarantine period for coronavirus COVID-19",
- "What is a range of incubation period for coronavirus SARS-CoV-2 COVID-19 in humans",
- "What is an effective quarantine period for coronavirus COVID-19"?

During each batch of retriever fitting, predictions of the queries simultaneously are done and stored in the form of a data frame. Later on, the answer with the best score used to produce the output. We have kept the retriever score as 0.35 for the predictions that presented the best weight on the development set of SQuAD 1.1-open.

Each of the question predictions were stored in a separate data frame have columns such as:
"answer","probability","start","end","qas_id","title", "paragraph","retriver_score","final_score".

For getting the final results the data frame is sorted upon "final_score" to obtain the 15 best predictions. For evaluating the model, it requires three steps.

Firstly, the panda's data frame converted into JSON format with SQuAD version 1.1 formats.

Then, using a cdQA-annotator, we prepared an annotated JSON format file to add ground truth question-answer pairs for evaluation of the model.

The third and final step is to evaluate both the JSON file prepared in the above steps to get the Exact Match (EM) and F1 score.

5.5 Metrics

To measure the accuracy in problems such as QA, EM (Exact Match), and F1 scores are regularly used. Exact Match: This metric is one of the most widely used metrics for evaluating the results of question-answering tasks. For each pair of question and answer if the words in the predicted answers are exactly matched with the ground truth answer then EM score 1 otherwise 0. It is straightforward as it sounds.

F1 score calculates the similarity to the ground truth. F1 is harmonic mean among recall and precision, as shown in Equation 5, 6, and 7, respectively. F1 score measured using the following criteria's:

True Positive (TP): This is the case where the positive class predicted correctly by the model.

True Negative (TN): In this, the negative class correctly predicted by the model.

False Negative (FN): Here, the model incorrectly predicts the negative class.

False Positive (FP): Using this, the model outputs the positive class as a negative class.

Here we will show the recall and precision formulas based on the notions above:

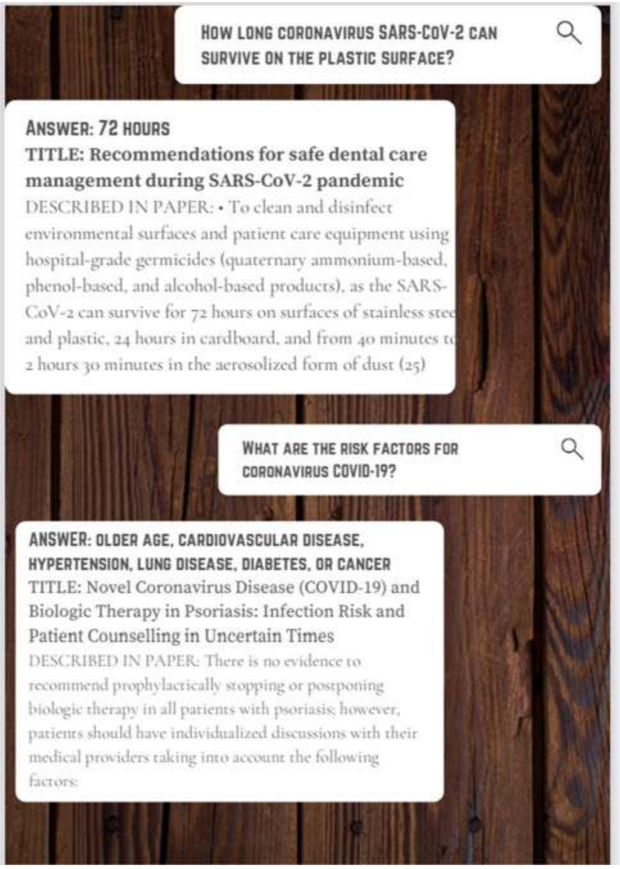
$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$F_1 = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

6 Results And Comparison Analysis

Figure 5 shows an example of the two queries when given as input to our model. The system outputs not only an answer but also the title of the document/article, the paragraph where the answer was found. In addition, every query model outputs three answers. For the sake of simplicity, we have presented only a single instance of the answer to the query. This can be set to multiple answers also. We observed that the model outputs more accurate answers in the query containing frequent terms related to coronavirus. Apart from score evaluation, through the use of Metrics described in Sect. 5.5, multiple such queries were checked on by multiple users for a better evaluation by the author. And both the results were satisfying according to the feature generation capabilities of the COBERT system. In the screenshot, we can see the answers



HOW LONG CORONAVIRUS SARS-COV-2 CAN SURVIVE ON THE PLASTIC SURFACE?

ANSWER: 72 HOURS
TITLE: Recommendations for safe dental care management during SARS-CoV-2 pandemic
DESCRIBED IN PAPER: • To clean and disinfect environmental surfaces and patient care equipment using hospital-grade germicides (quaternary ammonium-based, phenol-based, and alcohol-based products), as the SARS-CoV-2 can survive for 72 hours on surfaces of stainless stee and plastic, 24 hours in cardboard, and from 40 minutes t 2 hours 30 minutes in the aerosolized form of dust (25)

WHAT ARE THE RISK FACTORS FOR CORONAVIRUS COVID-19?

ANSWER: OLDER AGE, CARDIOVASCULAR DISEASE, HYPERTENSION, LUNG DISEASE, DIABETES, OR CANCER
TITLE: Novel Coronavirus Disease (COVID-19) and Biologic Therapy in Psoriasis: Infection Risk and Patient Counselling in Uncertain Times
DESCRIBED IN PAPER: There is no evidence to recommend prophylactically stopping or postponing biologic therapy in all patients with psoriasis; however, patients should have individualized discussions with their medical providers taking into account the following factors:

Fig. 5 Example of output of the COBERT system with the input query

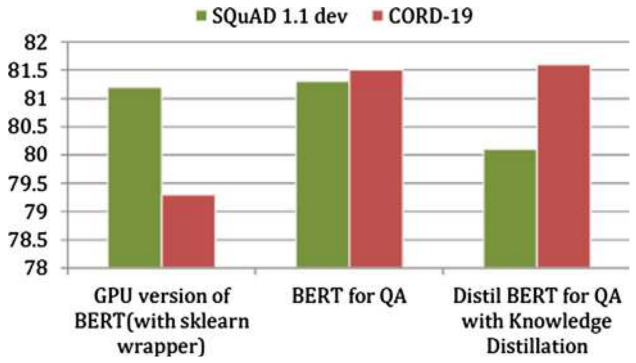


Fig. 6 Exact match (EM) comparison of models

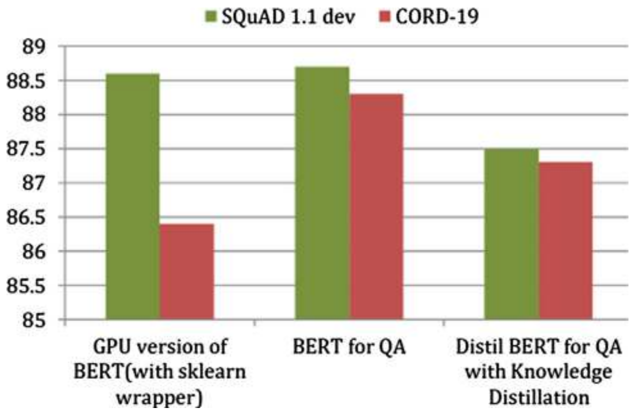


Fig. 7 F1 score comparison of models

generation quality that COBERT is handling, this capability is likely to increase along with the content of literature given to our system.

The model comparison based on EM and F1 score shown in Table 3, Figs. 6, and 7. GPU version of BERT (with sklearn wrapper) is a version of the BERT model trained on SQuAD 1.1 runnable on GPU. It is available only with a sklearn wrapper achieving an EM score of 81.2% and 88.6%, whereas after fitting the pipeline on the CORD-19 corpus, the model achieves 79.3% EM and 86.4% F1-score. BERT for the QA model is CPU/GPU agnostic. The model was loaded in a machine with support for CUDA. It automatically sends the model to GPU for computations. This version of the model achieves 81.3% EM and 88.7% F1-score on SQuAD 1.1 dev, and after fitting the pipeline on CORD-19 corpus, the model achieves 81.5% EM and 88.3% F1-score. Here there is an increase in EM, and the F1 score is comparable. DistilBERT model trained on SQuAD 1.1 using Knowledge Distillation and Bert-large-uncased-whole-word-masking-finetuned-squad as a teacher. This version of DistilBERT achieves 80.1% EM and 87.5% F1 while being much faster and lighter. Whereas, after fitting the pipeline on the CORD-19 corpus, the model achieves 81.6% EM and 87.3% F1-score. Here, we achieved EM highest among all the pre-trained models on SQuAD 1.1 dev and fitted model of CORD-19 as well, whereas achieved a comparable F1 score.

Table 3 Comparative Analysis Of Models

MODEL	Dataset	EM	F1
GPU version of BERT(with sklearn wrapper)	SQuAD 1.1 dev	81.2	88.6
BERT for QA	SQuAD 1.1 dev	81.3	88.7
Distil BERT for QA with Knowledge Distillation	SQuAD 1.1 dev	80.1	87.5
GPU version of BERT(with sklearn wrapper)	CORD-19	79.3	86.4
BERT for QA	CORD-19	81.5	88.3
Distil BERT for QA with Knowledge Distillation	CORD-19	80.6	87.3

Although both the DistilBERT's EM score improved on the fine-tuned data, due to large size input document size of the CORD-10 dataset as compared to SQuAD1.1 but still significant increase in score of 0.4 shows that DistilBERT performs better on large input sequence size dataset. Whereas, it's not the case with the BERT uncased version which saw a decline in EM in GRU version but still improved in non-GPU version. Therefore, we can say that our Distil BERT for QA with Knowledge Distillation after considering all the above models performs better than BERT on data with large input sequence with 60 per cent faster performance in terms of speed and 40 percent smaller size than BERT-uncased whose reason is due to its triple loss combining language modeling.

7 Limitations

As the system is developed on literature based dataset of CORD-19 hence the answerability of the model on unstructured dataset is still a domain of research which is not incorporated in the current study. Also, we attempted to set a limit for the scores of answers, the conveyance of scores were very conflicting across various queries. The proposed model captures the context-sensitive features only from the generated embedding.Our system is pre-trained on a closed domain of COVID-19 literature that too in English language only. There is also a scope to integrate the system with various trusted websites and articles too.

8 Conclusion

In this paper, COBERT- question answering system architecture was proposed where our outline was to tackle challenges of COVID-19 to aid researchers and clinical workers to get the authentic scientific information at ease. This model was based on a three component architecture of Retrieval, Reader and Ranker to automatically answer .We show that our architecture is able to capture similarity among query and large number of text documents .Due to the portrayed attributes of our methodology portability is guaranteed which is a critical favorable position in correlation to past methodologies. We have indicated the enormity of our methodology in a broad assessment over various benchmarks. We have fine-tuned the BERT uncased and DistilBERT version of BERT in the Reader phase on the CORD-19 dataset then compared documents on the basis of cosine similarity to get the best answer. The results of the system were tested on a diverse set of questions and found to be consistent by the users. We have shown with help of empirical results that DistilBERT is able to capture features dependencies of long documents in a better way due to its triple loss combining language modeling obtaining a EM/F1 score of 81.6/87.3. Our proposed model is additionally simple to be summed up to general area and has the option to enormously help the endeavors of scientists to viably help to battle the current pandemic.

9 Future Works

We would like to consider creating knowledge based graphical representation to map the similarity of extracted documents to integrate them on the basis of ranking and referral perspective. To incorporate unstructured text an extra component called verifier can be included which tactfully extracts the answer collectively from the unstructured data too. To capture not only context-sensitive features but also semantics which are structured which can provide rich language representation using techniques such as role labeling before feeding the input sequence to BERT then using self-attention to capture word-level representation. Also, we want to explore the multilingual capabilities of the model along with adding real-time content from resources such as trusted websites.

Funding This research recieved no external funding.

Availability of data and material Source mentioned.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code availability We can upload as per requirement.

References

1. Acter, T.; Uddin, N.; Das, J.; Akhter, A.; Choudhury, T.R.; Kim, S.: Evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as coronavirus disease 2019 (COVID-19) pandemic: A global health emergency. Science of the Total Environment p. 138996 (2020)
2. Jeyaprakash, K.; Velavan, S.; Ganesan, S.; Arjun, P.: Covid-19-molecular transmission and diagnosis-Review article. Asian J. Innov. Res. **5**(2), 1 (2020)
3. da Silveira, M.P.; da Silva Fagundes, K.K.; Bizuti, M.R.; Starck, É.; Rossi, R.C.; e Silva, D.T.d.R.: Physical exercise as a tool to help the immune system against COVID-19: an integrative review of the current literature. Clinical and experimental medicine p. 1 (2020)
4. Shen, I.; Zhang, L.; Lian, J.; Wu, C.H.; Fierro, M.G.; Argyriou, A.; Wu, T.: In search for a cure: recommendation with knowledge graph on CORD-19. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, p. 3519 (2020)
5. Soni, S.; Roberts, K.: An evaluation of two commercial deep learning-based information retrieval systems for covid-19 literature. J. Am. Med. Inform. Assoc. **28**, 132–137 (2020)

6. Hofstätter, S.; Zlabinger, M.; Sertkan, M.; Schröder, M.; Hanbury, A.: Fine-Grained Relevance Annotations for Multi-Task Document Ranking and Question Answering. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, p. 3031 (2020)
7. Emre, K.; Demirkan, K.; Serhat, Ü.: Knowledge and attitudes among hospital pharmacists about COVID-19. Turkish J. Pharm. Sci. **17**(3), 242 (2020)
8. Kricka, L.J.; Polevikov, S.; Park, J.Y.; Fortina, P.; Bernardini, S.; Satchkov, D.; Kolesov, V.; Grishkov, M.: Artificial intelligence-powered search tools and resources in the fight against covid-19. Ejifcc **31**(2), 106 (2020)
9. Wise, C.; Ioannidis, V.N.; Calvo, M.R.; Song, X.; Price, G.; Kulkarni, N.; Brand, R.; Bhatia, P.; Karypis, G.: COVID-19 Knowledge Graph: Accelerating Information Retrieval and Discovery for Scientific Literature. arXiv preprint arXiv:2007.12731 (2020)
10. Diefenbach, D.; Both, A.; Singh, K.; Maret, P.: Towards a question answering system over the semantic web. Semantic Web (Preprint) **1**,(2020)
11. Yu, W.; Wu, L.; Deng, Y.; Mahindru, R.; Zeng, Q.; Guven, S.; Jiang, M.: A Technical Question Answering System with Transfer Learning. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, p. 92 (2020)
12. Abacha, A.B.; Hasan, S.A.; Datla, V.V.; Liu, J.; Demner-Fushman, D.; Müller, H.: VQA-Med: overview of the medical visual question answering task at ImageCLEF 2019. In: CLEF (Working Notes) (2019)
13. Nogueira, R.; Cho, K.: Passage re-ranking with bert (2020)
14. Gao, L.; Dai, Z.; Callan, J.: Understanding BERT Rankers Under Distillation. In: *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, p. 149 (2020)
15. Carrino, C.P.; Costa-jussà, M.R.; Fonollosa, J.A.: Automatic spanish translation of the squad dataset for multilingual question answering. arXiv preprint arXiv:1912.05200 (2019)
16. Qu, C.; Yang, L.; Qiu, M.; Croft, W.B.; Zhang, Y.; Iyyer, M.: BERT with history answer embedding for conversational question answering. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 1133 (2019)
17. Yang, W.; Xie, Y.; Lin, A.; Li, X.; Tan, L.; Xiong, K.; Li, M.; Lin, J.: End-to-end open-domain question answering with bertserini. arXiv preprint arXiv:1902.01718 (2019)
18. Yang, W.; Zhang, H.; Lin, J.: Simple applications of BERT for ad hoc document retrieval. arXiv preprint arXiv:1903.10972 (2019)
19. Esteva, A.; Kale, A.; Paulus, R.; Hashimoto, K.; Yin, W.; Radev, D.; Socher, R.: Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization. arXiv preprint arXiv:2006.09595 (2020)
20. Abacha, A.B.; Demner-Fushman, D.: A question-entailment approach to question answering. BMC Bioinform. **20**(1), 511 (2019)
21. Tang, R.; Nogueira, R.; Zhang, E.; Gupta, N.; Cam, P.; Cho, K.; Lin, J.: Rapidly bootstrapping a question answering dataset for COVID-19. arXiv preprint arXiv:2004.11339 (2020)
22. Lee, J.; Yi, S.S.; Jeong, M.; Sung, M.; Yoon, W.; Choi, Y.; Ko, M.; Kang, J.: Answering questions on covid-19 in real-time. arXiv preprint arXiv:2006.15830 (2020)
23. Su, D.; Xu, Y.; Yu, T.; Siddique, F.B.; Barezi, E.J.; Fung, P.: CAiRE-COVID: A Question Answering and Multi-Document Summarization System for COVID-19 Research. arXiv preprint arXiv:2005.03975 (2020)
24. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
25. Chen, J.; Zhou, J.; Shi, Z.; Fan, B.; Luo, C.: Knowledge abstraction matching for medical question answering. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). p. 342, IEEE (2019)
26. Farias, A.; Mikaelian, F.; Amrouche, M.: Closed domain question answering (2019). https://cdqa-suite.github.io/cdQA-website/
27. Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; tau Yih, W.; Choi, Y.; Liang, P.; Zettlemoyer, L.: Quac : Question answering in context (2018)
28. McCarley, J.S.; Chakravarti, R.; Sil, A.: Structured pruning of a bert-based question answering model (2020)