Mitigating Data Shift of Biomedical Research Articles for Information Retrieval and Semantic Indexing

Nima Ebadi[1], **Anthony Rios,**[2] **and Paul Rad**[1,2]
[1]Department of Computer Science
[2]Department of Information Systems and Cyber Security
University of Texas at San Antonio
{nima.ebadi, anthony.rios, peyman.najafirad>}@utsa

Abstract

Researchers have explored novel methods for both semantic indexing and information retrieval of biomedical research articles. Moreover, most solutions treat each task independently. However, both tasks are related. For instance, semantic indexes are generally used to filter results from an information retrieval system. Hence, one task can potentially improve the performance of models trained for the other task. Thus, this study proposes a unified retriever-ranker-based model to tackle the tasks of information retrieval (IR) and semantic indexing (SI). Particularly, our proposed model can adapt to rapid shifts in scientific research. Our results show that the model effectively leverages task similarity to improve the robustness to dataset shift. For SI, the Micro f1 score increases by 8% and the LCA-F score improves by 5%. For IR, the MAP increases by 5% on average.

1 Introduction

The pandemic caused the rapidly evolving curation of scientific publications about COVID-19, resulting in an information crisis (Roberts et al., 2020). As a result, healthcare practitioners, policymakers, and other individuals fighting against COVID-19 require specialized information retrieval (IR) and semantic indexing (SI) systems to keep track of the ever-evolving literature landscape (Esteva et al., 2020; Wang et al.). Researcher's methods to address these tasks have generally focused on one task, IR or SI (Zhang et al., 2020; Colic et al., 2020).

IR and SI are related. For example, IR addresses the question, "what are the most relevant research papers, and why are they deemed relevant?" SI is essential to facilitate easy browsing and filtering of IR-retrieved manuscripts. For instance, if a user finds all relevant papers related to the question, "What vaccines are the most effective for COVID-19?", they can use SI to filter papers associated with a specific COVID-19 variant, e.g., MeSH terms on PubMed (Lipscomb, 2000). Hence, this paper proposes a novel architecture that jointly addresses both tasks.

There has been an array of research in IR and SI of biomedical documents. For instance, since 2012, there has been an annual competition where researchers compete to develop more accurate biomedical IR and SI methods (BioASQ[1]). The competition is essential to improve the National Library of Medicine's (NLM) infrastructure, which provides IR and SI systems for biomedical scientists and healthcare professionals to search for biomedical research articles. NLM manually indexes biomedical research articles with Medical Subject Headings (MeSH). MeSH terms are used for biomedical SI purposes (e.g., filtering search results), to facilitate hypothesis generation by biomedical scientists, and to help general knowledge discovery. Unfortunately, there are over 29 thousand MeSH terms. Thus manually identifying the subset of terms applicable to each article is difficult and expensive to complete in a timely manner. Hence, the competition has helped researchers introduce various methods for automated MeSH coding. For instance, many researchers have trained linear models, which still result in strong baselines (Liu et al., 2014; Rios and Kavuluru, 2015). For example, Liu et al. (2014) combined linear models with a learning-to-rank framework, which is still used today in combination with neural networks (Dai et al., 2020).

Similarly, BioASQ had a part in advancing biomedical IR systems. For instance, Pappas et al. (2020) used convolutional neural networks for biomedical snippet retrieval. Similar to BioASQ, recent IR efforts have focused on COVID-related IR as part of the annual TREC competition (TREC-COVID) (Roberts et al., 2020). For example, Soni and Roberts (2021) evaluated two commercial deep

[1]http://www.bioasq.org/

Proceedings of the Third Workshop on Scholarly Document Processing, pages 138–151
October 12–17, 2022.

learning IR systems on the TREC-COVID dataset, showing that both systems underperformed the expected results. Researchers have proposed other models beyond the commercial systems, including pre-trained transformer models for text ranking (Lin et al., 2020), along with zero-shot retrieval systems for COVID (MacAvaney et al., 2020). Some researchers have recently explored combining IR and SI. As an example, researchers have used an IR system as part of a KNN-based component of an ensemble model to improve MeSH identification (Liu et al., 2014; Dai et al., 2020). Nevertheless, to the best of our knowledge, no prior work has used SI to improve IR systems, especially in IR systems for COVID-related retrieval.

There are four major technical challenges with developing COVID-related IR and SI systems: sparse datasets, shifts in the data distribution, scale, and interpretability. The limited amount of labeled data and dynamic changes in the COVID-19 landscape has made it challenging to generalize IR and SI methods beyond the datasets used to train them (Shokraneh and Russell-Rose, 2020). Because information is quickly becoming outdated in research articles, understanding what is relevant is difficult for current IR methods. For example, expert human judgments did not identify 70% of the retrieved results as relevant (Voorhees et al., 2021). However, the manual assessment process is time-consuming. Therefore, it is important to improve current models and provide textual evidence for "why" it detected a document as relevant to facilitate easier manual assessments by human experts (Xun et al., 2019)—providing answers to "why" is useful, especially if we develop systems that work to help experts. For instance, Jin et al. (2018) shows that human indexers at NLM become significantly more efficient and accurate if they are provided semantically sensible associations between the input text and system outputs.

To address the technical challenges, we propose a specialized IR and SI approach that combines interpretability, multi-task learning, and a mechanism of using unlabeled data via self-supervised learning to improve model robustness. Overall, our model will allow for quick adaptation and robustness to the dataset shift problem, becoming suitable for the context of the pandemic. We summarize the major contributions of this paper below: **(1.)** We propose a novel interpretable, self-supervised, multi-task learning method to tackle the tasks of IR and SI COVID-19-related research articles. We devise a mechanism to train a unified retriever-ranker on a self-supervised *masked language modeling (MLM)*, SI, and an IR task. This joint training framework enables inter-document representation learning, quick adaptation to new changes in the data distribution, and interpretability, which we demonstrate to be important for the context of the pandemic. To the best of our knowledge, this is the first study to show the utility of joint training of SI and IR tasks—showing both tasks complement each other in a single model, not just one task helping the other one. **(2.)** We introduce a novel output layer transformation method that allows us to predict new concepts as they appear over time *without* retraining the model. **(3.)** Our study provides detailed quantitative and qualitative analysis of our model's interpretability and transfer learning components that highlight the dataset shift challenges of IR and SI tasks during a health crisis.

2 Related Work

Biomedical Semantic Indexing. NLM has collected biomedical literature from the last 150 years. As of 2020, the PubMed database contains about 30 Million biomedical journal citations. This number has risen from 12 Million citations in 2004 to 30 Million citations in 2020, having a growth rate of 4% per year. Through a laborious process, NLM curators fully examine every document and annotate it with a set of hierarchically organized terminologies developed by NLM called Medical Subject Headings (MeSH[2]) along with supplementary concepts for more fine-grained categorization (Papagiannopoulou et al., 2016). In 2019, more than 900K biomedical citations were added to PubMed and manually indexed to more than 29K MeSH concept categories[3].

Researchers have been trying to address biomedical natural language processing problems effectively for more than a decade, e.g. BioASQ (Tsatsaronis et al., 2015c), which has led to introduction of many models for IR and SI (Jin et al., 2018; Peng et al., 2016; Müller et al., 2017; Zavorin et al., 2016; Xun et al., 2019). A successful group of submissions involves deep learning models with substantial hand-coded features and supervision. DeepMesh (Peng et al., 2016), the best performing model in the BioASQ challenge, combines docu-

[2] https://www.nlm.nih.gov/mesh/meshhome.html

[3] https://www.nlm.nih.gov/pubs/techbull/mj18/brief/mj18_updates_2018_baseline_stats.html

ment to vector models with crafted features from the document and MeSH indexes, along with ensemble models fed by those features. Other deep learning approaches include UIMA concept extractor links (Peng et al., 2016), and AUTH, which also uses a document-to-vector approach with an ensemble of machine learning classifier (SVM) fed with document-MeSH features (Papagiannopoulou et al., 2016). Jin et al. (2018) and Xun et al. (2019) combined retrieval systems with deep recurrent neural networks and attention mechanism and also provide explainability for MeSH indexing decisions. The amount of hand-crafted features and supervision required for these models make it difficult to scale up as the biomedical databases change during pandemic crises (Foroughi Pour et al., 2020).

Most SI models are developed to perform well in normal situations across a broad range of biomedical concepts. Researchers evaluate SI models based on their overall performance on all major MeSH indices (Nentidis et al., 2019). In the pandemic situation, however, the focus of the literature has drastically shifted toward the specific concepts and subconcepts related to the current Coronavirus disease. The number of published documents related to Coronavirus has risen from a few articles per month to more than 10K articles in June 2020—roughly 1 out of every 11.5 citations are about Coronavirus these days. Chen et al. (2020) The rapidly growing and evolving literature on COVID-19 causes challenges for automatic SI models (Shokraneh and Russell-Rose, 2020). Previously introduced SI models are based on supervised learning approaches and heavily hand-coded features. Therefore, they require large amounts of labeled data for a specific concept to perform well. They also have challenges scaling up to newly introduced terminologies and sub-concepts. Hence, they are unsuitable for emergencies, like the ongoing health crisis. In this paper, we focus on measuring and improving shifts in this setting.

Biomedical Information Retrieval. As previously mentioned, BioASQ challenge (Tsatsaronis et al., 2015a) is the largest challenge for SI and IR. Since 2015. BioASQ have shared a set of question—answering-related datasets every year. IR systems work in two phases. First, a broad (simple) method is used to retrieve the initial candidate's articles, and the second stage is to re-rank the candidates using a more complex method. The re-ranking model is usually based on the cross-attention model and fine-tuned for the binary classification task (Nentidis et al., 2020). For the first stage, many researchers use BM25 (Rosso-Mateus et al., 2018; Almeida and Matos, 2020; Kazaryan et al., 2020; Pappas et al., 2020). Likewise, several methods have been developed for the second stage. Rosso-Mateus et al. (2020) developed a system that takes as input learns distance metric to match question-passage pairs. Specifically, they use siamese and triplet networks to create a novel similarity learning method using a max-margin approach.

To the best of our knowledge, our study is the first to combine the two specific tasks of extraction of semantic indexes (which is essentially a multi-label text classification into a set of pre-defined, hierarchically organized semantic indexes) and IR (ranking a list of documents based on their relatedness to a query)—two tasks for which high-quality annotation by human experts exists compared to other domains. Other multi-task learning benchmarks mostly combine text problems that take a single piece of text as input rather than multiple documents, such as masked language modeling, NLU, and text classification (Raffel et al., 2019; McCann et al., 2018). Semantic search studies use the pre-trained models on such single input text problems, then fine-tune and use the representation of the document along with a similarity function or task-specific layers to compute the similarity between mid-level representations from the pre-trained encoder. These approaches cause discrepancies between the operations required for pre-text and downstream tasks. Therefore, they may not leverage the transfer learning (Ratner et al., 2018) effectively.

The most similar work to this paper is by Liu et al. (2019) which combines binary text classification with an information retrieval task via a multi-task learning framework. However, our work differs from Liu et al. (2019) in three major ways. First, we focus on semantic indexing, which is multi-label and contains more than 29k classes. Hence, rather than assigning a binary class to an instance (yes vs. no), our method must be able to assign a set of classes. Moreover, training large-scale multi-label models requires substantially different methodological choices beyond what binary classification needs. Second, their work does not focus on the biomedical domain, particularly biomedical-related scientific documents. Third, most of their

work focuses on single sentences rather than complete documents. Because of the sequence length limitation of BERT (Beltagy et al., 2020) multi-document and long-document analysis is only feasible by truncating the text. Hence, our approach can scale beyond sentence-level tasks.

3 Datasets

This paper uses three datasets: BioASQ Tasks 8a and 8b dataset, CORD-19 (Wang et al., 2020), and TREC-COVID. We describe each dataset below:

BioASQ 8a and 8b. First, we use the SI and IR datasets that were part of the BioASQ 8a and 8b competitions. Specifically, we use the PubMed articles from BioASQ's (Tsatsaronis et al., 2015b) Task 8a dataset, which includes almost 15 million article abstracts and titles. We select 8M recent articles published from 2007 to 2019.

For IR, we use BioASQ's Task 8b dataset, which includes 3,243 questions paired with related article abstracts. We use validation sets for each task for hyperparameter tuning. We also use the validation dataset as a pretraining procedure for the COVID-19-related corpora. But, we ensure there is no overlap between these general sets and their corresponding COVID-19 datasets.

CORD-19. The models are trained and/or evaluated on the following three COVID-19 datasets corresponding to the three tasks: SI, IR, and Masked Language Modeling (MLM). For our Semantic Indexing task, we use CORD-19 dataset (Wang et al., 2020) which includes 200K research articles about Coronavirus published in peer-reviewed venues and archival services such as bioRxiv[4] and medRxiv[5]. We select CORD-19 articles whose MeSH indexes are manually annotated in PubMed. We crawl and collect each article's MeSH indexes. The COVID-related SI dataset contains 17K articles which we chronologically sort and split into 13.6K for training (the oldest 80%) and 3.4K for testing (the latest 20%)—the number of articles is less than 200k because NLM has not yet indexed many articles.

During a data crisis, such as what is occurring with COVID-19, it is likely that we will collect unlabeled data quickly. However, it is unclear how to best use the unlabeled data. In response to this issue, we add an unsupervised task of incorporating COVID-related information into our models. Specifically, we perform a self-supervised pre-text task similar to Masked Language Modeling in (Devlin et al., 2019) to introduce knowledge about the pandemic. The masked article is treated as a query, and masked tokens are selected from a list of COVID-19-related terms[6]. The model attempts to detect articles that include the masked term(s), which allows our model to learn context matching using intra- and inter-document information (Cohan et al., 2020). To train this task, we use the entire CORD-19 training dataset, even the articles that have not been indexed yet at the time of the experiments.

TREC-COVID (Information Retrieval). As for the COVID-19-specific IR task, we use the TREC-COVID dataset (Roberts et al., 2020), which is an IR dataset for question answering similar to BioASQ QA task 8b. TREC-COVID includes 50 topics as queries represented by (concept, question, narrative) tuples. It also includes a dataset of 191K candidate documents from CORD-19. Experts manually evaluated the relevance of 69,317 topic-document pairs and annotated with three labels: unrelated, partially related, and related. Our task is to return a list of related articles, which include the target answer using the topic assigned to each question and given question as a query. This task structure is the same as used in BioASQ IR task 8b.

4 Methods

Intuitively, our method reformulates the semantic indexing task as IR such that we can train a single model—with a single output layer—that can perform both indexing and retrieval. Furthermore, our method does not require learning class-level parameters, thus allowing it to adapt easily to changes in the data distribution. Specifically, our method has three main phases: 1. Given an input document, we query all similar PubMed articles using a robust IR approach (combining BM25 and document embeddings). 2. We generate document embeddings that combine information from the input document (query) with each candidate (similar) document returned in step 1 (the initial retrieval phase) 3. Finally, given the query-candidate joint embeddings, we introduce a novel output layer that can apply to both the Semantic Indexing (classification) and

[4] https://www.biorxiv.org
[5] https://www.medrxiv.org

[6] We have used the list of related terms published by NLM https://www.nlm.nih.gov/pubs/techbull/nd20/nd20_mesh_covid_terms.html

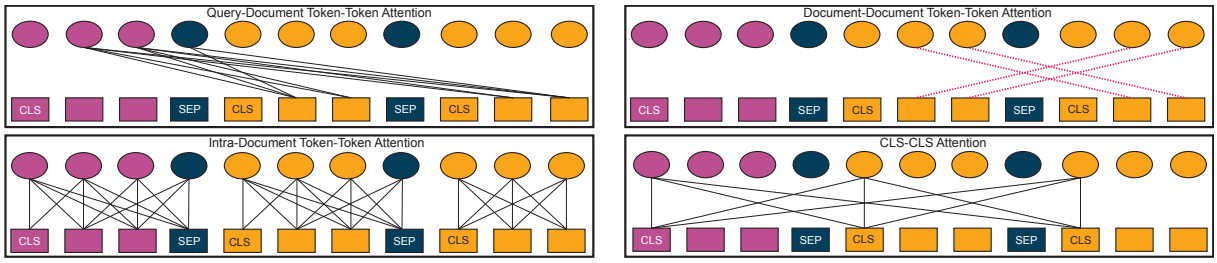

Figure 1: Intuitive attention modification diagram between the query Q (i.e., the purple items in the Figure) and candidates D (i.e., the orange items). The candidates (document-document) attentions are masked in our model.

IR tasks. Given the novel output layer, we take advantage of multi-task learning, jointly training the model for SI, IR, and additional tasks (self-supervised learning) to improve performance further. We describe each step in the following subsections.

Initial Retrieval. For the first stage, we use an initial retrieval system to identify a subset of related articles along with their task-specific annotations (for example, for extraction of semantic indexes, the task annotations include each candidate article's MeSH terms). Our initial retrieval system combines a document-level embedding model of SPECTER (Cohan et al., 2020) with a Bag-of-Words representation fused with BM25 following the schema of (Jin et al., 2018) and (Esteva et al., 2020). We initialize sPECTER with SciBERT (Beltagy et al., 2019) and trained on a bipartite graph of citations to capture document-level relatedness and minimize a triplet loss between related articles and maximize over unrelated ones. We further pre-train SPECTER on PubMed articles and fine-tune it on the COVID-19 dataset only. In addition, we use a BM25 weighted sum of article tokens to compute a keyword-based representation as well.

Formally, the input query and each candidate document are described as sequences of word tokens, denoted as $Q = \{q_i\}_{i=1}^{n}$ and $D = \{d_j\}_{j=1}^{m}$, respectively. For every candidate article, we also track associated metadata such as manually assigned MeSH terms defined as $L_D = \{l_j\}_{j=1}^{U_D}$, where U_D is the total number of MeSH terms assigned to candidate document C. We represent every article as an embedding $\mathbf{c} \in \mathbb{R}^z$ defined as

$$\mathbf{c}_d = \frac{\sum_{i=1}^{n} \text{Score}(w_i, D) \cdot \mathbf{v}_{w_i}}{\sum_{i=1}^{n} \text{Score}(w_i, D)} \qquad (1)$$

where z is the size of the SPECTER embedding, n is the number of tokens in document D, Score() $\in \mathbb{R}$ represents a token-level BM25 score, w_i is the i-th word in article d, and $\mathbf{v}_{w_i} \in \mathbb{R}^z$ is the token-level embeddings from the pre-trained model. Equation 1 is used to represent every document D which is used to represent both query $\mathbf{d}_Q$ and candidate $\mathbf{d}_D$ documents.

Next, we use the cosine similarity scores between each input query representation $\mathbf{d}_Q$ and every candidate article representation $\mathbf{d}_D$ in our database to find the top K most relevant articles $\mathcal{C} = \{D_1, \ldots, D_K\}$.

Transformer-based Representations and Reranking Next, given a query document Q and a set of candidate documents $\mathcal{C} = \{D_1, \ldots, D_K\}$, we use a BERT-like transformer model to rerank each candidate document D_i with respect to the input query. Specifically, we first concat the query Q with each candidate D_i to form a long sequence $[\text{CLS}, Q, \text{SEP}, \text{CLS}, D_1, \ldots, \text{SEP}, \text{CLS}, D_K]$, where each candidate is separated with a CLS and SEP token. Next, we predict a score for each candidate $\hat{y}_i = \sigma(\text{CLS}_{D_i})$, where σ represents a sigmoid function and CLS_{D_i} represents the CLS token directly preceding the start of candidate D_i's sequence of tokens.

At each level of the BERT representation, our input structure provides the ability to interpret that model in three unique ways using attention scores: token-to-token, token-to-document, and document-to-document. A high-level depiction of the attention scores is shown in Figure 1. First, the token-to-token scores between words within each query Q or within each candidate document D (i.e., the self-attention scores in Figure 1) calculates the importance of each word. For instance, the model can learn that the word "the" is unimportant for the downstream task. The *token-to-token* scores are also calculated between the tokens in the query and each candidate document (i.e., the token-to-token cross-attention scores in Figure 1), which can be interpreted as a similarity between the two words

across two documents regarding the downstream tasks.

Given that we care about relations between the query and candidates, but we do not care about token-to-token relations between two candidates, we mask the attention weights at each level of the BERT representation such that they are ignored. Next, the *CLS-to-token* attention is calculated between each token within a query or candidate document and the CLS representing each other document, which can be interpreted as the importance of how similar that token is regarding the topical content of another document. Finally, the *CLS-to-CLS* attention scores can be interpreted as a similarity score between each document—either between the query and each candidate or between each candidate, respectively. For instance, for semantic indexing of MeSH terms, the model should learn to give large CLS-to-CLS attention scores (for attention scores between the query and each candidate) to candidate documents with many MeSH terms that should be assigned to the query. Finally, because of the input sequence size, we use the Longformer model (Beltagy et al., 2020).

Output Layer Transformation. The output of the reranker transformer model is a set of sigmoid scores representing similarities between each candidate document and the input query. However, while these scores can directly be used to train the IR models to detect relevant documents for reranking purposes, we propose to use these identical scores to generate other types of output, such as MeSH code predictions for semantic indexing. Specifically, we propose a simple output layer transformation and training procedure to handle this task. Intuitively, our model is a Transformer-weighted k-NN, where scores of the scores for each "neighbor" is learned and contextual.

Formally, given a candidate score $\hat{y}_i$ for each candidate $D_j \in \mathcal{C}$, we generate a score for MeSH term as

$$\hat{l}_i = \sum_{j=1}^{K} \hat{y}_j \cdot \mathbb{1}[l_i \in L_j]$$

where $\hat{y}_j$ represents the sigmoid score for candidate $D_j \in \mathcal{C}$, l_i represents the j-th MeSH code, L_j represents the set of MeSH codes assigned to candidate D_j, and $\hat{l}_i$ is the final prediction score for MeSH code l_i with respect to the input query Q. At inference time, we optimize the thresholds to maximize the micro-f1 score (Pillai et al., 2013). For the SSL task, we generate scores for each of the COVID-related terms that are masked within the candidate documents.

To train the model, we first sample a task randomly, then sample training instances for the task, apply the output transformation, and train using Binary Cross-Entropy loss. For instance, for MeSH prediction, we train the model as

$$L = \sum_{i=1}^{U} l_i \log(\hat{l}_i) + (1 - l_i) \log(1 - \hat{l}_i)$$

where l_i represents the ground-truth label (1 or 0) for the i-th MeSH term and $\hat{l}_i$ is the prediction for the i-th term.

Note that for the IR task, we also train on relevance using binary cross-entropy. Hence, instead of using l_i as the ground-truth and $\hat{l}_i$ as the prediction, we use $\hat{l}_j$ and l_j, where $\hat{l}_j$ is the sigmoid output described in Section 4 that scores the relevance between the query and the j-th candidate and l_j is the ground-truth relevance (1 if relevant, 0 otherwise). Overall, this output transformation procedure has two major advantages. First, we do not need to learn any label-specific parameters. Many MeSH terms appear infrequently. As new MeSH terms are added, models must be retrained to predict them. However, our method can hypothetically predict terms as soon as new terms are used to annotate existing documents *without* retraining the model. Second, the output layer can predict any meta-data manually assigned or computed (as is the case for the SSL task) to the candidate database instances.

5 Results

Evaluation Metrics. To evaluate the performance of SI we use two sets of evaluation measures; i) flat measures such as micro- and macro-f1 scores, and ii) hierarchical measures such as Lowest Common Ancestor F-measure (LCA-F) (Kosmopoulos et al., 2015) for which we leverage BioASQ suggested algorithm[7].

For evaluation of IR tasks, we leverage *trec_eval*, the evaluation metrics and algorithms provided by TREC-COVID [8]. The evaluation metrics include normalized discounted cumulative gain (nDCG@N), P@N, Mean Average Precision

[7] https://github.com/BioASQ/Evaluation-Measures/tree/master/hierarchical

[8] https://trec.nist.gov/trec_eval/

Model	Micro F1
Medical Text Indexer (MTI) (default)	.658
MTI (first line indes)	.649
Average top score	**.714**
R+TR (base) (full attention)	.553
R+TR (base) (w/o multi-task)	.660
R+TR (base) (w/ multi-task)	.667
R+TR (large) (w/o multi-task)	.698
R+TR (large) (w/ multi-task)	**.705**

(a)

Model	MAP
Average top score	**.464**
R+TR (base) (full attention)	.191
R+TR (base) (w/o multi-task)	.328
R+TR (base) (w/ multi-task)	.344
R+TR (large) (w/o multi-task)	.355
R+TR (large) (w/ multi-task)	**.410**

(b)

Table 1: Semantic Indexing (a) and Information Retrieval (b) performances of our models, Retriever and Transformer-based Ranker (R+TR), along with the baselines (best performing models of BioASQ Task 8a for SI, and Task 8b Phase A for IR). The baseline scores are the average of their provided Micro F1 and Mean Average Persision (MAP) for IR and SI, respectively. The results are averaged across all test batches. Our model Retriever and Transformer-based Ranker is abbreviated as R+TR.

(MAP), and Binary preference (Bpref) (Esteva et al., 2020).

Baselines. We compare to three baseline models: the default MTI, MTI first line index, and the top models from the BioASQ competition. We explore MTI with base and "first line" parameters. MTI is a pre-trained model that is for SI of biomedical articles by the US National Library of Medicine. The first line version is the current version used by NLM that partially automates the standard indexing process at the US National Library of Medicine before human annotators further fine tune the indexes. We also report the scores for the best BioASQ team in each batch as "Average top score". Finally, to compare state-of-the-art methods on the COVID data, we retrain Attention MeSH (Grishchenko et al., 2020).

Hyperparameters and Model Variations. We optimize hyperparameters using a held-out validation dataset. For the SI experiments, K (i.e., the number of relevant articles retrieved) is set to 512. For the IR experiments, we set K to 1024. We use two versions of our re-ranker, a longformer base version (4 layers, 256 hidden size, 8 heads) and a large version (6 layers, 512 hidden size, 8 heads). Furthermore, we evaluate different attention mechanism on the base model. We also experiment with a naïve full attention mechanism (R+TR (full attention)) to compare the effect of the specific attention mechanism suggested by (Beltagy et al., 2020)[9]. All hyperparameters were chosen using the validation data. Refer to the Appendix for a comprehensive list of hyperparameters we searched over in our experiments.

[9]R+TR (full attention) requires truncation of the input documents, resulting in poor performance. However, Longformer uses *dilated sliding window* attention to avoid truncation. Dilation and window sizes are the target hyperparameters here. See the appendix for results with various dilation parameters.

	All Training Data				Micro F1			
Model	LCA-F	MiF	MaF	Accu.	0%	5%	10%	20%
MTI (default)	.563	.730	.506	.491	.222	.332	.459	.564
MTI (first line indes)	.553	.722	.501	.507	.218	.309	.462	.578
Attention MeSH	.579	**.764**	.529	.558	.271	.396	.504	.619
R+TR (base) (w/o ssl & mt)	.540	.700	.492	.485	.307	.433	.504	.591
R+TR (base) (w/ ssl)	.552	.728	.506	.510	.380	.486	.616	.663
R+TR (base) (w/ ssl & mt)	.563	.755	.511	.523	**.485**	.592	.656	**.724**
R+TR (large) (w/o ssl & mt)	.562	.742	.502	.523	.363	.474	.559	.595
R+TR (large) (w/ ssl)	.597	**.777**	.532	.569	**.490**	.619	**.698**	**.733**
R+TR (large) (w/ ssl & mt)	.612	**.810**	.558	.586	**.564**	**.676**	**.741**	**.789**

Table 2: Semantic indexing performance of our proposed models in comparison with baselines and ablation studies. For ablation, we experiment with (w/) and without (w/o) self-supervised learning (ssl) and multi-task learning (mt). For evaluation, we use Micro F1 (MiF) and Macro F1 (MaF). The second half of the table shows the MiF score based on the size of the COVID-19 training dataset, ranging from 0% (zero-shot) to 20% (few-shot).

BioASQ Experiments (Non-COVID). We analyze several design decisions for our transformer-based ranking system, such as the effect of multi-task learning on the general datasets and experimentally compare our use of the masked attention mechanisms. We report the results of each design decision in Table 5aa for BioASQ SI Task 8a, and Table 1bb for the BioASQ IR Task 8b, respectively. Our model Retriever and Transformer-based Ranker is abbreviated as R+TR.

Overall, for both IR and SI, we find that the full attention mechanism requires truncating the input documents, resulting in poor performance. The multi-task learning improves the performance of IR without affecting the SI's performance. Such improvement is expected not only because of the ef-

fect of transfer learning but also because the SI task improves retrieval and reinforces the latent space to be closer to those of the semantic indexes which human experts believed to be a better representation. For the IR results, we find that the multi-task improvement is higher for larger versions of our ranker (.328→.344 vs. .355→.410), showing that the knowledge transfer capability increases with the size of the transformer model. We do not report the effect of the self-supervision results here because it is only for COVID-19 datasets and disregarded in our ablation analysis on the general data. However, its effect is analysed in the following sections. Overall, the SI results in match the top contestants in the BioASQ competition (.714 vs .705). This is substantial given the submissions use ensemble methods while we are just training a single model. Similarly, for the IR task, we do not match the best contestants. However, we will show in the next results sections our model generalizes better to out of domain data related to COVID-19.

COVID-19 Semantic Indexing Experiments. Table 2 shows the SI performance of our models and baselines on the COVID-19 SI test set. Results on the left side (All Training Data) show the performance of the models once trained on the entire COVID-19 SI training set. The baselines are similarly fine-tuned with the training data for fair comparison. Our proposed "R+TR(large) w/o SSL & MT" model (i.e., without self-supervied learning and without multi-task learning) performs similar to the state-of-the-art baselines without leveraging the proposed self-supervised task and multi-task learning with IR (.742 Micro F1 vs .764). However, when combined, each of these transfer learning techniques substantially improves the SI performance. Leveraging the self-supervised learning task contributes and multi-task learning (SI + IR) helps because R+TR models gets acquainted with the context of the novel pandemic and its distributions, improving the Micro F1 score to 0.810. Overall, this experiment supports our hypothesis that IR tasks with SI improves model performance, particularly for COVID-related data.

The right side of Table 2 shows the performances based on the size of the COVID-specific training data. We chronologically sort the data and train the SI models with a proportion of them from the beginning. As shown in Table 2 the partitions include:0% which represents the zero shot learning ability, 5%, 10%, and 20% denoting the few-shot learning. 0% represents a model only trained on the original BioASQ dataset (i.e., no COVID-specific data). Our large R+TR's zero-shot micro-f1 score is significantly higher than the baselines, by 0.32 on average. It achieves 97% of its optimum performance by using only 20% of the training data. Again, providing evidence that our SI + IR multi-task learning framework can adapt better across domains.

Model	nDCG@20	P@20	Bpref	MAP
top score	**.850**	**.876**	**.638**	**.473**
ranke#1 in nCDG@20	**.850**	**.876**	.637	.472
ranke#1 in P@20	**.850**	**.876**	.637	.472
ranke#1 in Bpref	.850	.870	**.638**	**.473**
ranke#1 in MAP	.850	.870	**.638**	**.473**
R+TR (base) (w/o ssl & mt)	.792	.838	.602	.455
R+TR (base) (w/ ssl)	.821	.856	.626	.468
R+TR (base) (w/ ssl & mt)	**.857**	.870	**.642**	.464
R+TR (large) (w/o ssl & mt)	.805	.849	.620	.457
R+TR (large) (w/ ssl)	.830	.861	.633	.475
R+TR (large) (w/ ssl & mt)	**.889**	**.891**	**.657**	**.492**
R+TR (base) (w/ ssl & mt) (w/ f.t.)	**.899**	**.915**	**.664**	**.506**
R+TR (large) (w/ ssl & mt) (w/ f.t.)	**.924**	**.946**	**.691**	**.523**

Table 3: Information retrieval performance of our model with and without pre-training on self-supervised and semantic extraction tasks.

Information Retrieval on COVID-19 Experiments.

Table 3 shows the IR performance of our models evaluated on TREC-COVID round 5 dataset.[10]. Our model trained without SSL and Multi-Task learning (R+TR (base) (w/o ssl & mt) was only trained on the BioASQ QA dataset (i.e., No COVID-specific data), hence, it shows inferior performance which is because of the inconsistencies between two tasks. However, leveraging SSL and multi-task learning, our base model beats the top nDCG@20 and Bpref scores. This shows how the proposed transfer learning framework improves model's ability to scale up to a new domain. Our large R+RT achieves significantly superior performance in every metric score.

To analyze the zero- and few-shot learning ability of our model, we fine-tune our SSL multi-task learning models with TREC-COVID dataset. We choose round 3 dataset for training which has 40 topics identical to the first 40 topics in round 5. This is because the competition stated from 30 topics in round 1 and every time added 5 topics for the next round. We leave the last 10 topics of round 5 for evaluation.

[10]See for other baselines https://ir.nist.gov/covidSubmit/archive.html

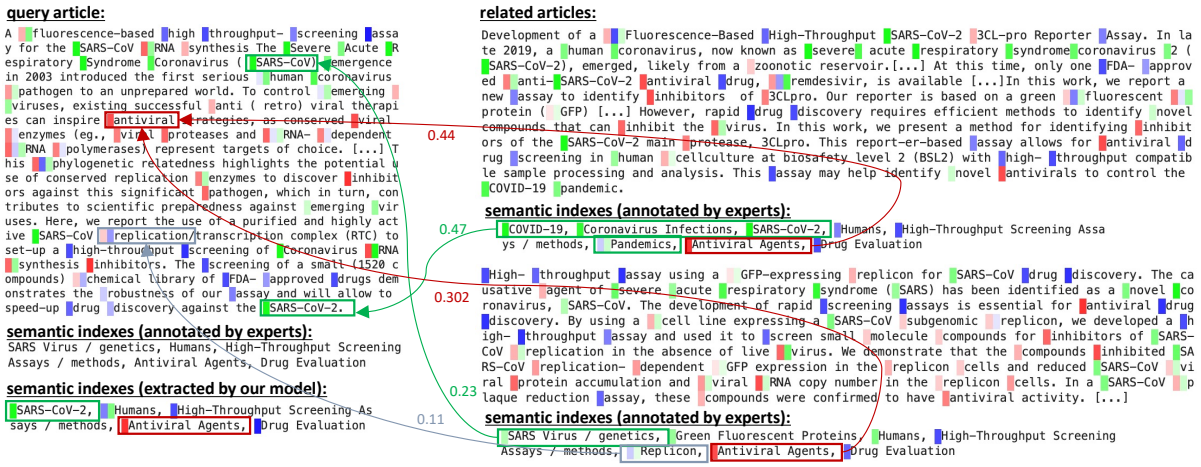

Figure 2: Illustration of attention weights between the input query and candidate articles along with the extracted outputs. The intensity and the color of the highlights denotes attention weights' values which is averaged and set to three scalar between the highly correlated terms.

We also expand the 40 samples of *(topic , set of candidate documents)* to 1,530 samples by randomly selecting a subset of 128 candidate documents for every given topic rather than 1024. As shown in the bottom of Table 3, our base and large model can leverage such fine-tuning and achieve significantly better scores than the top ones, by 0.05 MAP score. Note that in TREC-COVID challenge also participants could use results from previous rounds.

Interpretability.

As mentioned in Section 1, if our models improve human productivity, it is important for them to be interpretable. The interpretability can help human experts comprehend the decision making of a model and what has caused a mistaken output. As shown in Figure 2, the local-global attention of our model can assist human experts even when it makes an error by providing evidence for the mistaken output and suggesting other alternatives. The model extract the semantic index of SARS-CoV-2 while the manual annotator believes the article is about the general SARS viruses rather than a specific variant. Highlights in the figure show the global attention between the related articles and the query article, and the local attention within the query article. The weights are averaged and set to three scalar values, following (Sarker et al., 2019), to make the visualization simple (Lei et al., 2017). As depicted by Figure 2, the extraction of SARS-CoV-2 is because of the highly matched context about COVID-19 (the top related article) and the last sentence. However, the global attention provides another related article along with suggestions for the correct index. Knowing these, one can quickly identify and fix the error.

The interpretability can also help to understand the performance of the model in mitigating the challenges of COVID-19 infodemic. Please refer to A for more interpretability analysis.

6 Conclusion

In this study, we have unified the tasks of IR for question answering with the extraction of semantic indexes and with a self-supervised pre-text task. Our approach allows us to *simultaneously* train on downstream tasks and unlabeled data to maximize the advantages of transfer learning in addressing the data efficiency, generalization, and dataset shift issues. Compared to benchmarks, our model learns with less labeled data (it does not even need to learn class-specific parameters) and shows a substantially higher zero-shot (out-of-domain) performance. Overall, our study brings focus towards state-of-the-art remedies to the current challenges of the pandemic, which opens up new doors to a more systematic analysis of each of these challenges and more sophisticated algorithms.

As future research, we will look to combine more IR and SI-related tasks as more data is being annotated and prepared for the domain-specific environment of the pandemic. To better evaluate the performance of the global-local interpretability, we plan to perform qualitative analysis by providing this tool to human experts. The goal is for the tool to improve their time efficiency and perfor-

mance when they are performing manual indexing of biomedical research articles.

References

Tiago Almeida and Sérgio Matos. 2020. Bit. ua at bioasq 8: Lightweight neural document ranking with zero-shot snippet retrieval. In *CLEF (Working Notes)*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Q. Chen, A. Allot, and Z. Lu. 2020. Keep up with the latest coronavirus research. *Nature*, 579(7798):193.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282.

Nico Colic, Lenz Furrer, and Fabio Rinaldi. 2020. Annotating the pandemic: Named entity recognition and normalisation in covid-19 literature. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Suyang Dai, Ronghui You, Zhiyong Lu, Xiaodi Huang, Hiroshi Mamitsuka, and Shanfeng Zhu. 2020. Fullmesh: improving large-scale mesh indexing with full text. *Bioinformatics*, 36(5):1533–1541.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. 2020. Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization. *arXiv preprint arXiv:2006.09595*.

Ali Foroughi Pour, Maciej Pietrzak, Lori A Dalton, and Grzegorz A Rempała. 2020. High dimensional model representation of log-likelihood ratio: binary classification with expression data. *BMC bioinformatics*, 21:1–27.

Ivan Grishchenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. 2020. Attention mesh: High-fidelity face mesh prediction in real-time. *arXiv preprint arXiv:2006.10962*.

Zhiwen Hu, Zhongliang Yang, Qi Li, and An Zhang. 2020. The covid-19 infodemic: Infodemiology study analyzing stigmatizing search terms. *Journal of medical Internet research*, 22(11):e22639.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2018. Attentionmesh: Simple, effective and interpretable automatic mesh indexer. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 47–56.

Ashot Kazaryan, Uladzislau Sazanovich, and Vladislav Belyaev. 2020. Transformer-based open domain biomedical question answering at bioasq8 challenge. In *CLEF (Working Notes)*.

Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. 2015. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865.

Tao Lei et al. 2017. *Interpretable neural models for natural language processing*. Ph.D. thesis, Massachusetts Institute of Technology.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: Bert and beyond. *arXiv preprint arXiv:2010.06467*.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Ke Liu, Junqiu Wu, Shengwen Peng, Chengxiang Zhai, and Shanfeng Zhu. 2014. The fudan-uiuc participation in the bioasq challenge task 2a: The antinomyra system. In *CEUR Workshop Proceedings*, volume 1180, pages 1311–1318. CEUR-WS.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2020. Sledge: A simple yet effective zero-shot baseline for coronavirus scientific knowledge search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4171–4179.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Bernd Müller, Christoph Poley, Jana Pössel, Alexandra Hagelstein, and Thomas Gübitz. 2017. Livivo–the vertical search engine for life sciences. *Datenbank-Spektrum*, 17(1):29–34.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2019. Results of the seventh edition of the bioasq challenge. In

Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 553–568. Springer.

Anastasios Nentidis, Anastasia Krithara, Konstantinos Bougiatiotis, and Georgios Paliouras. 2020. Overview of bioasq 8a and 8b: Results of the 8th edition of the bioasq tasks a and b. In *CLEF(Working Notes)*.

Toan Q Nguyen and Julian Salazar. 2019. Transformers without tears: Improving the normalization of self-attention. *arXiv preprint arXiv:1910.05895*.

Eirini Papagiannopoulou, Yiannis Papanikolaou, Dimitris Dimitriadis, Sakis Lagopoulos, Grigorios Tsoumakas, Manos Laliotis, Nikos Markantonatos, and Ioannis Vlahavas. 2016. Large-scale semantic indexing and question answering in biomedicine. In *Proceedings of the Fourth BioASQ workshop*, pages 50–54.

Dimitris Pappas, Petros Stavropoulos, and Ion Androutsopoulos. 2020. Aueb-nlp at bioasq 8: Biomedical document and snippet retrieval. In *CLEF (Working Notes)*.

Shengwen Peng, Ronghui You, Hongning Wang, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2016. Deepmesh: deep semantic representation for improving large-scale mesh indexing. *Bioinformatics*, 32(12):i70–i79.

Ignazio Pillai, Giorgio Fumera, and Fabio Roli. 2013. Threshold optimisation for multi-label classifiers. *Pattern Recognition*, 46(7):2055–2065.

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*.

Alex Ratner, Braden Hancock, Jared Dunnmon, Roger Goldman, and Christopher Ré. 2018. Snorkel metal: Weak supervision for multi-task learning. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, pages 1–4.

Anthony Rios and Ramakanth Kavuluru. 2015. Analyzing the moving parts of a large-scale multi-label text classification pipeline: Experiences in indexing biomedical articles. In *2015 International Conference on Healthcare Informatics*, pages 1–7. IEEE.

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*.

Andrés Rosso-Mateus, Fabio A. González, and Manuel Montes-y Gómez. 2018. MindLab neural network approach at BioASQ 6B. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

Andrés Rosso-Mateus, Fabio A. González, and Manuel Montes-y Gómez. 2020. A deep metric learning method for biomedical passage retrieval. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6229–6239, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Abeed Sarker, Ari Z Klein, Janet Mee, Polina Harik, and Graciela Gonzalez-Hernandez. 2019. An interpretable natural language processing system for written medical examination assessment. *Journal of biomedical informatics*, 98:103268.

Farhad Shokraneh and Tony Russell-Rose. 2020. Lessons from covid-19 to future evidence synthesis efforts: first living search strategy and out of date scientific publishing and indexing industry (submitted). *Journal of Clinical Epidemiology*.

Sarvesh Soni and Kirk Roberts. 2021. An evaluation of two commercial deep learning-based information retrieval systems for covid-19 literature. *Journal of the American Medical Informatics Association*, 28(1):132–137.

G. Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, M. Zschunke, M. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, D. Polychronopoulos, Y. Almirantis, John Pavlopoulos, Nicolas Baskiotis, P. Gallinari, T. Artières, A. N. Ngomo, N. Heino, Éric Gaussier, L. Barrio-Alvers, M. Schroeder, Ion Androutsopoulos, and G. Paliouras. 2015a. An overview of the large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015b. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015c. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk

Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1).

Jingqi Wang, Huy Anh, Frank Manion, Masoud Rouhizadeh, and Yaoyun Zhang. Covid-19 signsym–a fast adaptation of general clinical nlp tools to identify and normalize covid-19 signs and symptoms to omop common data model. *ArXiv*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William. Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Guangxu Xun, Kishlay Jha, Ye Yuan, Yaqing Wang, and Aidong Zhang. 2019. Meshprobenet: a self-attentive probe net for mesh indexing. *Bioinformatics*, 35(19):3794–3802.

Ilya Zavorin, James Mork, and Dina Demner-Fushman. 2016. Using learning-to-rank to enhance nlm medical text indexer results. In *Proceedings of the Fourth BioASQ workshop*, pages 8–15.

Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang, et al. 2020. Covidex: Neural ranking models and keyword search infrastructure for the covid-19 open research dataset. *arXiv preprint arXiv:2007.07846*.

A Interpretability

As a case study, to analyze the performance of our model in handling the shift in the topics and terminologies of COVID-19 related literature, we look at attention weights between the various stigmatized and standard terms for the novel Coronavirus over the time. The stigmatized terms include those which have been used prior to the provisional standard term "2019-nCoV", such as "Wuhan Coronavirus," "Chinese Virus," "Wuhan Novel Pneumonia" to name a few (Hu et al., 2020). We use the aggregated weights[11] when these terms attend to or get attended by the standard ones (i.e. COVID-19 and SARS-CoV-2). We use the chronologically sorted dataset and looked at the weights as the model gets trained over the different time frames.

As shown in Figure 3, as the distribution of terminologies changes over time, the attention mechanism learns to relate to the well-established terms mitigating the effect of the dataset shift. In the beginning, the model shows high attention weights toward SARS-CoV as it is another variant of Coronavirus, which has also originated from China. This finding shows that the model matches the new context. Specifically, the model quickly relates stigmatized terms even prior to introducing their standard terms. With the standard terms, the model pays less attention to stigmatized and provisional terms. The attention over SARS-CoV-1 and other related variants decreases as the model dissolves the confusion between them.

[11] Summed and averaged over all sample queries and candidate articles, using both local and global attentions.

Hyperparameter	Value(s)
$\|V\|$	20M, **30M**
K	128, 256 , **512**, 1024
w (sliding window size)	32,..., 512, $\mathbf{inc}[32:512]$, $\mathrm{dec}[32:512]$
dilation	0, 1, 2, 3, $\mathbf{inc}[0:3]$
dilation heads	1, **2**, 3
dorpout	0.1, **0.2**, 0.3, **0.4***
batch size	8, **16**, 32, 64 (gpu memory limit)
output vector size	512, **1024**, 2048
w.e. size	128, **256**, **512***
hidden size	128, **256**, **512***
#layers	**4**, 5, **6***, 7, 8
learning rate	0.001, **0.0005**, 0.00025, **0.0001**

Table 4: Hyperparameter values. w.e.: embedding size for initial retrieval step. We use bold text for the optimal ones among all tried values. * refer to those for large ranker. Best dilation size is achieved by increasing it by 1 from first layer to the last.

B Hyperparameters

In Table 4, we list all of the hyperparameters we search over in this study. The best hyperparameters we found on the validation dataset are marked via bold and an asterisk (*). When training the transformer reranker model, we use a dropout value of 0.2, batch size of 16, 2 dilation heads, with a dilation varying from 0 to 3 from the first to last layer of the Longformer (increasing or decreasing every/every other layer).

C Dilation Results

In Table 5, we experiment with the longformer dilation parameter w, fixing it at 230, varying it from size 32 to 512 from the first to last layer, varying it from 512 to 32 from the first to last layer (i.e., in reverse), using dilation on two heads, and combining global dilation with dialated sliding windows. See Beltagy et al. (2020) for more details on the dilation parameters. Overall, we find that the combination of global and dilated sliding window

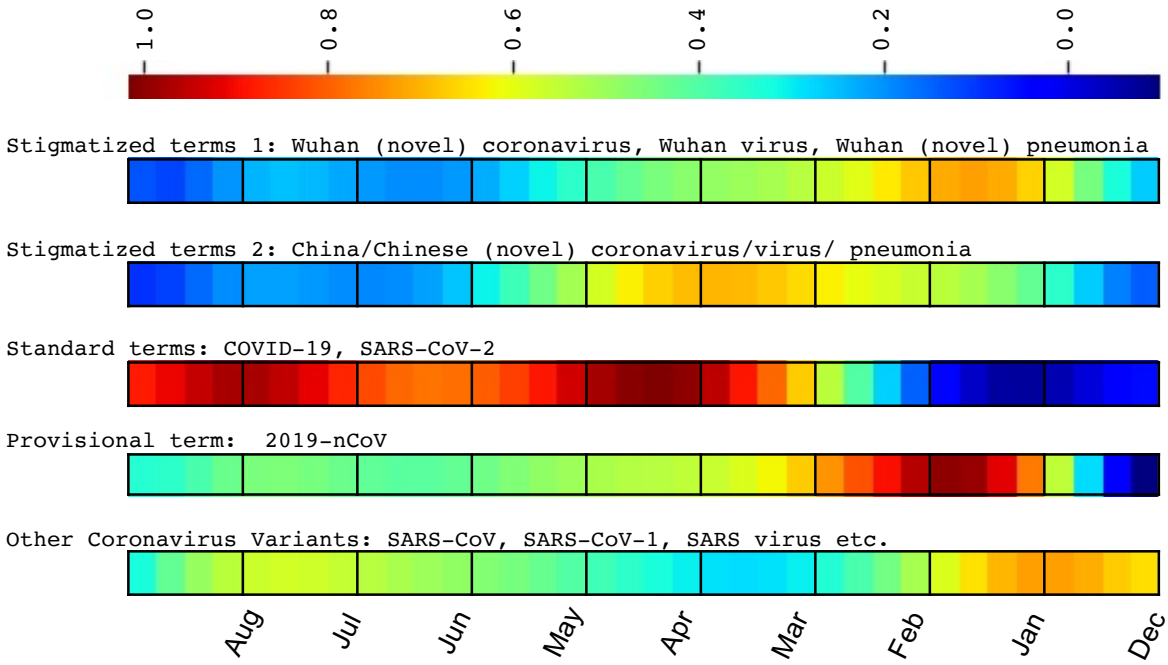

Figure 3: Attention weights of terms attending to COVID-19 and SARS-CoV-2 over different time frames. These weights are normalized for visualization purpose, following (Nguyen and Salazar, 2019)

with increasing window size shows better performance than other combinations in both IR and SI. However, the performance still does not match our custom attention filtering as shown in Tables 5a and 1b.

Model	Micro F1
R+TR (base) (full attention)	.553
R+TR (increasing w) (from 32-512)	.628
R+TR (fixed w) (=230)	.614
R+TR (decreasing w) (from 512-32)	.600
R+TR (increasing w) (dilation on 2 heads)	.633
R+TR (global + dilated sliding window*)	.660

(a)

Model	MAP
R+TR (base) (full attention)	.191
R+TR (increasing w) (from 32-512)	.293
R+TR (fixed w) (=230)	.280
R+TR (decreasing w) (from 512-32)	.258
R+TR (increasing w) (dilation on 2 heads)	.303
R+TR (global + dilated sliding window*)	.328

(b)

Table 5: Semantic Indexing (a) and Information Retrieval (b) performances of our models, Retriever and Transformer-based Ranker (R+TR), along with the baselines (best performing models of BioASQ Task 8a for SI, and Task 8b Phase A for IR). The baseline scores are the average of their provided Micro F1 and Mean Average Persision (MAP) for IR and SI, respectively. The results are averaged across all test batches. Our model Retriever and Transformer-based Ranker is abbreviated as R+TR.