AWS CORD19-Search: A Scientific Literature Search Engine for COVID-19

Parminder Bhatia, Kristjan Arumae, Nima Pourdamghani, Suyog Deshpande, Ben Snively, Mona Mona, Colby Wise, George Price, Shyam Ramaswamy, Taha Kass-hout

Amazon Web Services AI

Parmib, arumae, nimpourd, suyogd, snivelyb, monamo, colbywi, gwprice, tahak @amazon.com

Abstract

Coronavirus disease (COVID-19) has been declared as a pandemic by WHO with thousands of cases being reported each day. Thousands of scientific articles are being published on the disease raising the need for a service which can organize, and query them in a reliable fashion. To support this cause, we present AWS CORD-19 Search (ACS), a public COVID-19 specific search engine that is powered by machine learning. ACS with its capabilities such as topic based, natural language search queries, and reading comprehension and FAQ matching provides a scalable solution to COVID-19 researchers and policy makers in their search and discovery for answers to high priority scientific questions. We present evaluation and qualitative analysis of the system with specific examples to illustrate the capabilities of the system.

1 Introduction

With the global outbreak of Coronavirus disease (COVID-19) (Guan et al., 2020), the world is in turmoil. Medical researchers are required to work quickly to fully understand and to provide a form of intervention for the virus. Due to a large research focus on the disease, knowledge is published at a rapid rate throughout the world. One such repository of information is curated through the COVID-19 Open Research Dataset Challenge (CORD-19) (Wang et al., 2020). CORD-19 is a joint challenge put forth by Allen Institute (AI2), National Institutes of Health (NIH), and the United States federal government via the White House. The objective of the challenge is to make sense of and extract useful knowledge across thousands of scholarly articles related to COVID-19. CORD-19 aims to connect the machine learning community with biomedical domain experts and policy makers in a race to identify effective treatments and management policies for COVID-19. In accordance with this initiative our goal is to present a scalable solution aimed at aiding COVID-19 researchers and policy makers in their search and discovery for answers to high priority scientific questions. These questions should be understood in their natural language form; examples include: “What do we know about COVID-19 risk factors?” , “Which medications were most beneficial in the 2002 SARS outbreak?”, and “Which is the most referenced paper doing study for Hydroxy-chloroquine?” To appropriately answer these questions we require a system with deeper biomedical understanding of the natural queries and structured knowledge (Rotmensch et al., 2017).

AWS CORD-19 Search (ACS) provides an easy to use search interface where researchers query using natural language 1. ACS goes beyond keyword matching by understanding question semantics to efficiently find relevant answers. As illustrated in Figure 2, we provide the system with a natural language query inquiring about IL-6 inhibitors and showcase the system response with relevant query components highlighted. This query can confuse a traditional search engine that relies solely on text matching as it can only capture term overlap between a query and a document and may not necessarily be relevant to the researchers true intent (i.e. learning implicit relations). Our system, however, establishes the relationship between IL-6 and SARS-CoV-2 showing evidence where elevated IL-6 occur in a large number of patients with severe COVID-19 and higher mortality rates. In the following example, if tasked with understanding the epidemiology and transmission of COVID-19 a researcher may ask about salivary viral load. We provide a further advantage from an organizational perspective using topics with domain specific relations. The user can select clinical-treatments to

1https://cord19.aws/
refine search articles to understand that “...the highest load [is] assumed to be the day before symptoms appear.” Similarly if commissioned to understand convalescent plasma therapy and selecting the three topics as shown in the last example the system quickly identifies the most relevant article and answers the question in a highlight.

In the following sections we will discuss the various individual AWS services that allow ACS to provide this functionality, highlighting the value to scientists who can quickly query, validate their research, and advance their investigations.

2 System Overview

AWS CORD-19 is based on a deep learning-based semantic search model to return a ranked list of relevant documents. This system makes use of topic modeling, knowledge graphs, and natural language queries. Furthermore it leverages document ranking, reading comprehension, as well as FAQ matching. This provides a scalable solution to COVID-19 researchers and policy makers. In this section we present the overall architecture of the system and a closer look at several individual components.

2.1 Amazon Kendra

Amazon Kendra is an enterprise search service provided by AWS. Kendra allows customers to power natural language based searching across their own data. For the purposes of this challenge Kendra has also been toolled around answering questions regarding COVID-19, using the scholarly articles provided by CORD-19. We further boosted Kendra search by leveraging knowledge extracted using the AWS Comprehend Medical (CM) core NERe service. This involves building a knowledge graph using entities extracted using CM and indexing articles with this information.

Kendra allows customers to power natural language based searching across their own data. It primarily consists of three components.

- **Natural language & keyword support** - Amazon Kendra’s ability to understand natural language questions is at the core of its search engine returning the most relevant passage and related documents.

- **Passage retrieval (Rajpurkar et al., 2016) & FAQ matching** - Amazon Kendra can extract specific answers from unstructured data by identifying the closest question to the search query and return the corresponding answer.

- **Document ranking** - To complement the extracted answers, Amazon Kendra uses a deep learning based semantic search model to return a ranked list of relevant documents.

For the purposes of this challenge, Kendra has also been toolled around answering questions regarding COVID-19, using the scholarly articles provided by CORD-19. In order to improve Kendra search and make it clinically more relevant for medical researchers, we are leveraging knowledge extracted using the CM NERe service as well as topics extracted using a semi-supervised prior based LDA approach. While indexing Kendra, data is enriched by creating attributes and index based on

---

Figure 1: AWS CORD-19 Search results page.
"Are IL-6 inhibitors key to COVID-19?"

SARS-CoV-2 and COVID-19: is interleukin-6 (IL-6) the ‘culprit lesion’ of ARDS onset?

"...monoclonal antibody against IL-6, is being tested in a clinical trial against COVID-19 (Sarilimumab COVID-19). Another drug that showed potential inhibition of IL-6 related JAK/STAT pathway is glatiramer acetate which showed potential to downregulate both IL-17 and IL-6."

Clinical Treatment

"When is the salivary viral load highest for COVID-19?"

Elective, Non-urgent Procedures and Aesthetic Surgery in the Wake of SARS-COV-19 [...]

"Patients with COVID-19 have demonstrated high viral loads in the upper respiratory tract soon after their infection, with the highest load assumed to be the day before symptoms appear."

Clinical Treatment

"Is convalescent plasma therapy a precursor to vaccine?"

COVID-19 convalescent plasma transfusion

"...a passive immunotherapy, has been used as a possible therapeutic option when no proven specific vaccine or drug is available for emerging infections."

Clinical Treatment, Immunology, Lab Trials

---

Figure 2: Sample natural language queries and response using AWS CORD-19 Search. The response field is taken directly from the top result of this service. Also provided are the article titles where the answer is taken from as well as selected topics for the response.

medical entities from Comprehend Medical NERe API as well as topics created from Custom Classification.

### 2.2 Comprehend Medical

Comprehend Medical ³ (CM) (Bhatia et al., 2019), is a HIPAA eligible AWS product for medical domain entity recognition (Bhatia et al., 2018), relationship extraction (Singh and Bhatia, 2019) and normalization. Comprehend Medical supports entity types divided into five different categories (Anatomy, Medical Condition, Medication, Protected Health Information, and Treatment, Test & Procedure) and four traits (Negation, Diagnosis, Sign and Symptom). These entities are directly used to enrich the Kendra search.

### 2.3 COVID-19 Knowledge Graph

Knowledge graphs (KGs) are structural representations of relations between real-world entities in the form of triplets containing a head entity, a tail entity, and the relation type connecting them. KG based information retrieval has shown great success in the past decades (Dalton et al., 2014). The COVID-19 Knowledge Graph is a directed property graph constructed from the CORD19 Open Research Dataset of scholarly articles. Entities including scholarly articles, authors, author institutions, citations, extracted topics and comprehend medical entities are used to form relations in the CKG. The resulting KG continues to grow as the CORD19 dataset increases and currently contains over 335k entities and 3.3M relations. The CKG powers a number of features on AWS CORD-19 including: article recommendations, citation-based navigation, and search result ranking by author or institution publication count. Scientific article recommendations are made possible by a document similarity engine that quantifies similarity between documents by combining semantic embeddings obtained from a pre-trained language model (Beltagy et al., 2019) with document knowledge graph embeddings (?) capturing topological information from the CKG.

### 2.4 Topic Models

Topic modeling is a statistical discovery paradigm for generating topics that occur in a collection of documents. Perhaps the most widely used model for topic modeling is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a generative model which groups documents together by observed content, often giving each document a mixture of topics it belongs to. An extension of this work termed Z-label LDA (Andrzejewski and Zhu, 2009) utilizes priors to allow the model to force certain topics which the users have manually curated, or wish to see clustered together.

#### 2.4.1 Generating Topics

For the purposes of this work we experimented with 5, 10, and 20 topic models. The outputs of each clustering size were manually inspected and topic labels were provided by us when inspecting the top ten terms for each cluster. The final granularity of the topic models was chosen by manually deleting and merging topics from the 20 topic model. In general we were able to clearly extract groups which centered around important topics including virology, proteomics, epidemiology, and cellular biology to name a few. However when faced with 20 topics the less populated ones tended to be noisy,
Table 1: Performance of Amazon Cord-19 on annotated CovidQA dataset

| Search       | MAR | EM/F1       | Avg Ans | No Ansrate |
|--------------|-----|-------------|---------|------------|
| ACS          | 50.5| 91.7/93.6   | 1.61    | 0.27       |
| Covidex      | 45.5| 77.1/81.6   | 1.48    | 0.32       |
| COVID-19 RE  | 39.8| 55.7/63.8   | 1.25    | 0.42       |

and captured peripheral information present in the input, such as language (e.g. Spanish and French) or provide redundancies with existing topics (e.g. two topics of Influenza). As a control we ran a publicly available implementation of Z-label LDA with no priors which yields topics close to those extracted using Comprehend. Although similar we observed better definition in certain groups (such as pulmonary diseases, and policy/industry), and decided to use this as the curation entry-point. Our goal was to limit these topics to ten, and compile them in advance as much as possible. With the help of medical professionals we eliminated and combined topics to form the following: Vaccines/immunology, Genomics, Public health Policies, Epidemiology, Clinical Treatment, Virology, Influenza, Healthcare Industry, Pulmonary Infections, and Lab Trials (human).

2.4.2 Multi-Label Classification

Having to manually feed a topic model and re-model the entire corpus once new data becomes available is largely inefficient. We therefore used the topic model labels to train a multi-label classifier (Read et al., 2011). To evaluate the performance of this model we observe the average F1 of each held-out test sample by calculating the set overlap between our gold standard and system labels as follows:

Using this metric our trained model achieved an average F1 of 91.92, with on average 2.37 labels per document. Fewer than 1% of the documents in the test set received no label, using 0.5 as the confidence threshold.

3 Evaluation

Since no designated evaluation data exist, we created a set of 54 COVID queries which are sourced from CovidQA and compared results with Covidex and COVID-19 Research Explorer. The data was prepared by 3 annotators tasked with selecting the correct answer for the questions. We observed the following key insights from our system:

- For both natural language queries and keyword queries, ACS has the highest average number of question for each query, and least number that answer can not be found or determined.
- The annotators have higher agreement using ACS than Covidex or the COVID-19 Research Explorer.
- The mean average recall of relevant documents is 50.5 for CORD-19, comparing with 45.5 for Covidex and 39.8 for COVID-19 Research Explorer. The EM/F1 of selecting the top best answer for CORD-19 is 91.7/93.6, comparing with 77.1/81.6 for Covidex and 55.7/63.8 for COVID-19 Research Explorer.

4 Analysis

In this section we look into a number of sample queries to shed light on how different components of ACS help in improving search results. We begin by observing how small semantic differences in the query alter the results. The first sample in Figure 3 is specific to medications. While the top result does not include the term medication the system highlights ribavirin and corticosteroids. The CORD-19 system understands that these terms represent medications with the help of CM NERe engine. In the second example we change medications to measures and observe the top result discussing border control, and quarantine. This clearly demonstrates that Amazon Kendra has a deep comprehension of token and query meanings.

Finally, we take a look at the effects of topic modeling when grouping and filtering results. The last two examples in Figure 3 showcase the difference this makes in the top result. Without specifying any topic the resulting article discusses high level policy, specifically quarantine measures in Singapore. When we filter by clinical treatment the top result instead focuses on infections which is covered in the clinical setting. Furthermore the extracted text returned to the user still focuses on lessons learned staying true to the query.

5 Limitations and Future Directions

Amazon CORD-19 is an initial step towards helping medical researchers find relevant content in a
| Query                                                                 | Article                                                                 | Response                                                                 |
|----------------------------------------------------------------------|------------------------------------------------------------------------|-------------------------------------------------------------------------|
| “What medications were most beneficial in the SARS outbreak?”       | Development of chemical inhibitors of the SARS coronavirus: Viral heli-case as a potential target | “…spread of SARS, a number of broad-spectrum antiviral medications were empirically administered to the SARS patients during the SARS outbreak in 2003. These medications include ribavirin, HIV protease inhibitors, corticosteroids, and alpha-interferon (IFN-a). In a retrospective review of treatment…” |
| “What measures were most beneficial in the SARS outbreak?”          | Impact of quarantine on the 2003 SARS outbreak: A retrospective modeling study | “During the 2003 Severe Acute Respiratory Syndrome (SARS) outbreak, traditional intervention measures such as quarantine and border control were found to be useful in containing the outbreak…” |
| “What did we learn from the SARS outbreak?” (no topic)              | Use of quarantine in the control of SARS in Singapore                   | “The main lesson to learn from the SARS outbreak is the capability of an emerging infection to cause a pandemic in a short span of time and the paradigm shift needed to respond to such a disease.” |
| “What did we learn from the SARS outbreak?” (clinical-treatment)    | Characteristics of COVID-19 infection in Beijing                        | “We compared the epidemic features between COVID-19 and 2003 SARS to learn lessons and control the outbreak.” |

Figure 3: Sample queries demonstrating the semantic understanding of ACS, the use-fullness of Comprehend Medical NERe, and the utility of topic modeling for filtering.

timely and meaningful way. In order to improve the robustness, we see following areas as direction for future research.

**Feedback Loop** - Since ACS is a search engine the motivation would be to evaluate it as such; using well-established methodologies based on test collections comprising topics (information needs) and human annotations. Since no designated evaluation data exist, our initial focus is to capture different interactions and feedback. Currently, ACS lacks the feedback loop and federated learning approaches where the system would continuously learn and improve the search. However, the system captures feedback from the researchers in the form of implicit and explicit reactions. Implicit feedback evaluation consists of topics of interests, their CTR as well as the ranking of the results which were selected by medical researchers. Explicit feedback evaluation is captured by providing up-down rating associated with each search results. In the future results can be personalized based on this feedback. Now that we have a system in place, our efforts have shifted to broader engagement with potential stakeholders to solicit additional guidance, while trying to balance between the features and ranking.

**Q&A Curation** - Curation and normalization of questions have potential a use-case of presenting trending questions asked by the medical research community at a particular point. However, curation would involve capturing the questions asked as well as identifying similar questions that can be later normalized. Currently, there is no mechanism to curate the questions asked by the researchers.

**Summarization** - Currently, ACS outputs the relevant passage based on the query. It would be beneficial to get the overall summary of the paper. A potential future direction would be to generate summaries (Raffel et al., 2019) from paper abstracts and full body.

6 Conclusion

This paper describes our efforts in building AWS CORD-19 Search with its capabilities consisting of topic based, knowledge graph, and natural language search queries. This is further enhanced with reading comprehension and FAQ matching as well as document ranking providing a scalable solution to COVID-19 researchers and policy makers in their search and discovery for answers to high priority scientific questions. Our solution is powered by Amazon Kendra, Comprehend Medical and Neptune which incorporate the latest neural architectures to provide information access capabilities to the CORD-19 challenge. We hope that our solution can prove useful in the fight against this global pandemic, and that the capabilities we have developed can be applied to analyzing the scientific literature more broadly.

Acknowledgments

We acknowledge the broader collaboration with AI2 team and White house as well as the broader AWS CORD19 team including Tyler Stepke, Kingston Bosco, Victor Wang, Vaibhav Chaddha, Miguel Calvo, Ninad Kulkani, Kevin Longofer, Ray Chang, Adrian Bordone, Tony Nguyen, and Kyle Johnson.
References

David Andrzejewski and Xiaojin Zhu. 2009. Latent dirichlet allocation with topic-in-set knowledge. In Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, pages 43–48. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In EMNLP.

P. Bhatia, B. Celikkaya, M. Khalilia, and S. Senthivel. 2019. Comprehend medical: A named entity recognition and relationship extraction web service. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), pages 1844–1851.

Parminder Bhatia, Busra Celikkaya, and Mohammed Khalilia. 2018. Joint entity extraction and assertion detection for clinical text. arXiv preprint arXiv:1812.05270.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022.

Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity query feature expansion using knowledge base links. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pages 365–374.

Wei-jie Guan, Zheng-yi Ni, Yu Hu, Wen-hua Liang, Chun-quan Ou, Jian-xing He, Lei Liu, Hong Shan, Chun-liang Lei, David S.C. Hui, Bin Du, Lan-juan Li, Guang Zeng, Kwok-Yung Yuen, Ru-chong Chen, Chun-li Tang, Tao Wang, Ping-yan Chen, Jie Xiang, Shi-yue Li, Jin-lin Wang, Zi-jing Liang, Yi-xiang Peng, Li Wei, Yong Liu, Ya-hua Hu, Peng Peng, Jian-ming Wang, Ji-yang Liu, Zhong Chen, Gang Li, Zhi-jian Zheng, Shao-qin Qiu, Jie Luo, Chang-jiang Ye, Shao-yong Zhu, and Nan-shan Zhong. 2020. Clinical characteristics of coronavirus disease 2019 in china. New England Journal of Medicine, 382(18):1708–1720.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. Machine learning, 85(3):333.

Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. 2017. Learning a health knowledge graph from electronic medical records. Scientific reports, 7(1):1–11.

Gaurav Singh and Parminder Bhatia. 2019. Relation extraction using explicit context conditioning. arXiv preprint arXiv:1902.09271.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. Cord-19: The covid-19 open research dataset. arXiv preprint arXiv:2004.10706.