An Evaluation of Publicly Available Deep Learning Based Commercial Information Retrieval Systems to search Biomedical Articles related to COVID-19

Sarvesh Soni, Kirk Roberts
School of Biomedical Informatics
University of Texas Health Science Center at Houston
Houston TX, USA
{sarvesh.soni, kirk.roberts}@uth.tmc.edu

Abstract
The COVID-19 pandemic has resulted in a tremendous need for access to the latest scientific information, primarily through the use of text mining and search tools. This has led to both corpora for biomedical articles related to COVID-19 (such as the CORD-19 corpus (Wang et al., 2020)) as well as search engines to query such data. While most research in search engines is performed in the academic field of information retrieval (IR), most academic search engines—though rigorously evaluated—are sparsely utilized, while major commercial web search engines (e.g., Google, Bing) dominate. This relates to COVID-19 because it can be expected that commercial search engines deployed for the pandemic will gain much higher traction than those produced in academic labs, and thus leads to questions about the empirical performance of these search tools. This paper seeks to empirically evaluate two such commercial search engines for COVID-19, produced by Google and Amazon, in comparison to the more academic prototypes evaluated in the context of the TREC-COVID track (Roberts et al., 2020). To ensure a fair comparison, we limit the number of documents in the retrieved runs and also annotate additional documents. We find that the top-performing system from TREC-COVID on bpref metric performed the best among the different systems evaluated in this study on all the metrics.

1 Background and Significance
There has been a surge of scientific studies related to COVID-19 due to the availability of archival sources as well as the expedited review policies of the publishing venues. A systematic effort to consolidate the flood of such information content, in the form of scientific articles, along with studies from the past that maybe relevant to COVID-19 is being carried out as requested by the White House (Wang et al., 2020). This effort led to the creation of CORD-19, a dataset of scientific articles related to COVID-19 and the other viruses from coronavirus family. One of the main aims for building such a dataset is to bridge the gap between machine learning and biomedical expertise to surface insightful information from the abundance of relevant published content. The TREC-COVID challenge was introduced to target the exploration of the CORD-19 dataset by gathering the information needs of biomedical researchers (Roberts et al., 2020; Voorhees et al., 2020). The challenge involved an information retrieval (IR) task to retrieve a set of ranked relevant documents for a given query. Similar to the task of TREC-COVID, major technology companies Amazon and Google also developed their own systems for exploring the CORD-19 dataset. However, despite the popularity of these companies’ products, no formal evaluation of these systems is made available by the companies. Also, neither of these companies participated in the TREC-COVID challenge. In this paper, we aim to evaluate these two IR systems and compare against the runs submitted to TREC-COVID challenge to gauge the efficacy of what are likely high-utilized search engines.

2 Methods
2.1 Information Retrieval Systems
We evaluate two publicly available IR systems targeted toward exploring the COVID-19 Open Research Dataset (CORD-19)\textsuperscript{1} (Wang et al., 2020). These systems are launched by Amazon (CORD-19 Search\textsuperscript{2}) and Google (COVID-19 Research Explorer\textsuperscript{3}). We hereafter refer to these systems by

---
\textsuperscript{1}https://www.semanticscholar.org/cord19
\textsuperscript{2}https://cord19.aws
\textsuperscript{3}https://covid19-research-explorer.appspot.com
Table 1: Three example topics from Round 1 of the TREC-COVID challenge.

| Topic | Query | Question | Narrative |
|-------|-------|----------|-----------|
| 7     | serological tests for coronavirus | are there serological tests that detect antibodies to coronavirus? | looking for assays that measure immune response to coronavirus that will help determine past infection and subsequent possible immunity. |
| 10    | coronavirus social distancing impact | has social distancing had an impact on slowing the spread of COVID-19? | seeking specific information on studies that have measured COVID-19’s transmission in one or more social distancing (or non-social distancing) approaches. |
| 30    | coronavirus remdesivir | is remdesivir an effective treatment for COVID-19? | seeking specific information on clinical outcomes in COVID-19 patients treated with remdesivir. |

the names of their corporations, i.e., Amazon and Google. Both the systems take as input a query in the form of natural language and return a list of documents from the CORD-19 dataset ranked by their relevance to the given query.

Amazons system uses an enriched version of the CORD-19 dataset constructed by passing it through a language processing service called Amazon Comprehend Medical (ACM) (Kass-Hout and Snively, 2020). ACM is a machine learning-based natural language processing (NLP) pipeline to extract clinical concepts such as signs, symptoms, diseases, and treatments from unstructured text (Kass-Hout and Wood, 2018). The data is further mapped to clinical topics related to COVID-19 such as immunology, clinical trials, and virology using multi-label classification and inference models. After the enrichment process, the data is indexed using Amazon Kendra that also uses machine learning to provide natural language querying capabilities for extracting relevant documents.

Googles system is based on a semantic search mechanism powered by BERT (Devlin et al., 2019), a deep learning-based approach to pre-training and fine-tuning for downstream NLP tasks (document retrieval in this case) (Hall, 2020). Semantic search, unlike lexical term-based search that aims at phrasal matching, focuses on understanding the meaning of user queries for searching. However, deep learning models such as BERT require a substantial amount of annotated data to be tuned for some specific task/domain. Biomedical articles have very different linguistic features than the general domain, upon which the BERT model is built. Thus, the model needs to be tuned for the target domain, i.e., biomedical domain, using annotated data. For this purpose, they use biomedical IR datasets from the BioASQ challenges4. Due to the smaller size of these biomedical datasets, and the large data requirement of the neural models, they use a synthetic query generation technique to augment the existing biomedical IR datasets (Ma et al., 2020). Finally, these expanded datasets are used to fine-tune the neural model. They further enhance their system by combining term- and neural-based retrieval models by balancing the memorization and generalization dynamics (Jiang et al., 2020).

2.2 Evaluation

We use a topic set collected as part of the TREC-COVID challenge for our evaluations (Roberts et al., 2020; Voorhees et al., 2020). These topics are a set of information need statements motivated by searches submitted to the National Library of Medicine and suggestions from researchers on Twitter. Each topic consists of three fields with varying levels of granularity in terms of expressing the information need, namely, (a keyword-based) query, (a natural language) question, and (a longer descriptive) narrative. A few example topics from Round 1 of the challenge is presented in Table 1. The challenge participants are required to return a ranked list of documents for each topic (also known as runs). The first round of TREC-COVID used a set of 30 topics and exploited the April 10, 2020 release of CORD-19. Round 1 of the challenge was initiated on April 15, 2020 with the runs from participants and relevance judgements from organizers due April 23 and May 3, respectively.

4http://bioasq.org
We use the question and narrative fields from the topics to query the systems developed by Amazon and Google. These fields are chosen following the recommendations set forward by the organizations, i.e., to use fully formed queries with questions and context. We use two variations for querying the systems. In the first variation, we query the systems using the question field from the topics. In the second variation, we also append the narrative part of the topics to the question field while querying to provide more context.

As we accessed these systems in the first week of May 2020, the systems could be using the latest version of CORD-19 at that time (i.e., May 1st, 2020 release). Thus, we filter the list of returned documents and only include the ones from the April 10th release to ensure a fair comparison with the submissions to the Round 1 of TREC-COVID challenge. We compare the performance of these systems (by Amazon and Google) with the 5 top submissions to the TREC-COVID challenge Round 1 (on the basis of bpref scores). It is valid to compare Amazon and Google systems with the submissions from Round 1 because all these systems are built without using any relevance judgements from TREC-COVID.

Relevance judgements (or assessments) for TREC-COVID are carried out by individuals with biomedical expertise. The assessments are performed using a pooling mechanism where only the top-ranked results from different submissions are assessed. A document is assigned one of the three possible judgements, namely, relevant, partially relevant, or not relevant. We use relevance judgements from Rounds 1 and 2. However, even the combined scores from both the rounds may not ensure that the relevance judgements for top-n documents for both the evaluated systems exist. So, to create a level ground for comparison, we perform additional relevance assessments for the documents from evaluated systems that may not have been covered by the combined set of judgements from TREC-COVID. In total, 141 documents were assessed by 2 individuals who are also involved in performing the relevance judgements for TREC-COVID.

The runs submitted to TREC-COVID could contain up to 1000 documents per topic. Due to the restrictions posed by the evaluated systems, we could only fetch up to a total of 100 documents per query. This number further decreases when we remove the documents that are not covered as part of the April 10th release of CORD-19. Thus, to ensure a fair comparison of the evaluated systems with the runs submitted to TREC-COVID, we calculate the minimum number of documents per topic (we call it topic-minimum) across the different variations of querying the evaluated systems (i.e., question or question+narrative). We then use this topic-minimum as a threshold for the maximum number of documents per topic. This ensures that there are the same number of documents returned for each topic.

We use the standard measures in our evaluation as employed for TREC-COVID, namely, bpref (binary preference), NDCG@10 (normalized discounted cumulative gain with top 10 documents), and P@5 (precision at 5 documents). Here, bpref only uses judged documents in calculation while the other two measures assume the non-judged documents to be not relevant. Additionally, we also calculate MAP (mean average precision), NDCG, and P@10. Note that we can precisely calculate some of the measures that cut the number of documents at up to 10 since we have ensured that both the evaluated systems (for both the query variations) have their top 10 documents manually judged (through TREC-COVID judgements and our additional assessments as part of this study). We use the trec_eval tool\(^5\) for our evaluations, which is a standard system employed for the TREC challenges.

\(^5\)https://github.com/usnistgov/trec_eval

3 Results

The total number of documents used for each topic based on the topic-minimums are shown in the form of a box plot in Figure 1. Approximately, an average of 43 documents are evaluated per topic with a median number of documents as 40.5. This is another reason for using a topic-wise minimum
Table 2: Evaluation results after setting a threshold at the number of documents per topic using a minimum number of documents present for each individual topic. The relevance judgements used are a combination of Rounds 1 and 2 of TREC-COVID and our additional relevance assessments. Highest scores for the evaluated and TREC-COVID systems are underlined.

| System                | P@5  | P@10 | NDCG@10 | MAP  | NDCG | bpref |
|-----------------------|------|------|---------|------|------|-------|
| Amazon                |      |      |         |      |      |       |
| question              | 0.6733 | 0.6333 | 0.539 | 0.0722 | 0.1838 | 0.1049 |
| question + narrative  | 0.72  | 0.64  | 0.5583 | 0.0766 | 0.1862 | 0.1063 |
| Google                |      |      |         |      |      |       |
| question              | 0.5733 | 0.57  | 0.4972 | 0.0693 | 0.1831 | 0.1069 |
| question + narrative  | 0.6067 | 0.56  | 0.5112 | 0.0687 | 0.1821 | 0.1054 |
| TREC-COVID            |      |      |         |      |      |       |
| 1. sab20.1.meta.docs  | 0.78  | 0.7133 | 0.6109 | 0.0999 | 0.2266 | 0.1352 |
| 2. sab20.1.merged     | 0.6733 | 0.6433 | 0.5555 | 0.0787 | 0.1971 | 0.1154 |
| 3. UIowaS_Run3       | 0.6467 | 0.6367 | 0.5466 | 0.0952 | 0.2091 | 0.1279 |
| 4. smith.rm3         | 0.6467 | 0.6133 | 0.5225 | 0.0914 | 0.2095 | 0.1303 |
| 5. udel_fang_run3    | 0.6333 | 0.6133 | 0.5398 | 0.0857 | 0.1977 | 0.1187 |

rather than cutting off all the systems to the same level as the lowest return count (that would be 25 documents). Having a topic-wise cut-off allowed us evaluate the runs with maximum possible documents while keeping the evaluation fair.

The evaluation results of our study are presented in Table 2. Among the commercial systems that we evaluated as part of this study, the question plus narrative variant of the system by Amazon performed consistently better than any other variant in terms of all the included measures other than bpref. In terms of bpref, the question-only variant of the system from Google performed the best among the evaluated systems. Note that the best run from TREC-COVID challenge, after cutting off using topic-minimums, still performed better than the other four submitted runs included in our evaluation. Interestingly, this best run also performed substantially better than all the variants of both commercial systems evaluated as part of the study on all the calculated metrics. We discuss more about this system below.

4 Discussion

We evaluate two commercial IR systems targeted toward extracting relevant documents from the CORD-19 dataset. For comparison, we also include the 5 best runs from TREC-COVID challenge in our evaluation. We additionally annotate a total of 141 documents from the runs by the commercial systems to ensure a fair comparison between these runs and the runs from TREC-COVID challenge. We find that the best system from TREC-COVID in terms of bpref metric outperformed all the commercial system variants on all the evaluated measures including P@5, NDCG@10, and bpref, which are the standard measures used in TREC-COVID.

The commercial systems often employ cutting edge technologies, such as AMC and BERT used by Amazon and Google, while developing their systems. Also, the availability of technological resources such as CPUs and GPUs maybe better in industry settings than in academic settings. This follows a common concern in academia, namely that the resource requirements for advanced machine learning methods (e.g., GPT-3 (Brown et al., 2020)) are well beyond the capabilities available to the vast majority of researchers. However, instead these results demonstrate the potential pitfalls of deploying a deep learning-based system without proper tuning. The sabir (sab20.+) system does not use machine learning at all: it is based on the very old SMART system (Buckley, 1985) and does not utilize any biomedical resources. It is instead carefully deployed based on an analysis of the data fields available in CORD-19. Subsequent rounds of TREC-COVID have since overtaken sabir (based indeed on machine learning with relevant training data). The lesson, then, for future emerging health events is that deploying “state-of-the-art” methods without event-specific data may be dangerous, and in the face of uncertainty simple may still be best.

As evident from Figure 1, many of the documents retrieved by the commercial systems were...
not part of the April 10th release of CORD-19. We queried these systems after another version of the CORD-19 dataset was released. New sources of papers were constantly being added to the dataset alongside updating the content of existing papers and adding newly published research related to COVID-19. This may have led to the retrieval of more articles from the new release of dataset. However, for a fair comparison between the commercial and the TREC-COVID systems, we pruned the list of documents and performed additional relevance judgements.

5 Conclusion

We assessed the performance of two commercial IR systems using similar evaluation methods and measures as the TREC-COVID challenge. To facilitate a fair comparison between these systems and the top 5 runs submitted to the TREC-COVID, we cut all the runs at different thresholds and performed more relevance judgements beyond the assessments provided by TREC-COVID. We found that the top performing system from TREC-COVID on bpref metric remained the best performing system among the commercial and the TREC-COVID submissions on all the evaluation metrics. Interestingly, this best performing run comes from a simple system that is purely based on the data elements present in the CORD-19 dataset and does not apply machine learning. Thus, applying cutting edge technologies without enough target data-specific modifications may not be sufficient for achieving optimal results.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Ariel Herbert-Voss, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler,Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs].

Chris Buckley. 1985. Implementation of the SMART information retrieval system. Technical Report 85-686, Cornell University.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Keith Hall. 2020. An NLU-Powered Tool to Explore COVID-19 Scientific Literature.

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C. Mozer. 2020. Characterizing Structural Regularities of Labeled Data in Overparameterized Models. arXiv:2002.03206 [cs, stat].

Taha A. Kass-Hout and Ben Snively. 2020. AWS launches machine learning enabled search capabilities for COVID-19 dataset.

Taha A. Kass-Hout and Matt Wood. 2018. Introducing medical language processing with Amazon Comprehend Medical.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan MacDonald. 2020. Zero-shot Neural Retrieval via Domain-targeted Synthetic Query Generation. arXiv:2004.14503 [cs].

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R. Hersh. 2020. TREC-COVID: Rationale and Structure of an Information Retrieval Shared Task for COVID-19. Journal of the American Medical Informatics Association.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. ACM SIGIR Forum, 54:1–12.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jianguang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset. arXiv:2004.10706v2.