Literature Review Draft

The Surge in SARS-CoV-2 Literature:

For researchers, the task of finding appropriate literature can be challenging as there is an abundance of papers, journals and articles in existence, all of which may have potentially relevant information for their studies. This process increases in complexity when literature is published at an overwhelming rate, which has been the case during the SARS-CoV-2 pandemic (1). During the first 100 days of 2020, over 5000 documents were published which were related to COVID-19 (2) which is claimed to be the largest growth in the number of biomedical documents related to a specific subject ever recorded (1). When all new documents relating to COVID-19 are concatenated with all documents prior to 2020 which discuss similar themes, the result is the CORD-19 dataset which currently has a corpus of over 50,000 documents (3).

The CORD-19 Dataset:

CORD-19 was released on March 16, 2021 by a partnership of academic institutions and led by the Allen Institute for AI (3). While the dataset was introduced to promote experimentation of text mining and information retrieval techniques for COVID-19 literature, it also contains historical publications relating to a variety of coronavirus, most notable SARS and MERS (3). Since the first human coronaviruses were identified in the 1960s (4), there is an abundance of research contained in CORD-19 which could be considered of relevance to the current pandemic, as investigating past events can help to understand future outcomes. As a result, it has been reported that the dataset encompasses almost all coronavirus literature that is available online (5). More importantly, however, analysis of the dataset has revealed that the scope of CORD-19 expands further into areas of research unrelated to coronaviruses (5). With the inclusion of a large collection of unnecessary documents, the dataset does not fully solve the original problem that it was intended to. While an information retrieval system can be confident that the documents it is searching for are contained within the dataset, there is a plethora of superfluous documents which can increase the search time and cause degraded performances. Therefore, a recent study has emphasised the importance of enhancing CORD-19 by considering the use of Altmetrics and other online data sources (5). Altmetric is a start-up which monitors online activity related to publications, and assigns each one a real-value representing the extent to which it has been publicly discussed (6). With this in mind, the CoronaCentral resource was developed to augment the CORD-19 dataset, removing redundant literature and making the most of innovative tools to aid information retrieval.

The CoronaCentral Dataset:

The CoronaCentral tool combines the CORD-19 dataset with PubMed as well as papers regarding SARS-CoV and MERS-CoV, and applies machine learning techniques to leverage profitable information about the documents (1). This results in an enhanced corpus of 128,921 papers, which categorises documents by their topic, article type and Altmetric data (1). The categorising of documents by topic and article type was performed using BERT-based document multilabel classification and it is claimed that the model achieved a micro-F1 score of 0.68 (1). The creators of the tool, however, do acknowledge that the exact performance of each topic and article type can vary, stating that the "long-haul" topic is not very successful as there is little knowledge about this in the test set and in the scientific community as a whole (1).

Several important conclusions can be drawn from the CoronaCentral dataset which identify areas of future research. While the resource does an excellent job of categorising documents and consequently aiding navigation through the dataset for researchers, it does not provide a sufficient search tool. Currently, the search tool employs primitive keyword matching algorithms to retrieve appropriate documents, and so it is evident that there is room for improvement in this specific area. While information retrieval systems have been widely deployed in the wake of the SAR-CoV-2 pandemic, the majority of these have been utilized over the CORD-19 dataset in the TREC-COVID challenge (7). With the CoronaCentral dataset being a refined adaption of CORD-19, including several unique metrics such as topic and article type, it is clear that an investigation into the possibilities of an information retrieval (IR) system on this dataset has the potential to yield improved results.

TREC-COVID:

The task of IR on coronavirus literature has been widely researched following the pandemic, specifically the TREC-COVID challenge which aimed to encourage participants to construct appropriate IR systems to meet the demands of a pandemic (7). It proposed the task to build an ad hoc IR system deployed on the CORD-19 dataset over five competitive rounds. The first round consisted of 30 topics, with each subsequent round receiving five additional topics, influenced by numerous sources such as social media and medical library searches. Each topic has the three fields – a collection of keywords, a question-like query and a larger associated narrative. Expert judges annotated individual documents as "relevant", "partially relevant" or "not relevant" for each topic. The participants would submit the documents retrieved for each topic, and the document annotations were used to identify whether they were relevant to the topic. It is clear that the TREC-COVID challenge has several parallels with IR system for CoronaCentral and as a result, there are several key elements to note.

Firstly, the IR system is ad hoc as users must be able to retrieve documents instantly at the moment of querying. Secondly, the challenge provides an effective means of evaluating our IR system. While CoronaCentral is not annotated in terms of relevance to query, the documents can be reversely mapped to their corresponding documents in CORD-19, which the TREC-COVID challenge has annotated. Therefore, there is pre-defined method for analysing performance of the IR system over the CoronaCentral dataset. Despite this convenient evaluation method, Roberts et al (7) does acknowledge that the challenge has shortcomings, specifically that the rate of publications is too high for the judges to annotated every document. Nevertheless, a significant number of documents have been annotated and therefore, an effective evaluation can be carried out using the pre-defined topics.

TREC-COVID Solutions:

Due to the parallels drawn between TREC-COVID and the IR system I wish to build for the CoronaCentral dataset, it is beneficial to inspect successful solutions to the challenge and extract key strategies which produced outstanding results. The SLEDGE IR system was successful throughout the challenge, with the most promising results of the first round (2). This employs the very popular two-stage retrieval strategy, with the first stage selecting the initial documents, and the second stage applying a neural reranking policy with SciBERT to move the higher valued documents to the top of the results (2). The developers of the system assert that it is zero-shot so can be applied to any form of medical literature with careful filtering of the training collection (8).

Similarly, the COVIDEX tool achieved high-scoring results in several rounds of the TREC challenge, most notably triumphing in the third round at the top of the leaderboard (9). This solution made use of an elementary first stage retrieval using TF-IDF document embeddings, followed by applying a MonoT5 reranker (9). When evaluating a solution to the challenge, however, it is crucial not to consider the

performance metrics as the only factor which defines success. These solutions are aimed to be used by professionals in a research environment, and so latency is instrumental for an appropriate tool. While SLEDGE does not acknowledge this, COVIDEX refers to the fact that their reranking stage can be costly and consequently, it is only applied to the top 96 documents retrieved (9). On the other hand, SLEDGE could be favourable over COVIDEX as a resource, as its reranking model is SciBERT, which improves upon a simple BERT model as it is trained upon a collection of scientific documents (10) and theoretically should enhance any search over CORD-19 as it is a collection of clinical research papers.

Neural Reranking Methods:

It is crucial to consider the shortcomings of these tools in order to establish which methods can be employed to achieve optimal results in the desired environment. The CoronaCentral search tool must be able to provide ad hoc results almost instantaneously, something which is not often achievable with modern neural rerankers. Several technologies were implemented in TREC solutions aiming to tackle this problem such as CO-Search, which adopted a Siamese BERT model named SBERT (11). This particular model claims to provide a significant improvement over BERT with regards to computation time, alleging that the time taken to cluster 10,000 sentences can be reduced from 65 hours to 5 seconds with their model (12). In spite of this enhancement, the model will not outperform SciBERT on CoronaCentral as it has been trained to function over scientific datasets. Therefore, CO-Search is sacrificing accuracy in favour of performance, which is not desirable. It seems like an appropriate compromise would be a combination of the COVIDEX tool and the SLEDGE tool, applying SciBERT to rerank only the top set of documents retrieved in the first stage. Intuitively it seems that researchers will only be interested in the most highly ranked documents as the fundamental purpose of this search tool is to eliminate the need to examine many research papers.

Furthermore, SLEDGE illuminates the correlation between publication date and document relevance, making the suggestion to filter the older papers out of the dataset prior to ranking (2). The parallels between filtering by date and filtering by the most highly ranked documents in the first stage are unambiguous and inspires exploration into this approach. This concept of trying to predict document relevancy can be considered more deeply, with the idea of exploiting document metadata such as Altmetric scores.

Sequence-to-Sequence Models:

The consideration of typical encoder-only transformation such as BERT based models is important, but recent research into sequence-to-sequence models has suggested that they may be superior in solving specific problems, while simultaneously increasing computational efficiency (13). One key obstacle in the search for coronavirus literature is the lack training data available due to the field only being considered a priority from the beginning of 2020. The TREC-COVID challenge provided a collection of 50 sample queries and a set of judgements for a subset of documents in CORD-19, detailing whether they were relevant to a specific query, partially relevant or not relevant. There are 69318 judgements applicable to CORD-19 and of these, only 39075 judge documents within CoronaCentral. Considering there are 223533 documents in CoronaCentral, we only have judgements for 17.5% of the documents. In addition, this small percentage of judgements will be split up into training, validating and testing datasets, thus further reducing the amount of training data available. When a lack of training data available, it has been reported that sequence-to-sequence models can exceed the capabilities of a traditional BERT based model (13). This claim is supported by evidence gathered in the Covidex IR system which proves the advantage of sequence-to-sequence models such as T5 variations (9).

Alternative Neural Indexing Possibilities:

While numerous reranking possibilities are available, it is vital to note that they require a significant amount of pretraining and are also very resource intensive during the query stage. Once the first stage has retrieved a collection of documents, the model must asynchronously rerank each document which can increase latency and overall degrade the user experience. The TREC-COVID challenge was designed to be ad hoc, meaning that information must be available at an instant. Subsequently, various neural indexing strategies can be considered as alternatives to neural reranking. These neural indexing strategies are sequence-to-sequence models, which as previously conveyed, can be beneficial compared to BERT based models. This involves various forms of document expansion which can be carried out prior to the indexing stage. The Doc2Query model is trained using MS Macro and then tasked with predicting possible queries for each document (14). It will envision 10 queries per document and append them to the main body of the text (14). The fact that the model is trained using the MS Macro dataset highlights that it can be applied to almost any field, as this comprehensive dataset covers 500 thousand queries and corresponding documents from the Bing search engine (14). One avenue which could be explored further is the concept of further training the model on more specific, scientific datasets so that it can be more expertly aware of the nuances of academic writing. Nevertheless, the resource has been proven to enhance metrics significantly over basic retrieval methods, with its developers even stating that it can compete with state of the art neural rerankers (14). As the query expansion is carried out prior to indexing, the latency at query time is avoided, clearly illustrating the importance of such techniques. One problem, however, which is not addressed by the developers of the tool is the fact that it does not account for different styles of queries. The TREC-COVID challenge defines queries as topics, with three different sub topics named “title”, “description” and “narrative” (7) . Each form of query expands in length, with the title being keywords, the description being a corresponding question and the narrative adding additional details. The queries generated by Doc2Query are most similar to the descriptions provided by TREC-COVID, although in reality it could be anticipated that users would typically search for keywords. Therefore, one could argue that this technique would not transfer to a realistic situation. It would be constructive to study the queries submitted to CoronaCentral currently to gain an insight into the most common format.

In order to combat this issue, a different form of document expansion could be used such as DeepCT. While most techniques only consider the frequency of terms when constructing embeddings, DeepCT takes advantage of the contextual importance of terms (15). The model is trained on BERT, and assigns the terms in the text a weighting, which depicts their contextual importance (15) in the document. Therefore, once the documents are indexed, the IR system should capture the core topics discussed in the document. For example, it is noticeable that many abstracts begin with a generic sentence regarding the background of COVID-19. It would be desirable for DeepCT indexing to identify this as unimportance and assign it a low weighting, instead allowing the true topic of the document to be emphasised. The research into contextual weighting makes reference to query expansion using Doc2Query and alludes to the possibilities of future work combining them (15). For example, it would be intriguing to understand how the process of expanding text with contextual weightings can allow more relevant queries to be developed by Doc2Query.

As with many potential strategies, neural indexing techniques inevitably have flaws which must be scrutinized. It is noteworthy that the preprocessing time required to expand a large collection of documents like so can be very long. As such it may be beneficial to only consider expanding a subset of the documents, similarly to the previous section which discussed only applying neural rerankers to a subset of retrieved documents. By analysing the metadata associated with documents, such as

publication year and Altmetric score, it might be possible to make educated guesses about the relevance of documents and filter the dataset accordingly.

Analysing Metadata:

Lessons can be learned about the benefits of metadata in information retrieval by investigating earlier literature from the field, prior to the period of time when advanced language models were being considered. Early research concluded that the usage of document genre in information retrieval was still an area which required further inquiries (16). At this time, however, the potential of natural language processing had not yet been harnessed, and as a result, studies which aimed to evaluate the effectiveness of genre in information retrieval were unable to do so, as they could not classify documents in the first place (17). It seems appropriate to return to this area of research now with the advances in natural language processing allowing us to overcome this barrier. Coronacentral has already been successfully categorized, giving each document a set of topics. Therefore, it would be intriguing to pick up where previous work was forced to halt, in order to discover whether this additional genre information can improve IR.

This opens several threads of discussion when designing an IR system for CoronaCentral. In addition to the topic of each article, the dataset contains potentially beneficial metadata such as DOI, publication year, article type, whether it is a preprint and the journal in which it was published. There is very little literature which discusses such aspects of documents in information retrieval, especially during a pandemic. Evidently, this is an area of research which is still very new and has a great amount of potential, something I hope to take advantage of during this project.

It is vital not to rely too heavily on this metadata, however, which may be flawed due to the publication process during a pandemic. Reports have shown that academic journals have reduced the average time taken to publish an article by 49% since the discovery of COVID-19 (18). Obviously, researchers were impatient to make their work publicly known in order to aid the effort to combat COVID-19 which may suggest that certain shortcuts were taken to do so. For example, it may be harmful to consider a DOI as an indicator of document quality, since the process of obtaining one can be very time-consuming. Additionally, there has been a significant increase in the publication of preprint articles (18) and so any implementation must be mindful of this. On the other hand, there may not be a significant difference between the preprint papers and journal papers, since the formal peer review process has been accelerated during COVID-19 and so arguably does not increase the integrity of a paper the way that it did prior to the pandemic.

To summarise, the analysis of document metadata is a thought-provoking area which should be reviewed, notably due to the advances in classification of document topics. It was even suggested by the TREC-COVID specification to explore whether machine learning could be utilized to extract data from documents and improve the IR process (7). Nevertheless, it is crucial to understand how best to interpret this metadata given the circumstances in which the articles have been published.

Conclusions:

While there are several state-of-the-art techniques defined for information retrieval, this particular problem involves several intricacies which must be considered, in order to best find a solution which is appropriate for the uncharted, ever-evolving environment of a pandemic. Although the information retrieval stages are of paramount importance to deliver the most relevant documents to the user, one must first fully understand the nuances of the dataset in question. CoronaCentral can be considered an enhanced version of CORD-19, by filtering out the gratuitous documents and providing further information about the relevant ones. With careful analysis of this information, I plan to better

understand its behaviours and characteristics which may be counter-intuitive given the fact they were published during a global pandemic. This will hopefully allow me to filter the dataset further, before moving onto the IR system.

The overwhelming majority of TREC-COVID solutions adopt the modern two-staged approach to IR, with a retrieval and reranking process. The first stage will involve selecting the appropriate indexing technique, with a plan to investigate the possibilities of neural indexing such as Doc2Query or DeepCT. Despite the drawback of these technologies being that they involve a large amount of preprocessing, this should hopefully be alleviated by the fact that the dataset has been reduced by the previous filtering stage. It would also be valuable to understand the format of queries currently submitted to CoronaCentral, allowing the search tool to be adapted accordingly.

It has been claimed that neural indexing has the capability of improving upon state-of-the-art neural reranking models such as BERT. Therefore, I intend to analyse this myself and embrace the appropriate tactic. Since the introduction of BERT-based encoders, more recent advances in sequence-to-sequence models have been shown to outperform them in the reranking stage and so I will be researching this area as well.

My initial thoughts, however, are that the second reranking stage will be redundant following an effective first stage retrieval. The current CoronaCentral tool uses primitive keyword matching, and so by making use of the metadata indexing the documents in a sophisticated manner will hopefully yield a significant improvement in results. The main lesson learned from the TREC-COVID challenge is that the most successful solutions were also the simplest.

Sources

1. Lever J, Altman RB. Analyzing the vast coronavirus literature with CoronaCentral. Proc Natl Acad Sci. 2021 Jun 8;118(23):e2100766118.

2. MacAvaney S, Cohan A, Goharian N. SLEDGE: A Simple Yet Effective Baseline for COVID-19 Scientific Knowledge Search. ArXiv200502365 Cs [Internet]. 2020 Aug 3 [cited 2021 Oct 5]; Available from: http://arxiv.org/abs/2005.02365

3. Lu Wang L, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, et al. CORD-19: The Covid-19 Open Research Dataset. ArXiv. 2020 Apr 22;arXiv:2004.10706v2.

4. Kahn JS, McIntosh K. History and Recent Advances in Coronavirus Discovery. Pediatr Infect Dis J. 2005 Nov;24(11):S223.

5. Colavizza G, Costas R, Traag VA, Eck NJ van, Leeuwen T van, Waltman L. A scientometric overview of CORD-19. PLOS ONE. 2021 Jan 7;16(1):e0244839.

6. Adie E, Roe W. Altmetric: enriching scholarly content with article-level discussion and metrics. Learn Publ. 2013;26(1):11–7.

7. Roberts K, Alam T, Bedrick S, Demner-Fushman D, Lo K, Soboroff I, et al. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. J Am Med Inform Assoc. 2020;

8. MacAvaney S, Cohan A, Goharian N. SLEDGE-Z: A Zero-Shot Baseline for COVID-19 Literature Search. ArXiv201005987 Cs [Internet]. 2020 Oct 12 [cited 2021 Oct 29]; Available from: http://arxiv.org/abs/2010.05987

9. Zhang E, Gupta N, Tang R, Han X, Pradeep R, Lu K, et al. Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset. ArXiv200707846 Cs [Internet]. 2020 Jul 14 [cited 2021 Oct 12]; Available from: http://arxiv.org/abs/2007.07846

10. Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. ArXiv190310676 Cs [Internet]. 2019 Sep 10 [cited 2021 Oct 12]; Available from: http://arxiv.org/abs/1903.10676

11. Esteva A, Kale A, Paulus R, Hashimoto K, Yin W, Radev D, et al. CO-Search: COVID-19 Information Retrieval with Semantic Search, Question Answering, and Abstractive Summarization. ArXiv200609595 Cs [Internet]. 2020 Jun 16 [cited 2021 Oct 12]; Available from: http://arxiv.org/abs/2006.09595

12. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv190810084 Cs [Internet]. 2019 Aug 27 [cited 2021 Nov 4]; Available from: http://arxiv.org/abs/1908.10084

13. Nogueira R, Jiang Z, Lin J. Document Ranking with a Pretrained Sequence-to-Sequence Model. ArXiv200306713 Cs [Internet]. 2020 Mar 14 [cited 2021 Oct 12]; Available from: http://arxiv.org/abs/2003.06713

14. Nogueira R, Yang W, Lin J, Cho K. Document Expansion by Query Prediction. ArXiv190408375 Cs [Internet]. 2019 Sep 24 [cited 2021 Oct 12]; Available from: http://arxiv.org/abs/1904.08375

15. Dai Z, Callan J. Context-Aware Term Weighting for First Stage Passage Retrieval. SIGIR 2020 - Proc 43rd Int ACM SIGIR Conf Res Dev Inf Retr. 2020;1533–6.

16. Crowston K, Kwasnik B. Can Document-Genre Metadata Improve Information Access to Large Digital Collections. Libr Trends. 2003 Jul 12;52.

17. Muresan G, Smith CL, Cole M, Liu L, Belkin NJ. Detecting Document Genre for Personalization of Information Retrieval. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06). 2006. p. 50c–50c.

18. Horbach SPJM. Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19. Quant Sci Stud. 2020 Aug 1;1(3):1056–67.