AWS CORD-19 Search: A Neural Search Engine for COVID-19 Literature

Parminder Bhatia, Lan Liu, Kristjan Arumae, Nima Pourdamghani, Suyog Deshpande,
Ben Snively, Mona Mona, Colby Wise, George Price, Shyam Ramaswamy, Xiaofei Ma,
Ramesh Nallapati, Zhiheng Huang, Bing Xiang, Taha Kass-Hout
Amazon Web Services AI
parmib, liuall, arumae, nimpourd, suyogd, snivelyb,
monamo, colbywi, gwprice, shyar, xiaofeim, rnallapa,
zhiheng, bxiang, tahak @amazon.com

Abstract

Coronavirus disease (COVID-19) has been declared as a pandemic by WHO with thousands of cases being reported each day. Numerous 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, neural search engine that is powered by several machine learning systems to support natural language based searches. ACS with capabilities such as document ranking, passage ranking, question answering, knowledge graph based ranking and biomedical topic classification 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 a quantitative evaluation and qualitative analysis of the system against other leading COVID-19 search platforms. ACS is top performing across these systems yielding quality results which we detail with relevant examples in this work.

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 collection of over 100,000 of COVID-19 scientific articles that is publicly available for research community to fight against coronavirus. It 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 provide a scalable solution to access insightful COVID-19 information easily using advanced NLP techniques. For example, these questions should be understood in their natural language form properly: "What are the recommended medications for COVID-19?" and "What is the average hospitalization time for patients?" To retrieve answers and relevant information for these questions, we require a system with a strong biomedical understanding of the natural questions (Rotmensch et al. 2017).

AWS CORD-19 Search (ACS) provides an easy to use search interface where researchers can query using natural language questions in addition to traditional keyword-based search throughout CORD-19 corpus[1]. It performs document ranking to retrieve a ranked list of COVID-19 articles based on relevancy. In addition, it supports passage ranking and question answering, which extract and highlight the answer directly from the top relevant passages. Both components are built with deep learning models.

Figure 1 shows a screenshot of the ACS interface of querying with a natural language question "What is the mean value of R0 for COVID-19?" In the top suggested answer area, the answer *3.28* is highlighted within the top passage. There are at most 3 suggested answers extracted from different article sources. In the bottom retrieved document results, the article is presented with its author, year, journal, institution and citation. It is worthwhile to highlight the browsing features. The users can select topic to filter the result, and sort the result based on best match, most citation, most published authors and institutions. These capabilities are supported by the COVID-19 literature topic modeling and knowledge graph, respectively.

Leveraging CORD-19 corpus, there are a few web search engines that support COVID-19 search. Google Covid-19 Research Explorer[2] and Neural Covidex[3] are the other two leading semantic search web interfaces that are based on advanced NLP techniques and also support document ranking and question answering. While these engines can gain high traction from the public, there is lack of formal evaluation about their performance. In this paper, we also perform a systematic study of ACS performance against these two engines.

In the following sections, we will present a system overview about the various individual Amazon Web Services (AWS) products that supports ACS, a performance evaluation, and future directions to improve ACS.

[1] https://cord19.aws/

[2] https://covid19-research-explorer.appspot.com/

[3] https://covidex.ai/

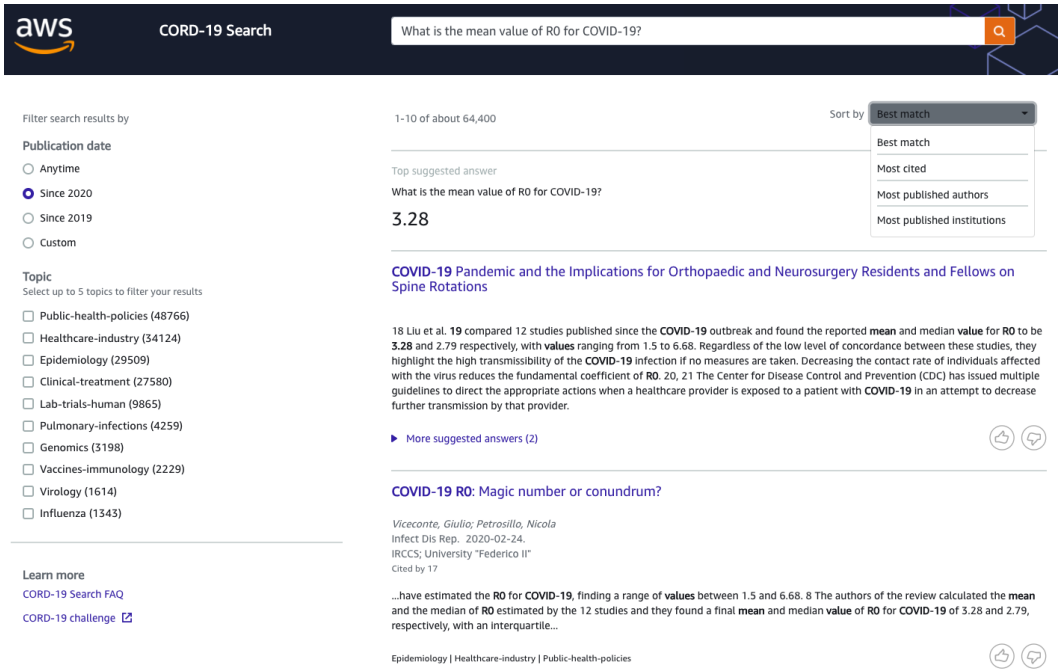

Figure 1: AWS CORD-19 Search results page.

System Overview

AWS CORD-19 Search (ACS) is a semantic search engine that performs document ranking, passage ranking and question answering. Additionally, it leverages knowledge graphs and topic modeling to enrich the biomedical data and make the search to be more clinically relevant. In this section we present the overall architecture of the system and a closer look at several individual components. Figure 2 provides an overview of this architecture.

Amazon Kendra

Amazon Kendra[4] is a semantic search and question answering service provided by AWS for enterprise customers. Kendra allows its customers to power natural language based searching across their own data. As response to the worldwide COVID-19 pandemic, Kendra has also been tooled to support COVID-19 related search using the CORD-19 article corpus. The end-to-end Kendra system consists of several components.

- **Document ranking (DR)** - Like any traditional search engine, Kendra returns a ranked list of relevant documents based on the user's query to fulfill their information needs. A deep semantic search model is used to understand natural language questions in addition to the keyword search.
- **Passage ranking (PR) & Question Answering (QA)** - Kendra ranks the passages and tries to extract the answer from the top relevant passages with a deep reading comprehension model.
- **FAQ Matching (FAQM)** - If there exists Frequently Answered Questions database, Kendra will automatically match a new coming query with FAQs and extract the answer if a strong match is found.

In order to improve Kendra CORD-19 search and make it clinically more relevant for medical researchers, we leverage knowledge extracted using the Amazon Comprehend Medical (CM) core NERe service as well as topics created from Amazon Comprehend Custom Classification [5]. Both are used to enrich the data when indexing CORD-19 corpus.

Comprehend Medical

Amazon Comprehend Medical [6] (CM) (Bhatia et al. 2019), is a HIPAA eligible AWS service for medical domain entity recognition (Bhatia, Celikkaya, and Khalilia 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 Test, Treatment, & Procedure) and four traits (Negation, Diagnosis, Sign and Symptom). These entities are directly used to enrich the Kendra search.

COVID-19 Knowledge Graph

Knowledge graphs (KGs) are structural representations of relations between real-world entities in the form of triplets

[4] https://aws.amazon.com/kendra/

[5] https://docs.aws.amazon.com/comprehend/latest/dg/how-document-classification.html

[6] https://aws.amazon.com/comprehend/medical/

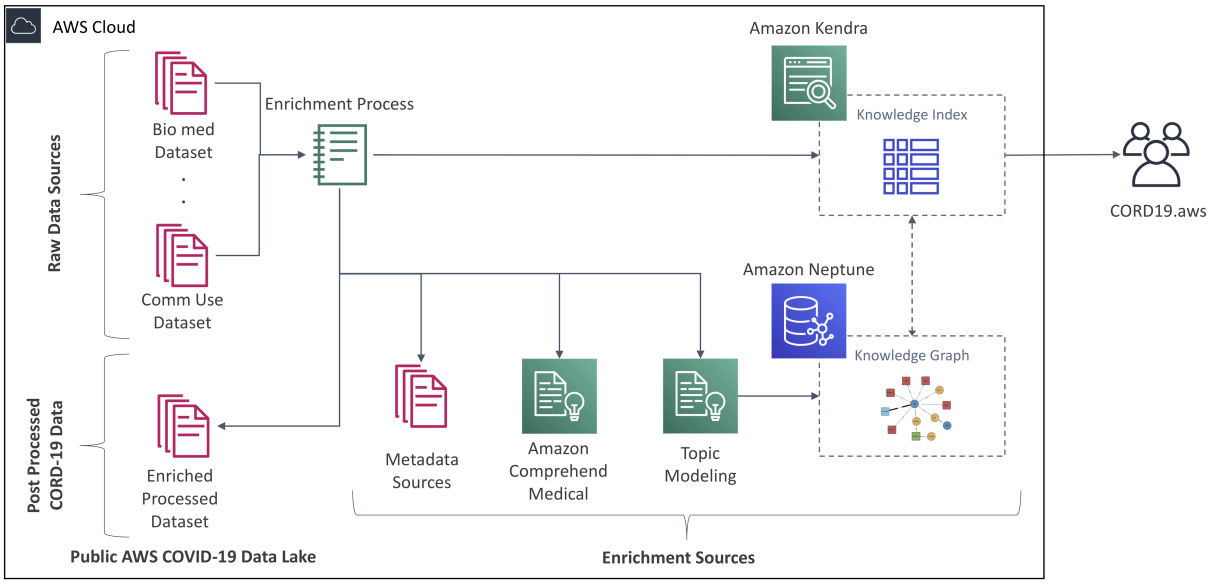

Figure 2: System Architecture.

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, Dietz, and Allan 2014). The COVID-19 Knowledge Graph (CKG) (Fig 3) is a directed property graph constructed from the CORD-19 Open Research Dataset of scholarly articles (Wise et al. 2020). 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 CORD-19 dataset increases and currently contains over 335k entities and 3.3M relations. The CKG powers a number of features on ACS including: article recommendations, citation-based navigation, and search result ranking by author or institution publication count.

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, Ng, and Jordan 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.

Generating Topics For the purposes of this work we experimented with 5, 10, and 20 topic models [7]. 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, 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 for Influenza). As a control we ran a publicly available implementation of Z-label LDA [8] with no priors which yields topics close to those extracted using Comprehend Topic Modeling. 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).

Multi-Label Classification Having to manually feed a topic model and re-train on 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 calculate the average F_1 across test samples by calculating the set overlap between our gold standard (topic model) and system labels (multi-label classification). This held-out test set contains 20% of the CORD-19 data available at the time.

Using this metric our trained model achieved an average F_1 of 91.92, with on average 2.37 labels per document.

[7] These models were trained using CORD-19 data available as of April 6th, 2020.

[8] http://pages.cs.wisc.edu/~andrzeje/research/zl_lda.html

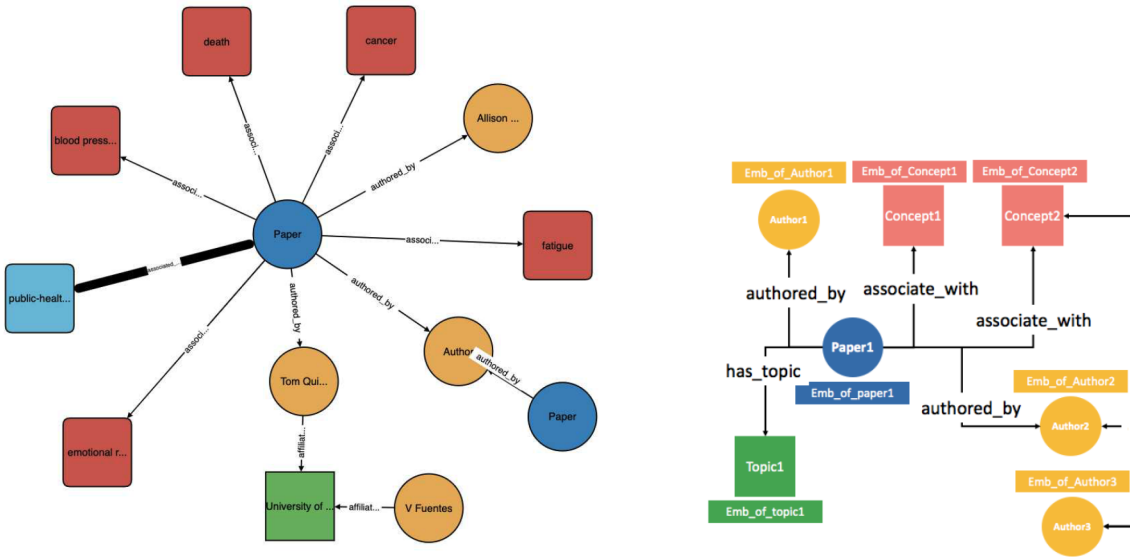

Figure 3: Visualization of COVID-19 Knowledge graph and Knowledge Graph Representation.

Query:	coronavirus origin
Question:	what is the origin of COVID-19?
Narrative:	seeking range of information about the SARS-CoV-2 virus's origin, including its evolution, animal source, and first transmission into humans.

Table 1: Example topic from TREC-COVID challenge.

Fewer than 1% of the documents in the test set received no label, using 0.5 as the confidence threshold.

Evaluation

ACS supports document ranking (DR), passage ranking (PR) and question answering (QA) using CORD-19 corpus. Similarly, Google COVID-19 Research Explorer (COVID-19 RE) and Neural Covidex (Covidex) (Zhang et al. 2020) are another two semantic searching web engines that facilitate DR and answer highlighting for COVID-19 questions. COVID-19 RE is powered by BERT (Devlin et al. 2019) and Covidex is based on T5-base model (Nogueira, Jiang, and Lin 2020). Both incorporate the latest NLP techniques to perform advanced semantic search. In this section, we compare the overall performance of ACS against COVID-19 RE and Covidex for their DR, PR and QA components, respectively.

To assess DR performance, we use the TREC-COVID challenge track (Voorhees et al. 2020), which contains 40 topic sets along with their document relevance judgement. The topic sets are written by its organizers with biomedical training, and motivated by search submitted to the National Library of Medicine and social media. As shown in Table 1, each topic consists of three fields with different levels of granularity, a keyword-based query (KQ), a more precise natural language question (NQ), and a longer descriptive narrative. The DR annotation are performed following the TREC pooling mechanism. Hundreds participants submitted ranked lists of documents for each topic set, based on which a depth of 10 to 20 documents are pooled and combined as a collection of (q, D) pairs. The pairs are then assessed by annotators with in-domain clinical expertise. More explicitly, the assessors are given a topic set and a list of documents to be judged. The assessors mark each document in the list as either 'Relevant', 'Partially Relevant' or 'Not Relevant'. There are three rounds of judgement available that corresponding to three different versions of CORD-19 corpus. To ensure sufficient coverage of annotation, we aggregate all rounds results into 33,064 (q, D) relevance judgements. Since the document id may change across versions, we map ids from each round to the May 19 release of CORD-19 corpus.

To collect DR results from the three systems, we crawled the top 50 articles by querying the engines with KQ and its NQ variation from the topic sets on June 15, 2020. The crawled data, which are characterized by the article title and link, are mapped to May 19 CORD-19 corpus. Note that articles that cannot be found in the corpus are removed to ensure fair comparison.

We use the standard DR metrics in our evaluation, namely, the precision and recall at k (P@k, R@k) and normalized discounted cumulative gain in the top k documents (NDCG@k). Note that we evaluate with k up to 20 since TREC-COVID has a pooling depth with 20 at most. Table 2 presents the DR performance of the engines over KQ and NQ, respectively. ACS performs consistently better than the

Search Engine	P@1	P@5	P@10	P@20	R@10	R@20	ndcg@20
Keyword Queries							
ACS	0.5250	**0.5650**	**0.5325**	**0.4775**	**0.0260**	**0.0459**	**0.4380**
Covidex	0.3421	0.2316	0.2079	0.1658	0.0109	0.0173	0.1633
COVID-19 RE	**0.5750**	**0.5650**	0.4775	0.4412	0.0236	0.0429	0.4022
Natural Language Questions							
ACS	**0.8750**	**0.7000**	**0.6400**	**0.5550**	**0.0345**	**0.0582**	**0.5357**
Covidex	0.4750	0.4800	0.4225	0.3625	0.0204	0.0356	0.3229
COVID-19 RE	0.6000	0.5300	0.4925	0.4600	0.0267	0.0474	0.4133

Table 2: Evaluation results on TREC-COVID dataset

2*Search Engine	Top3		Top30	
	EM	F1	EM	F1
ACS	**11.7**	**35.6**	**26.0**	**50.4**
Covidex	0.90	18.9	3.60	27.6
COVID-19 RE	10.0	31.8	18.2	43.8

Table 3: Top results variation from KQ to NQ on TREC-COVID dataset

other engines on NQ and mostly on KQ, and all engines perform better on NQ comparing with KQ.

In addition, we are able to evaluate how robust each system is against query variation from KQ to NQ. Ideally, the results shall remain unchanged with query variation that requests the same information. We define exact match (EM) and F1 score among top k results to evaluate the robustness. Let Q be the topic sets, and NQ_q^k, KQ_q^k denote the top k searching results of natural language question and keyword query corresponding to the same topic $q \in Q$, respectively. As shown in Eq.(1), we take the top k results KQ_q^k as ground truth, and compute the average exact match and F1 score of the top k results NQ_q^k, and then average over all queries. The article title string is used as comparison key, and EM and F1 are standard that the maximum is taken over all the ground truth articles. Table 3 demonstrates that ACS has the best capability to provide consistent results with query variation.

$$\begin{aligned} &EM(NQ, KQ, k) \\ =&\frac{1}{|Q|}\sum_{q\in Q}\frac{1}{k}\sum_{i=1}^{k} em(NQ_q^k(i), KQ_q^k) \end{aligned} \tag{1}$$

Next, we evaluate the performance of PR and QA using CovidQA dataset (Tang et al. 2020). CovidQA contains 27 questions which are selected by in-domain volunteers as the most promising COVID-19 research directions. The CovidQA answers are from judgement of 124 question-article pairs only. To ensure a sufficient annotation coverage over the 3 engines, we leverage our internal annotation resources to make PR and QA judgement. As shown in Figure 4, we crawled the top 3 results characterized by the article, displayed passage and the highlighted text snippet on June 15, 2020 from ACS, Covidex and COVID-19 RE, respectively. After that, the annotators assess PR and QA in terms of whether the displayed passage contains relevant information and whether the highlighted text snippet answers the given question. To avoid bias, the crawled results are combined and shuffled randomly, therefore, the annotators do not have access to the source engine and the rank position information during the judgement. We allocate 2 annotators to make independent judgements, and take average of their assessment as the final label.

Table 4 presents precision of top PR and QA results. We use P@k instead of EM and F1 to evaluate QA since ACS highlights the answer while Covidex and COVID-19 RE highlight the sentence that contains the answer. Instead of matching the answer, it is more reasonable to judge whether the highlighted text answers the question and compute precision. ACS achieves better accuracy on both PR and QA on most metrics. Note that ACS highlights answer snippet for at most three passages , while Covidex and COVID-19 RE highlight all displayed passages. This explains why ACS P@3 underperforms COVID-19 RE. Explicitly with an example, Figure 5 displays the topmost results of querying "What is the incubation period of virus" from the three systems. ACS highlights the answer in the passage. In contrast, Covidex presents an article published at 2004 with information of virus which is irrelevant with coronavirus, and COVID-19 RE does not answer the question at all.

As evident from the above results, ACS is one of the top-performing systems that provides high quality informative results over CORD-19 search.

Paper recommendation

CKG consists of multiple type of nodes related to author, affiliation, topics, entities etc. Knowledge Graph Embedding(KGE) models generate embeddings solely by taking into account the structure of the graph. In order to capture semantic information across the CORD-19 scientific articles we leverage SciBERT to generate Semantic embeddings. Thus, for each paper we captured both semantic as well as KGE. (Wise et al. 2020) provides the details around the creation and analysis of Covid Knowledge Graph. Scientific article recommendations are made possible by a document similarity engine that quantifies similarity between docu-

Query	Doc	passage	Is the passage relevant	Is Answer Highlighted
What is the proportion of patients who were asymptomatic?	Clinical characteristics of asymptomatic and symptomatic patients with mild COVID-19	... cases of COVID-19 was noted in early reports on the outbreak [4,5] and a large study on **72 314 Chinese patients reported that 1% were asymptomatic [6].** However, the majority of those cases were included from Hubei province...	Yes	Yes
	Clinical progression of patients with COVID-19 in Shanghai, China	...two repeated tests. The estimated median time from initiation of symptoms to PCR negative was 11 days (95 CIs: 10-12 days) in all the **patients**. In **asymptomatic patients**, PCR turned to be negative 2(1-3) days...	No	No

Figure 4: PR and QA annotation against CovidQA queries and top 3 engine results.

2*Search Engine	PR			QA		
	P@1	P@2	P@3	P@1	P@2	P@3
ACS	**0.4074**	**0.5370**	**0.4938**	**0.3333**	**0.2593**	0.2099
Covidex	**0.4074**	0.4444	0.4074	0.1481	0.1852	0.2099
COVID-19 RE	**0.4074**	0.3704	0.3580	0.2963	0.2407	**0.2593**

Table 4: Passage ranking (PR) and question answering (QA) performance on CovidQA dataset

ments by combining semantic embeddings obtained from a pre-trained language model (Beltagy, Lo, and Cohan 2019) with document knowledge graph embeddings (Wang et al. 2017) (Zheng et al. 2020) capturing topological information from the CKG. Due to lack of supervised signals for selecting the paper recommendation model, we used topic similarity between the source and the top $top - 5$ similar papers to select the right embeddings. From the Table 5, we note KGE embedding achieves a comparatively lower score than RGCN. Finally, the combination of semantic and KGE embeddings achieves the lowest Jaccard score.

Table 5: Topic similarity (Wise et al. 2020) (Jaccard distance) of recommendations vs random baseline.

Method	**Topic Distance**$_{Jaccard}$
Random	.821
Semantic$_{Sem}$	.360
Graph$_{KGE}$	.345
Graph$_{RGCN}$	.654
Sem. & KGE	.311

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 6 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 6 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.

Limitations and Future Directions

AWS CORD-19 is an initial step towards helping medical researchers find relevant content in a 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 clicks as well as the ranking of the results which were selected by medical researchers. Explicit feedback evaluation is cap-

ACS	Covidex	COVID-19 RE
Importance of Social Distancing: Modeling the spread of 2019-nCoV using Susceptible-Infected-Quarantined-Recovered-t model	Deadly viral syndrome mimics	Prediction of the virus incubation period for COVID-19 and future outbreaks
"Studies on the nature of the **virus** have suggested different **incubation periods** of the **virus**, and reports have suggested a median **incubation period of 5-6 days** and a very high symptom probability **period** of 14 days [5] "	"...**The incubation period is typically 3 to 14 days with the symptoms of an acute nonspecific, flulike illness developing suddenly**...**The incubation period is 7 to 10 days before the onset of symptoms [37]. This incubation period provides the potential for worldwide exposure because a person harboring the virus can expose the world at large via air travel**..."	"...while minimizing the negative consequences of the quarantine. 70 **71 The length of the incubation period varies both across and within virus families 4 . To our knowledge, 72 genomic features (if any) that correlate with the incubation**..."

Figure 5: Top-1 result of article title and displayed passage by querying 'What is the incubation period of virus?'

tured 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 generates summaries (Raffel et al. 2019) from paper abstracts and full body.

Conclusion

This paper describes our efforts in building AWS CORD-19 Search Engine with its capabilities consisting of document ranking, passage ranking and question answering. The deep semantic search model is leveraged to support querying with natural language questions. The search is further enhanced with topic modeling and knowledge graph. 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. By a systematic comparison with other public semantic searching engines that utilize the advanced NLP techniques and support similar component over COVID-19 search, we have demonstrated that ACS is one of the top-performing engines and is powerful on both its document ranking and question answering components. 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 access the scientific literature more easily and broadly.

Acknowledgments

We acknowledge the broader collaboration with AI2 team and White house as well as the broader AWS CORD-19 team including Kendra Science team, Tyler Stepke, Kingston Bosco, Victor Wang, Vaibhav Chaddha, Miguel Calvo, Ninad Kulkani, Kevin Longofer, Ray Chang, Adrian Bordone, Tony Nguyen, and Kyle Johnson.

References

Andrzejewski, D., and Zhu, X. 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*, 43–48. Association for Computational Linguistics.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

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

Bhatia, P.; Celikkaya, B.; and Khalilia, M. 2018. Joint entity extraction and assertion detection for clinical text. *arXiv preprint arXiv:1812.05270*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

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

Query	Article	Response
"What *medications* were most beneficial in the SARS outbreak?"	Development of chemical inhibitors of the SARS coronavirus: Viral helicase 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 for learn lessons and control the outbreak."

Figure 6: Sample queries demonstrating the semantic understanding of ACS, the use-fullness of Comprehend Medical NERe, and the utility of topic modeling for filtering.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)* 17–24.

Guan, W.-j.; Ni, Z.-y.; Hu, Y.; Liang, W.-h.; Ou, C.-q.; He, J.-x.; Liu, L.; Shan, H.; Lei, C.-l.; Hui, D. S.; Du, B.; Li, L.-j.; Zeng, G.; Yuen, K.-Y.; Chen, R.-c.; Tang, C.-l.; Wang, T.; Chen, P.-y.; Xiang, J.; Li, S.-y.; Wang, J.-l.; Liang, Z.-j.; Peng, Y.-x.; Wei, L.; Liu, Y.; Hu, Y.-h.; Peng, P.; Wang, J.-m.; Liu, J.-y.; Chen, Z.; Li, G.; Zheng, Z.-j.; Qiu, S.-q.; Luo, J.; Ye, C.-j.; Zhu, S.-y.; and Zhong, N.-s. 2020. Clinical characteristics of coronavirus disease 2019 in china. *New England Journal of Medicine* 382(18):1708–1720.

Nogueira, R.; Jiang, Z.; and Lin, J. 2020. Document ranking with a pretrained sequence-to-sequence model. *ArXiv* abs/2003.06713.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2011. Classifier chains for multi-label classification. *Machine learning* 85(3):333.

Rotmensch, M.; Halpern, Y.; Tlimat, A.; Horng, S.; and Sontag, D. 2017. Learning a health knowledge graph from electronic medical records. *Scientific reports* 7(1):1–11.

Singh, G., and Bhatia, P. 2019. Relation extraction using explicit context conditioning. *arXiv preprint arXiv:1902.09271*.

Tang, R.; Nogueira, R.; Zhang, E.; Gupta, N.; Cam, P.; Cho, K.; and Lin, J. 2020. Rapidly Bootstrapping a Question Answering Dataset for COVID-19. *arXiv preprint arXiv:2004.11339*.

Voorhees, E.; Alam, T.; Bedrick, S.; Demner-Fushman, D.; Hersh, W. R.; Lo, K.; Roberts, K.; Soborof, I.; and Wang, L. L. 2020. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. *arXiv preprint arXiv:2005.04474*.

Wang, Q.; Mao, Z.; Wang, B.; and Guo, L. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering* 29(12):2724–2743.

Wang, L. L.; Lo, K.; Chandrasekhar, Y.; Reas, R.; Yang, J.; Eide, D.; Funk, K.; Kinney, R.; Liu, Z.; Merrill, W.; et al. 2020. Cord-19: The covid-19 open research dataset. *arXiv preprint arXiv:2004.10706*.

Wise, C.; Ioannidis, V. N.; Calvo, M. R.; Song, X.; Price, G.; Kulkarni, N.; Brand, R.; Bhatia, P.; and Karypis, G. 2020. Covid-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature.

Zhang, E.; Gupta, N.; Nogueira, R.; Cho, K.; and Lin, J. 2020. Rapidly deploying a neural search engine for the covid-19 open research dataset: Preliminary thoughts and lessons learned. *arXiv preprint arXiv:2004.05125*.

Zheng, D.; Song, X.; Ma, C.; Tan, Z.; Ye, Z.; Dong, J.; Xiong, H.; Zhang, Z.; and Karypis, G. 2020. Dgl-ke: Training knowledge graph embeddings at scale. *arXiv preprint arXiv:2004.08532*.