TopiQAL: Topic-aware Question Answering using Scalable Domain-specific Supercomputers

Hamsa Shwetha Venkataram
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, USA
venkatar@jpl.nasa.gov

Chris A. Mattmann
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, USA

Scott Penberthy
Applied AI, Google
Mountain View, USA

***Abstract*—We all have questions. About today's temperature, scores of our favorite baseball team, the Universe, and about vaccine for COVID-19. Life, physical, and natural scientists have been trying to find answers to various topics using scientific methods and experiments, while computer scientists have built language models as a tiny step towards automatically answering all of these questions across domains given a little bit of context. In this paper, we propose an architecture using state-of-the-art Natural Language Processing language models namely Topic Models and Bidirectional Encoder Representations from Transformers (BERT) that can transparently and automatically retrieve articles of relevance to questions across domains, and fetch answers to topical questions related to COVID-19 current and historical medical research literature. We demonstrate the benefits of using domain-specific supercomputers like Tensor Processing Units (TPUs), residing on cloud-based infrastructure, using which we could achieve significant gains in training and inference times, also with very minimal cost.**

I. Introduction

Text has been the largest contributor to the increase in unstructured content there is today, thanks to smarter search engines, micro-blogging sites, social media, and indirectly due to improved research capabilities and affordable storage. As a consequence, Natural Language Processing (NLP) research boundaries are being constantly pushed in the past decade, to reach newer levels of text understanding and tapping the hidden potential that text has to offer.

Butting up against this is the hard truth that domain-specific tasks we intend to solve can be increasingly complicated, from trying to predict from posts what constitutes hate speech, and/or extremism, or trying to comprehend what is truthful on the Internet, versus that which is likely to be spoofed or falsely generated. Even more applicable is the task of gathering news posts that have to do with a global pandemic forcing us throughout the world to confront isolation, our health, and to rapidly search anything and everything on the Internet to seemingly find a cure.

In these trying times, attempting to match human capabilities for natural language understanding in these areas has been nearly impossible to date, and even more so when trying to reduce the amount of time for human-in-the-loop (HiTL) in cases of news analysis, social media contextualization and interpretation of scientific literature. Each domain, for example biomedical science, poses unique challenges, making it impossible to build a generic model as a panacea for all problems.

In this paper, we focus on literature related to SARS-CoV-2, or the COVID-19 pandemic. We explore the wealth of medical research literature that could possibly hold cures, relationships, key findings, and further contextualization for understanding the virus. We develop new text processing techniques using state of the art Bidirectional Encoder Representations from Transformers (BERT) language models and topic modeling using Latent Dirichlet Allocation (LDA). Finally we develop methodologies demonstrating how these techniques can be combined to rapidly identify the usable and extremely relevant information from the medical research literature that we explored as part of the Kaggle CORD-19 challenge run by the Office of Science and Technology Policy (OSTP) and the White House.

A. CORD-19 Dataset

The COVID-19 Open Research Dataset [1] or *CORD-19* [1] is a vast collection of extensive machine-readable literature with more than 29,000 scholarly articles upon initial release in March 2020, and more than 200k articles at the time of writing this paper. This dataset jointly curated by Allen Institute for AI (AI2), the White House Office of Science and Technology Policy (OSTP), the National Library of Medicine (NLM), the Chan Zuckerburg Initiative (CZI), Microsoft Research, and Kaggle, coordinated by Georgetown University's Center for Security and Emerging Technology (CSET), consists of rapidly growing collection of research literature about COVID-19 and historical SARS and MERS coronaviruses.

B. Problem Statement

A call to action was issued by the White House to the nation's NLP community to develop data and text mining tools to answer key scientific questions drawn from [2], [3] related to COVID-19.

C. Assumptions

Given the nature of the task, criticality of COVID-19, and based on medical research literature landscape, certain reasonable assumptions on the architecture and resulting models

[1] https://www.semanticscholar.org/cord19

DOI 10.1109/DLS51937.2020.00011

were made. The primary assumption was that the community should develop and distribute transparent models, that can be easily understood and easier to communicate. Each participant in the challenge should develop solutions that ideally be self explanatory, and require minimal NLP knowledge. Secondly, the systems built by various groups in the community should also serve as a tool for curating knowledge of relevance for further analysis, and that can critically impact the results obtained since those participating in the challenge weren't just working together in the challenge, but for the good of the world as it races to understand the virus.

D. Contributions

We propose an interpretable, unsupervised, generic and fused machine learning and deep learning architecture named TopiQAL that hierarchically sifts articles via (*question, abstract*) and (*question, paragraph*) topic matching, and feeds the resultant set of articles to the fine-tuned biomedical domain-specific BERT [4] question-answering [5] model. Our solution aggregates answers and presents them to the user with confidence scores. To do so, we train small footprint topic models at different degrees of granularity and leverage dominant topic signals to extract relevant abstracts and thus papers of interest. This effectively facilitates filtering the articles of interest based on the mixture of topics in the questions, while also reducing the number of tokens that are fed into the BERT model at the paragraph level. We demonstrate how we leverage two unsupervised language models, one interpretable for fine-grained controls, and the other pretrained model with a deeper sense of context, both beneficial for faster iterations, minimal compute and cost.

The purpose of this architecture is indeed threefold:

- Perform topic discovery in the corpus by seeding task-specific word priors to the topics and information filtering to obtain relevant articles of interest based on short text clues with lesser context, in this case, questions. This may further lead to document classification or other language modeling tasks.
- Conforming to pretrained or domain-specific BERT's sequence length constraints by breaking down larger text into multiple chunks of shorter paragraphs using hierarchical topic matching, thereby avoiding retraining BERT models which is expensive and time-consuming. More of this will be discussed later in Section IV.
- Leverage domain-specific cloud-based supercomputers for training and performing inference on expensive deep learning models like BERT.

The paper is structured as follows: Section II presents recent work and other literature in this domain, Section III gives a rationale behind TopiQAL, and Section IV talks about TopiQAL in detail. The domain-specific supercomputing is discussed in Section V and experiments in Section VI. We discuss the shortcomings, advantages and error analysis of architecture's responses in Section VII, and we wrap up with conclusions in Section VIII and present plans for future work in Section IX.

II. Related Work

Ever since the CORD-19 dataset was published, multiple applications and research studies have been carried out to find answers for the scientific questions. Research has spanned across building knowledge graph-based applications [6], semantic text similarity datasets [7], COVID-19 benchmark question answering datasets [8] to evaluating language models suitable for NLP tasks [9]. Multiple works have been carried out in fusing topic modeling and BERT together, mostly by ways of modifying Transformer architecture to provide topic information [10], [11], and for various applications.

The major impediment to processing and understanding larger text in this context is the constraints of BERT's maximum sequence length. This may either be overcome by performing operational changes, such as reduction in the number of tokens, or architectural changes, by modifying the length of the sequence and retraining the BERT model from scratch. Building longer text understanding capabilities is currently a major area of research, and BERT-AL [12] is one such direction. In this paper, owing to the fact that architectural changes take considerable time and compute, we choose to perform operational changes and truncate larger articles into length of 512 tokens each, details of which we will see in ensuing sections.

Our approach fuses topic modeling and BERT together in a single unified and understandable approach and allows each technique to be iteratively improved independently and combined together to provide for optimal tuning and analysis as we will demonstrate in the remainder of the paper.

III. Background

At the time of initial release of CORD-19 dataset and with Kaggle challenge underway, there were 29,000 scholarly articles, including over 13,000 with full text. In addition, there were high priority scientific questions that would benefit the medical experts. But to locate answers, articles are unannotated, and for that matter, section headers are inconsistent. Absence of subject matter experts (SMEs) in the loop, added to the inherent complexity of the task. For a system to be able to fetch answers, we had to develop a methodology to sift through the tremendous number of publications. Even at a scale of 10s of thousands, there was still too much information even for hundreds to thousands of independent researchers to sift through owing to the complexity of subject matter at hand since we all were not virologists or immunologists, and those participating were mostly computer scientists and data scientists.

Our systems had to trade between presentation of all of the relevant articles given a set of research questions and providing capability to capture articles such that interested parties could:

Decide what questions to ask based on underlying topic distributions of articles.

Subset the article to relevant sections and reduce the number of paragraphs in the articles that we ask questions on, since as we will see later, BERT imposes restrictions on sequence length of context.

A. Abstracts or Full articles ?

But do topic distributions of articles help? We asked ourselves this question given that abstracts - one of our starting points provided by the CORD-19 dataset - are generally sources of dense information, and succinct in explaining main topics of a given paper. One may then argue that the underlying topic distributions for the abstracts are not a very accurate measure of topic distributions in articles themselves. However, topics in the articles tend to be widely distributed in aspects of related works, experiments, datasets, which may not be descriptive of the crux of the articles. Hence we perform topic modeling at degrees of granularity - abstract and body of text, thereby filtering articles and paragraphs of interest.

B. Peer-reviewed vs Preprints

In the field of medical research, there have been subjective concerns about the advantages and disadvantages of scientific discoveries published on preprint servers versus conventional peer-reviewed journals. While some argue that preprints can accelerate and help in faster dissemination of results [13], some concerns still exist [14]. During the spread of Ebola and Zika viruses, there were merely 5% percent articles published as preprints, and owing to the slower process of peer-reviews, most of the literature was published post reduction in virus spread [13]. Recent advocacy for preprints may result in increases in potentially questionable articles being published going forward (indeed at the time of writing of this paper, over 32 papers have been retracted recently written about COVID-19)[2]. The CORD-19 corpus contains research articles from different data sources - PubMed Central, CZI, bioRxiv, medRxiv, out of which latter three are preprint servers.

To account for this, we create a delineated set of models, keeping domain and source overlap to a minimum, while facilitating researchers to trace back to answer sources. On the other hand, upon qualitative analysis, topic models gathered coherent topic terms in this process than the models trained on entire corpus. This gives scientists a choice of answer sources, thus building confidence and trust over the answers and chosen articles of relevance.

C. BERT

BERT [15], arguably the most popular pretrained Transformer model [16], instantly changed the NLP research landscape in the past year and has positively impacted downstream tasks like Named Entity Recognition (NER), Question Answering (QA) and Relation Extraction (RE). In this paper, we exclusively look at the task of QA to answer the questions posed in Kaggle tasks. BERT or any domain-specific BERT is fine-tuned on SQuaD 2.0 dataset [5], such that it learns to identify *[start]* and *[end]* span of an answer in the context.

Topic Modeling ensures interpretability of the results, especially for a domain as critical as COVID-19 research. The models ensure that BERT [15] is presented with articles retrieved using a set of keywords for fetching answers. In other words, we choose topical or relevant articles of interest based on the question cues, and hence the name of our system as the TopiQAL platform and is depicted in Fig. 1. We describe it in more detail in the ensuing section.

IV. TopiQAL

A. Topic Models

As noted earlier, Topic Modeling or *probabilistic topic modeling* [17] has been in the forefront of breakthrough NLP techniques in the past few years. Topic models accelerate summarizing and organizing a massively large collection of documents, which would otherwise be next to impossible by human annotation. [17] argues that topic models automatically discover main themes that pervade the large collection of documents, and topics are said to emerge with analysis of original text under the algorithm's statistical assumptions.

In this paper, we use Latent Dirichlet Allocation or *LDA* which is, according to authors [18], a generative probabilistic model for collections of discrete data such as text corpora. LDA is the simplest topic model algorithm that tries to capture the intuition that documents exhibit multiple topics, statistically models this intuition, and is widely used in the NLP community.

Notations: We would like to introduce notations and terminologies that are typical to LDA [18], and minor adaptations in our current context.

[2] https://retractionwatch.com/retracted-coronavirus-covid-19-papers/

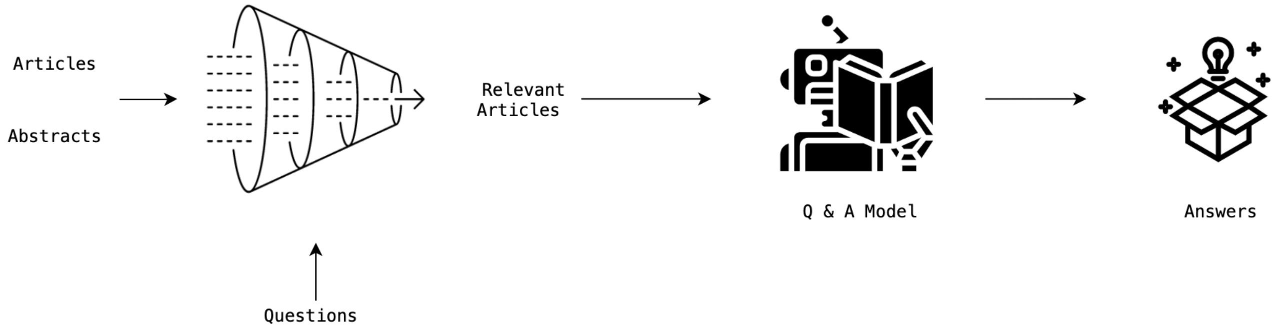


Fig. 1. TopiQAL Architecture

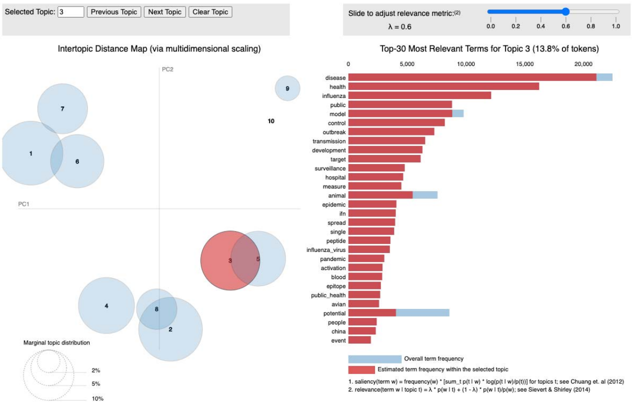
- A *term* or *word*, used interchangeably in the paper, signify basic unit of our text corpora, and is indexed in the vocabulary. The vocabulary also contains words from the task questions, which are then used to seed priors for model training.
- A *document,* $\mathbf{d} = (w_1, w_2, ..., w_N)$ is a sequence of words, where $w_{\rm i} \in \mathbb{R}^d$ is the $i^{\rm th}$ word in the sequence, and denotes a research articles with full-text. We do not consider other research articles in the scope this paper.
- An *archive,* $\mathcal{A} = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_M\}$ is a preprint server or peer-reviewed journal which is a collection of M full-text articles, where $\mathbf{d}_{\rm j}$ is the $j^{\rm th}$ full-text article, published under archive $\mathcal{A}$.
- A *corpus,* $\mathcal{C} = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_P\}$ is a collection of archives, where $\mathbf{a}_{\rm k} \in \mathcal{A}$ is the $k^{\rm th}$ archive in $\mathcal{C}$.

The choice of number of topics t are crucial when building topic models. There are about 10 thematic tasks on Kaggle, and every task signifies a theme, for example, risk factors, vaccines, transmission, origin etc. with questions that further elaborate on the theme with certain domain-specific terms. Hence, we choose $t = 10$, each set to represent a task and its related questions.

A topic is a distribution over a fixed vocabulary, according to the formal definition in [17]. It is possible that we set priors to particular word-topic combinations, such that we nudge models to converge in a required direction. This forms a way of incorporating domain knowledge to each topic using questions as cues. Thematic question along with the sub-questions are tokenized by removing stop words, and are used to set priors as a word-topic combination probability. Specifically, the *eta* hyperparameter, an A-priori belief on word probability and matrix of shape *(num_topics, num_words)* [19], is assigned a probability q for each word-topic combination specified. This topic x word matrix *eta* is initialized with default values, $p = 1/t$. A subset of words that we used to set priors are captured in Table. I. On the other hand, there is also implicit ingestion of $\mathcal{A}$-specific domain knowledge since we choose to train topic models one for each $\mathbf{a}_{\rm k}$ in $\mathcal{C}$, thus establishing clear distinctions between the models in $\mathcal{C}$, and supporting the assumptions in Section III.B.

Data Preprocessing: The dataset was prepared from the provided metadata *Comma Separated Values* file and JSON files by removing citations, combining all sections, and divided them based on the archive sources. The corpus is said to have been partly created using Optical Character Recognition (OCR) technique, and we noticed the parsing of the text content was not consistent, and missing metadata values in the JSON files. We take this into account during the error analysis of our architecture.

Training: For every $\mathbf{a}_{\rm k}$ in $\mathcal{C}$, $\mathcal{A} = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_M\}$, each $\mathbf{d}_j$'s abstract is converted to a Bag-of-Words format and phrases of *uni*-grams and *bi*-grams are detected from the sentences based on the collocation counts [20], with a minimum count of 10. LDA model is trained with hyperparameters *eta* as mentioned earlier, *alpha* set to "symmetric" over each topic, *chunksize*=20, *epochs*=10 and 100 iterations through the entire $\mathcal{A}$. Upon completion of training, dominant topic in the topic distribution of each $\mathbf{d}_j$ abstract is assigned to the article.


Selected Topic: 3

Slide to adjust relevance metric:(2)
λ = 0.6

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 2. PubMed Central Abstracts Topic Model using pyLDAVis [21]

Similarly, each $\mathbf{d}_j$'s body texts undergo similar processing and training phases. However, the topics are not assigned, instead the models are stored for hierarchical inference which will be described shortly. Fig. 2 is a visualization of PMC abstracts topic model clusters along two principal component axes.

B. Fine-tuning for QA

It is highly beneficial for our task to have domain specificity in the language model, since original BERT [15] is trained on general corpora. BioBERT [4] is the first domain-specific BERT pretrained on biomedical domain corpora consisting of PubMed abstracts and PMC full-text articles. With minimal architectural modifications, BioBERT has been proven to be effective when fine-tuned for downstream tasks NER, QA and RE [4].

SQuaD 2.0 dataset: SQuaD or Stanford Question Answering Dataset [5] is a reading comprehension dataset containing crowdsourced questions based on a set of Wikipedia articles, and the answer is a part of text from a given passage. SQuaD 2.0 consists of 50k+ unanswerable questions deliberately made to look strikingly similar to answerable ones, and model is expected to answer suitably in such cases. Instead of factoid BioASQ dataset [22] mentioned in [4], SQuaD dataset is used for fine-tuning BioBERT. Evaluation of fine-tuned BERT-Large, Uncased (24-layer, 1024-hidden, 16-heads, 340M parameters) and BioBERT-Large v1.1 (+ PubMed 1M) on SQuaD dev set resulted in 76% and 84% F1-scores respectively. Hence, TopiQAL architecture incorporates BioBERT QA model for question answering component.

C. Hierarchical Inference

In this subsection, we present the steps for performing inference given a question. When TopiQAL is presented with say,

Are there geographic variations in the rate of COVID-19 spread ?,

Theme	Keywords
Transmission, Incubation, and Environmental Stability	asymptomatic, shedding, surface stainless steel, decontamination, persistence, contagious. Adhesion, hydrophilic, phenotypic, nasal discharge, seasonality, transmission
COVID-19 Risk Factors	epidemiological, smoking, pulmonary, morbidity, neonate, pregnant, socio economic, reproductive number, susceptibility, infection, respiratory, virulent
Vaccines and Therapeutics	genome, genetic, sequencing, binding, receptor, reservoir, epidemic, pathogen, animal, livestock, farm, strain, infected
Virus, Genetics, Origin, and Evolution	inhibitor, naproxen, clarithromycin, antibody, ADE, prophylaxis, vaccination, recipient, assay, antiviral, predictive
Medical Care	nursing, mobilization, extracorporeal, membreane, oxygenation, ventilation, extrapulmonary, cardiomyopathy, N95 mask, intervention, workforce, outcome, respirator, telemedicine, hospital
Non-Pharmaceutical Interventions	NPIs, social, distancing, control, spread, housing, status, community, barrier, geographic, location, comply, public, health, advice, programmatic, alternative
Geographic Variation	geographic, rate, covid, spread, variation, mortality, evidence, virus, mutation
Diagnostics and Surveillance	antibody, PCR, testing, genetic, drift, mutations, serosurveys, rapid, influenza, cytokines, viral load, CRISPR, evolutionary hosts, reagents
Concerning ethical considerations for research	school closure, misinformation, anxiety, stigma, ethical principles, outbreak, fear, social medium, multidisciplinary
Information sharing and inter-sectoral collaboration	high-risk, inter-sectoral, government, public health, equity, preparedness, health surveillance, treatment, care, funding, marginalized, disadvantaged

TABLE I
KAGGLE THEMATIC QUESTIONS AND SUBSET OF WORD PRIORS

the question is tokenized and the vectorized question is posed to all abstract LDA models for archives $\mathbf{a}_\mathrm{k}$ in $\mathcal{C}$ for inference. Each $\mathbf{a}_\mathrm{k}$ returns a topic distribution as tuple *(topic_ID, probability)*. We only choose top *topic_IDs* such that

$$p_\mathrm{i} \geq threshold, i = 1, 2, ..., n \tag{1}$$

where p_i is i^th topic probability in the returned list of tuples ordered by their decreasing probabilities. In our experiments, $threshold = 0.2$ is a hyperparameter set to only retrieve topics that contribute significantly, but can be modified operationally, or one can set top *n* topics, whichever suits best. This results in a filtered list of abstracts as a union of all [$\mathbf{a}_\mathrm{k}$, *(topic_ID, probability)*]. Below are the topics that respective models inferred for the given question.

Examples:

Source: CZI
Topics: [(3, 0.30274478), (1, 0.22156551)]
Keywords:

Topic 3: (day, estimated, transmission, infection, time, period, contact, onset, rate, estimate)
Topic 1: (outbreak, epidemic, data, january, health, risk, time, wuhan, country)

Source: bioRxiv
Topics: [(2, 0.46331868)]
Keywords:

Topic 2: (model, time, individual, rate, data, disease, transmission, population, outbreak, parameter)

The second phase involves further reduction using *(question, paragraph)* topic matching and is done as follows. The same question is now asked to all body text topic models trained on full-text articles of archives $\mathbf{a}_\mathrm{k}$ in $\mathcal{C}$, and *topic_IDs* are retrieved following eq. 1. In each $\mathbf{a}_\mathrm{k}$ containing filtered list of abstracts, corresponding full-text article $\mathbf{d}_j$ is split into multiple paragraphs based on *newline* character. Each paragraph is presented to the $\mathbf{a}_\mathrm{k}$ body text topic model, and dominant topic is compared with question's list of *topic_IDs*. All matching paragraphs are concatenated and added to list of lists, along with the question, to be further posed to BioBERT-QA model on Google Cloud's Tensor Processing Unit (TPU) version 3 (v3) pods made available during our research. Fig. 3 captures the hierarchical filtering of articles using source-wise abstract and body text topic models.

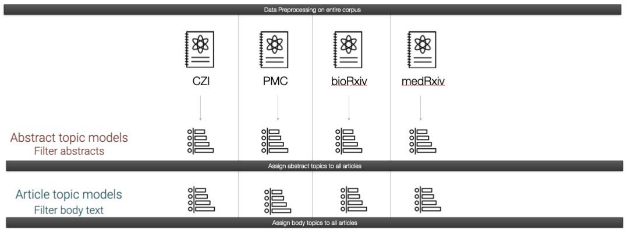

Data Preprocessing on entire corpus
CZI
PMC
bioRxiv
medRxiv
Abstract topic models
Filter abstracts
Assign abstract topics to all articles
Article topic models
Filter body text
Assign body topics to all articles

Fig. 3. Hierarchical Topic Modeling

In the next section we describe the interesting properties of our Cloud-based environment for training TopiQAL using TPUs on Google Cloud.

V. DOMAIN-SPECIFIC SUPERCOMPUTING

Tensor Processing Units or *TPU*s [23] are next-generation hardware accelerators specifically designed for domain-specific neural modeling. To train state-of-the-art language

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
cell	protein	virus	data	mouse	cell	sample	patient	antibody	health
virus	structure	cov	study	level	protein	pcr	study	cell	disease
viral	sequence	strain	time	lung	concentration	gene	infection	vaccine	public
infection	binding	mers	model	day	antibody	dna	respiratory	response	country
protein	domain	sars	population	tissue	assay	rna	virus	immune	human
expression	residue	sequence	individual	cell	anti	primer	clinical	antigen	pathogen
replication	acid	human	rate	study	medium	assay	influenza	mouse	animal
rna	region	bat	influenza	animal	incubated	detection	child	virus	outbreak
gene	interaction	viral	contact	response	culture	sequence	hospital	specific	public health
infected	peptide	host	infection	disease	activity	virus	pneumonia	epitope	surveillance
ifn	amino	genome	transmission	infection	plate	rt	age	serum	dog
host	complex	mers cov	analysis	treatment	control	analysis	symptom	human	risk
pathway	based	specie	network	increased	min	method	rsv	hiv	infectious
antiviral	mutation	sars cov	period	cytokine	pb	time	severe	infection	including
mrna	amino acid	study	outbreak	inflammatory	experiment	genome	acute	titer	potential

TABLE II
PMC BODY TEXT TOPIC MODEL

models like BERT, we need massive computational capability that has faster training times, which means specialized hardware. Supervised learning is an effective machine learning technique, where we present *(input, result)* pairs and the machine learns to adjust its weights or parameters to move from randomly initialized weights to trained weights, such that it can match the expected result. Common domain wisdom is that bigger the machines, bigger are the breakthroughs in deep learning. The computations also contain number of matrix multiplications, convolutions, activation functions at every step of the optimization process to achieve the desired result. Such computational challenges and constraints can be overcome with minimal cost and maximum training accuracy using TPUs on Google's cloud platform infrastructure [24]. TPUv2 and TPUv3 are cheaper when compared to GPUs in terms of price, but offer speedups in performance. BERT alone has 2.2x speedup, and training time has reduced from 3 days to 76 minutes [24]. Our proposed model architecture's BioBERT QA component was fine-tuned using cloud TPUv3 Pods with minimal training time (*approx.* 2 hours), and achieved a higher F1-score of 84% when compared to the original BERT model. Inference times for a question on TPUv3 Pods varied, from 1 minute for shorter 5-10 articles, to about 20 minutes to few hours when posed with longer articles with above 20k tokens.

VI. EXPERIMENTS

In this section, we present the results of the TopiQAL architecture, and the heterogeneous infrastructure we leveraged along with domain-specific supercomputers.

A. Infrastructure

Experimental setup for training topic model and fine-tuning of BERT has been detailed in Section IV, and in this subsection, we briefly present the usage of multiple cloud-based infrastructures in conjunction with local workstation for training and inference of models. Fig. 4 brings out the heterogeneity of the setup. In lieu of this paper's scope, we do not delve into details of summarizer component. In short, BERT extractive summarizer shown in Fig. 4, an open-source implementation of [25], was incorporated to obtain summaries of the articles retrieved using topic models and component integration is still a work in progress.

B. Experimental Results

Evaluation: Before we analyse the responses, we briefly present our evaluation framework. BERT and BioBERT fine-tuned on SQuaD Dev Set v2.0 described in [5] uses F1-Score as a metric, and we report 76% and 84% respectively. However, evaluation of answers from our model for CORD-19 questions was solely based on qualitative analysis owing to exploratory nature of this task. At the time of building this solution, there were no benchmark answers to compare ours with. Truncating articles into multiple chunks of 512 tokens poses certain challenges for a comprehensive evaluation, since every chunk may possibly offer an answer for the same question. At this juncture, we aggregate all answers for an article, ordered by BERT model's probability scores, and present them to the user for evaluation. This can bootstrap a feedback loop which can then be integrated into a robust evaluation framework with SMEs [26].

BioBERT Responses: We tabulate some of the model responses in Table III. The responses are fetched from individual paragraphs of articles further broken down into 512 tokens. Evaluation column is added to signify qualititive analyses using the articles in question, and validated them as right/wrong (according to our knowledge based on articles), or needs further evaluation from SMEs.

BioBERT model is also capable of returning `None` as an answer simply because there was no answer in that context, or model lacked global context. In such cases, the predictions contain -

```
Question =
{
    "text": "empty",
    "probability": 1.0,
    "start_logit": 0.0,
        "end_logit": 0.0
    }
```

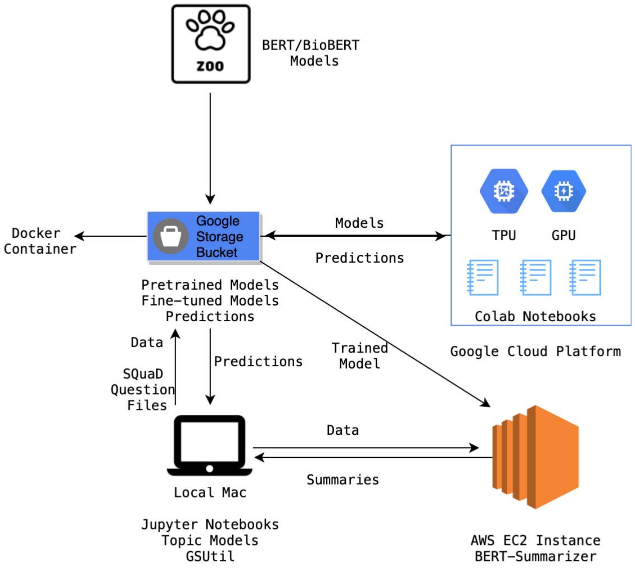


Fig. 4. Heterogeneous Infrastructure

VII. Discussion

Prior knowledge of the data sources effectively prevents early, at times contradicting, results to be consumed as authoritative answers. Dividing a corpus of research articles $\mathcal{C}$ into archives $\mathcal{A}$, gives researchers a fine-grained control over contexts they choose to fetch answers from. We believe this subjective decision is best when left to the medical researchers rather than building one giant model for the sake of generalization, that may lead to negative impacts on understanding science discoveries. As new data sources emerge, this facilitates multiple iterations over training topic models, and thus fetch answers quickly.

Overcoming BERT's maximum sequence length challenge can accelerate integration of other downstream tasks such as extractive summarization into our architecture seamlessly. However, TopiQAL can largely benefit from an architectural change to BERT that can capture the global context and perform better co-reference resolution in a larger setting.

But as one might notice, the current architecture still suffers from the loss of global context due to the reduced length of tokens from an article. To address the concern, we choose to ask the same set of questions to all chunks, aggregate the results for further qualitative analysis by the subject matter experts. We also intend to employ a human-in-the-loop interface based on our previous research work [26], that facilitates easy interaction with the SMEs and quicker iterations to retrain topic models and pass them through BERT-QA models for improved answers.

VIII. Conclusion

In this paper, we proposed a generic architecture called TopiQAL that can automatically retrieve relevant articles of interest from a large corpus of research literature by modeling the topic distributions, and return answers using fine-tuned BioBERT model. We presented this architecture to find answers for critical, topical questions related to coronaviruses using the COVID-19 Open Research Dataset. We demonstrated how one can benefit by having fine-grained control over choosing articles that may contain answers with high probability, and also by reduction in the number of research papers that we ask questions on. We also presented the benefits of running models on domain-specific supercomputers like TPUs and usage of pretrained, domain-specific language models that can together reduce the compute time on training expensive models like BERT.

IX. Future Work

We demonstrated a novel architecture of two unsupervised, state-of-the-art language models towards the task of topic discovery and question answering. In the future, we would like to leverage discovered topics for ranking of answers fetched by the BERT model, and explore the abstractive summarization of topic clusters for building a better model understanding. We also would like to demonstrate the extensibility of this architecture for other downstream tasks, and make it generic across any domain.

Acknowledgments

This effort was supported in part by JPL, managed by the California Institute of Technology on behalf of NASA, and additionally in part by the DARPA Memex/XDATA/D3M/ASED/SafeDocs/LwLL/GCA programs and NSF award numbers ICER-1639753, PLR-1348450 and PLR-144562 funded a portion of the work. The authors would like to thank Annie Didier for her extensive exploratory data analysis, data insights, and preparing the dataset for further ingestion into our architecture.

References

[1] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. M. Kinney, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. Wade, K. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier, "Cord-19: The covid-19 open research dataset," *ArXiv*, 2020.

[2] [Online]. Available: http://www.who.int/blueprint/priority-diseases/key-action/en/

[3] [Online]. Available: https://www.nationalacademies.org/event/03-11-2020/standing-committee-on-emerging-infectious-diseases-and-21st-century-health-threats-virtual-meeting-1

[4] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinform.*, vol. 36, no. 4, pp. 1234–1240, 2020. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz682

[5] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100, 000+ questions for machine comprehension of text," *CoRR*, vol. abs/1606.05250, 2016. [Online]. Available: http://arxiv.org/abs/1606.05250

[6] C. Chen, I. A. Ebeid, Y. Bu, and Y. Ding, "Coronavirus knowledge graph: A case study," *ArXiv*, vol. abs/2007.10287, 2020.

Questions	Answer	Evaluation
Guidance on ways to scale up NPIs in a more coordinated way (e.g., establish funding, infrastructure and authorities to support real time, authoritative (qualified participants) collaboration with all states to gain consensus on consistent guidance and to mobilize resources to geographic areas where critical shortfalls are identified) to give us time to enhance our health care delivery system capacity to respond to an increase in cases.	No silver bullet to use, no standard-operating-procedure (SOP) to follow, and no measures without negative consequences to control the epidemic Unpack the diverse factors social, political, environmental, medical, and legal - that facilitated the escalation of the Ebola crisis Quarantine	SME SME Right
Effectiveness of drugs being developed and tried to treat COVID-19 patients.	2019 The efficacy of computationally designed 44 vaccines has been recently approved for use in defined human vaccines	Wrong SME
What are the possible signs of HBoV?	Fever, sputum production, crackles, wheezing, rhinorrhea, cyanosis, vomiting, and diarrhea	Right
Alternative models to aid decision makers in determining how to prioritize and distribute scarce, newly proven therapeutics as production ramps up. This could include identifying approaches for expanding production capacity to ensure equitable and timely distribution to populations in need.	AI and deep learning Vaccines Auto Regressive Integrated Moving Average Model	Right Wrong SME

TABLE III
TopiQAL Responses

[7] X. Guo, H. Mirzaalian, E. Sabir, A. Jaiswal, and W. AbdAlmageed, "Cord19sts: Covid-19 semantic textual similarity dataset," *ArXiv*, vol. abs/2007.02461, 2020.

[8] R. Tang, R. Nogueira, E. Zhang, N. Gupta, P. Cam, K. Cho, and J. Lin, "Rapidly bootstrapping a question answering dataset for covid-19," *arXiv preprint arXiv:2004.11339*, 2020.

[9] D. Oniani and Y. Wang, "A qualitative evaluation of language models on automatic question-answering for covid-19," *ArXiv*, 2020.

[10] N. Peinelt, D. Nguyen, and M. Liakata, "tBERT: Topic models and BERT joining forces for semantic similarity detection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7047–7055. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.630

[11] M. Ramina, N. Darnay, C. Ludbe, and A. Dhruv, "Topic level summary generation using bert induced abstractive summarization model," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2020, pp. 747–752.

[12] R. Zhang, Z. Wei, Y. Shi, and Y. Chen, "{BERT}-{al}: {BERT} for arbitrarily long document understanding," 2020. [Online]. Available: https://openreview.net/forum?id=SklnVAEFDB

[13] M. A. Johansson, N. G. Reich, L. A. Meyers, and M. Lipsitch, "Preprints: An underutilized mechanism to accelerate outbreak science," *PLOS Medicine*, vol. 15, no. 4, pp. 1–5, 04 2018. [Online]. Available: https://doi.org/10.1371/journal.pmed.1002549

[14] J. Kaiser, "The preprint dilemma," *Science*, vol. 357, no. 6358, pp. 1344–1349, 2017. [Online]. Available: https://science.sciencemag.org/content/357/6358/1344

[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[17] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[19] [Online]. Available: https://radimrehurek.com/gensim/models/ldamodel.html

[20] [Online]. Available: https://radimrehurek.com/gensim/models/phrases.html

[21] [Online]. Available: https://pyldavis.readthedocs.io/en/latest/readme.html

[22] G. Balikas, A. Krithara, I. Partalas, and G. Paliouras, "Bioasq: A challenge on large-scale biomedical semantic indexing and question answering," in *Revised Selected Papers from the First International Workshop on Multimodal Retrieval in the Medical Domain - Volume 9059*. Berlin, Heidelberg: Springer-Verlag, 2015, p. 26–39. [Online]. Available: https://doi.org/10.1007/978-3-319-24471-6_3

[23] N. P. Jouppi, D. H. Yoon, G. Kurian, S. Li, N. Patil, J. Laudon, C. Young, and D. Patterson, "A domain-specific supercomputer for training deep neural networks," *Commun. ACM*, vol. 63, no. 7, p. 67–78, Jun. 2020. [Online]. Available: https://doi.org/10.1145/3360307

[24] Y. You, J. Li, J. Hseu, X. Song, J. Demmel, and C. Hsieh, "Reducing BERT pre-training time from 3 days to 76 minutes," *CoRR*, vol. abs/1904.00962, 2019. [Online]. Available: http://arxiv.org/abs/1904.00962

[25] D. Miller, "Leveraging BERT for extractive text summarization on lectures," *CoRR*, vol. abs/1906.04165, 2019. [Online]. Available: http://arxiv.org/abs/1906.04165

[26] H. S. Venkataram, I. Colwell, S. Liu, P. Southam, C. A. Mattmann, and T. Soderstrom, "Names don't fly: Smart filters for profanity detection and classification in user-generated content," in *1st Workshop on Data Science with Human in the Loop (DaSH), Knowledge Discovery and Data Mining*. ACM, Aug 2020. [Online]. Available: https://bit.ly/3mm1ILx