A Search Engine for Discovery of Scientific Challenges and Directions

Dan Lahav[1], **Jon Saad Falcon**[2,4], **Bailey Kuehl**[2], **Sophie Johnson**[2], **Sravanthi Parasa**[6], **Noam Shomron**[1], **Duen Horng Chau**[4], **Diyi Yang**[4], **Eric Horvitz**[5], **Daniel S. Weld**[2,3], **Tom Hope**[2,3]

[1]Tel Aviv University, [2]Allen Institute for AI, [3]University of Washington,
[4]Georgia Institute of Technology, [5]Microsoft, [6]Swedish Medical Group

Abstract

Keeping track of scientific challenges, advances and emerging directions is a fundamental part of research. However, researchers face a flood of papers that hinders discovery of important knowledge. In biomedicine, this directly impacts human lives. To address this problem, we present a novel task of extraction and search of scientific challenges and directions, to facilitate rapid knowledge discovery. We construct and release an expert-annotated corpus of texts sampled from full-length papers, labeled with novel semantic categories that generalize across many types of challenges and directions. We focus on a large corpus of interdisciplinary work relating to the COVID-19 pandemic, ranging from biomedicine to areas such as AI and economics. We apply a model trained on our data to identify challenges and directions across the corpus and build a dedicated search engine. In experiments with 19 researchers and clinicians using our system, we outperform a popular scientific search engine in assisting knowledge discovery. Finally, we show that models trained on our resource generalize to the wider biomedical domain and to AI papers, highlighting its broad utility. We make our data, model and search engine publicly available.

1 Introduction

> "We can only see a short distance ahead, but we can see plenty there that needs to be done."
>
> *– Alan Turing, 1950*

Success in scientific efforts hinges on identifying promising and important problems to work on, developing novel and effective solutions, and formulating hypotheses and directions for further exploration. Each new scientific advance helps address gaps in knowledge, including potential extensions and refinements of prior results. New advances often lead to new challenges and directions. With millions of scientific papers published every year, sets of challenges and potential directions for addressing them grow rapidly. A striking recent example is that of literature pertaining to the COVID-19 pandemic (Wang et al. 2020), which exploded in unprecedented volume with researchers from across diverse fields exploring the many facets of the disease and its societal ramifications. As the pandemic continues worldwide, it is especially urgent to provide scientists with tools for staying aware of advances, problems, and limitations faced by fellow researchers and medical professionals, and of emerging hypotheses or early indications of potential solutions.

Unfortunately, due to the immense scale and siloed nature of the scientific community, it can be difficult for researchers to keep track of their own specialty areas, let alone discover relevant knowledge in areas outside their immediate focus (Hope et al. 2017, 2020, 2021; Portenoy et al. 2021). This can result in poor awareness of failures or limitations reported in recent studies, wasting redundant resources and leading to clinical decision-making uninformed about shortcomings of interventions (Chalmers et al. 2014). Disturbingly, there have been many cases where problems in treatments had been reported but not picked up by sectors of the clinical community (Clarke and Hopewell 2013; Robinson and Goodman 2011; Cooper, Jones, and Sutton 2005) leading to higher rates of morbidity and mortality (Ker et al. 2012; Gilbert et al. 2005; Sinclair 1995).

Our goal is to bolster the ability of researchers and clinicians to **keep track of difficulties, limitations and emerging hypotheses**. This could help clinical decision making be well-informed, accelerate innovation by surfacing new opportunities to work on, inspire new research directions, and match challenges with potential solutions from other communities (Hope et al. 2020). In the face of challenging medical scenarios, such as the rise of a novel virus or situations where standard treatments fail, rapidly finding reports of similar challenges and directions to address them could have dramatic effect (Longhurst, Harrington, and Shah 2014). Finally, at the macro level, this ability could assist policymakers and funding agencies (e.g., NIH, NSF) seeking to identify important challenges and promising directions to prioritize research programs; in times of crisis this process needs to be done rapidly but demands substantial human effort.

To address this problem and facilitate discovery of scientific knowledge, we make the following key contributions:

- **Novel Task: Extraction and Search of Scientific Challenges and Directions.** We define semantic categories for 'challenges' and 'directions' that generalize across many types of difficulties, limitations, flaws and hypotheses or potential indications that an issue is worthy of investigation. We focus on COVID-19 literature as the main test bed for our task, as it is known to be highly in-

terdisciplinary (Hope et al. 2021) with research in many different fields (e.g., AI, climatology, engineering, economics) and relates to a global emergency that urgently demands tools to help researchers and clinicians keep track of challenges and new opportunities.

- **Expert-Annotated Dataset, Publicly Released.** We collect and publicly release a resource of 2.9K expert-annotated texts from full-length COVID-19 papers, labeled by experts for challenges and directions with high inter-annotator agreement. We use the data to train multi-label sentence classification models that achieve high accuracy scores. We analyze model errors, discovering that contextual information can both help and harm results. Based on this finding, we explore a simple technique that integrates multiple ways of encoding context.
- **Novel Scientific Search Engine For Researchers and Clinicians.**[1] We build a novel public search engine that indexes challenges and directions. We apply a model trained on our dataset and apply it to the full corpus of 550K COVID-19 papers to build an index of scientific challenges and potential directions. We create a search engine that allows users to search for combinations of entities (e.g., names of drugs, diseases, etc.) and retrieve challenge/direction sentences that mention them.
- **Evaluating Generality: Zero-Shot Generalization across Biomedicine and AI.** We demonstrate zero-shot generalization, obtaining a high MAP of over 95% when applying the model trained on COVID-19 papers to a broader corpus in the general biomedical domain, and to AI papers in computer science. This indicates the potential value of our resource beyond COVID-19, such as for future pandemics or crises, or for helping AI researchers handle the explosion of research in this area.
- **Evaluating Utility: User Studies with Researchers.** We conduct studies measuring utility. First, we evaluate the system's ability to help researchers with diverse backgrounds discover challenges and directions for a given query (e.g., directions in *drug discovery*). This could also be important for researchers looking into a new area, e.g., AI researchers seeking biomedical problems (Fig. 1). Second, we recruit nine *medical researchers working on COVID-19* in clinical practice and research. These users often require finding information on challenges and directions, during research or treatment planning. In both experiments, totalling 19 researchers and over 70 distinct queries, our prototype outperforms PubMed, the most widely used biomedical search tool, in both quality and utility for discovery of challenges and directions.

2 Task Overview & Definitions

2.1 Challenge and Direction Identification

The CORD-19 corpus (Wang et al. 2020) curates literature on COVID-19 and related diseases. With many thousands of papers, keeping track is generally hard, and mapping the landscape of scientific challenges and directions to address them is even harder. While "grand" challenges such as designing therapies and handling novel virus variants are broadly known, research focuses on *fine-grained* specific challenges, e.g., difficulties in functional analysis of specific viral proteins, or shortcomings of a specific treatment regime for children. Each challenge, in turn, is associated with potential directions and hypotheses.[2]

We present a novel task of automatically identifying sentences in papers that clearly state *scientific challenges and directions*. We consider the multi-label classification setting, where for a given sentence $\mathcal{X} = \{w_1, w_2, \ldots, w_T\}$ with T tokens, our goal is to output two labels $\mathcal{Y} = \{c, d\}$, where c and d are binary targets indicating if the sentence mentions a challenge/direction, respectively. Additionally, we are also given *context* sentences surrounding $\mathcal{X}$: $(\mathcal{X}_{\text{previous}}, \mathcal{X}_{\text{next}})$, for the previous and next sentences, respectively, which could be used as further input to models. The multi-label setting allows us to capture that in many cases, sentences refer to both challenges and directions at the same time (see Table 4). At a high level, our labels are defined as follows.

- **Challenge:** A sentence mentioning a problem, difficulty, flaw, limitation, failure, lack of clarity, or knowledge gap.
- **Research direction:** A sentence mentioning suggestions or needs for further research, hypotheses, speculations, indications or hints that an issue is worthy of exploration.

These categories allow us to capture important information for scientists that is not captured by existing resources (see §5). As part of data annotation we provide annotators with richer explanations and examples of each label (see §3.1) to make these definitions more concrete. Figure 1 shows examples for each category (also see Table 4 in Technical Appendix A.1 for more discussion[3]).

Many cases of challenges and directions are non-trivial for both humans and machines to identify. We demonstrate two main types of difficulties (see more discussion in Technical Appendix A.4) — cases of potentially misleading keywords, and cases where deep domain knowledge or context may be required.

- **Misleading keywords**. Consider the following sentence: *"The 15-30 mg/L albumin concentration is a critical value that could indicate kidney problems when it is repeatedly exceeded"*. This text mentions a diagnostic measure that is an indicator of a problem, rather than an actual problem. This is one example out of many other potentially misleading cases, such as cases where a negative outcome occurs to an entity we wish to harm (e.g., "the viral structural integrity is destroyed").
- **Context and domain knowledge**. *"BV-2 cells expressed Mac1 (CD11b) and Mac2 but were negative for the oligodendrocyte marker GalC..."* Deciding whether this sentence contains a challenge is highly non-trivial, since it requires more context and deep domain knowledge to understand whether this outcome is problematic or not.

[1] challenges.apps.allenai.org/.

[2] While many papers discuss future directions in their concluding section, our task involves capturing all mentions of directions/hypotheses/speculations/early indications appearing throughout *full* paper texts (e.g., in experimental analysis sections).

[3] Appendix can be found at https://arxiv.org/abs/2108.13751.

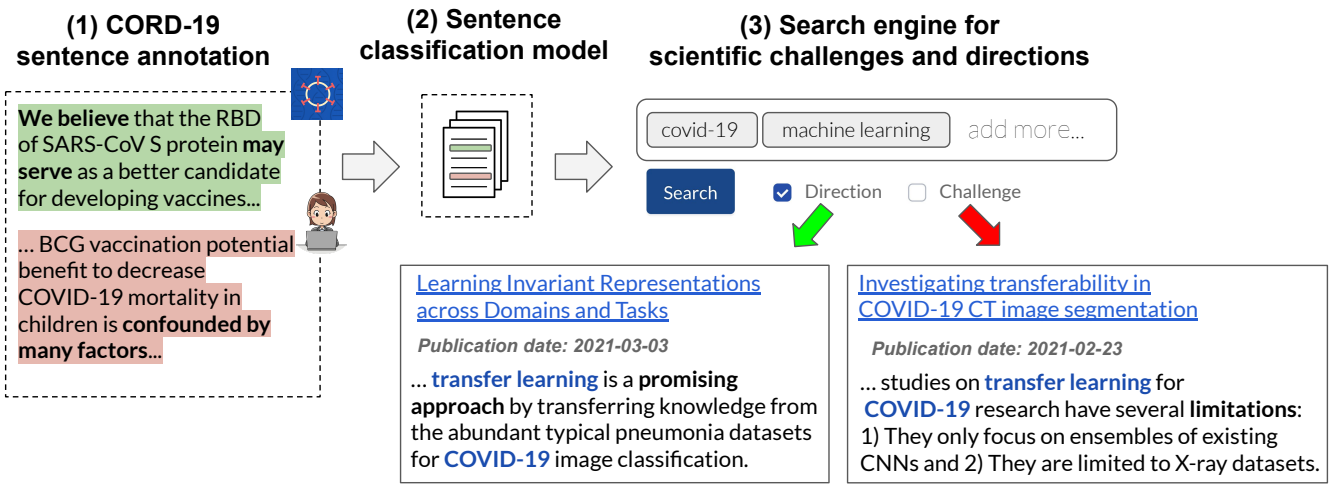

Figure 1: Overview of our system. (1) We collect expert annotations of sentences mentioning challenges and directions from across the CORD-19 corpus. (2) We train a sentence identification model on this data and apply it to the full corpus to extract high-confidence sentences. (3) We build a search engine indexing challenges and directions in COVID-19 literature, allowing users to search for entities and retrieve sentences with their contexts.

3 Data Collection and Models

We now describe our approach for identifying mentions of challenges and directions, starting with collecting expert annotations needed to train and evaluate models on our task.

3.1 Data

Data Collection & Annotation We recruited four expert annotators with biomedical and bioNLP backgrounds to annotate sentences sampled across CORD-19. Annotators were given detailed annotation guidelines [4] and had a one-hour training session for reviewing the guidelines and discussing more examples. The guidelines included simple explanations of challenges and directions along with introductory examples. We sampled sentences from full-text papers, aiming to capture diverse, fine-grained challenges/directions that often do not appear in abstracts. The subset of full-text papers in CORD-19 numbers roughly 180K papers with around 25 millions sentences.[5] We also provide surrounding sentences around the target sentence as context.

Randomly sampling sentences for annotation is highly unlikely to lead to enough challenge/direction cases. To increase this likelihood, two annotators curate 280 keywords or phrases with affinity to one of the two categories.[6] Sentences mentioning at least one keyword (lemmatized) are upsampled. For example, words such as *unknown, limit, however* provide weak signal indicating a potential mention of a challenge; words like *suggest, future work, explore* are weak indicators of a direction. To expand the list further, annotators made use of SPIKE (Taub-Tabib et al. 2020) which also has a vocabulary explorer that allows browsing keywords similar to an input term. Overall, the 280 keywords covered around a third of sentences in CORD-19, demonstrating their breadth. We note that for most keywords context can completely change their meaning; for instance, "limit" can appear in the context of "we limit the discussion" which has no relation to challenges. Our set of terms with weak correlation to the label (e.g., the word *may* that very weakly relates to directions) favors high recall rather than precision.

Finally, to further increase coverage, we sampled at random roughly a quarter of sentences from the remaining sentences that did *not* contain any of the keywords, obtaining in total 3000 sentences. We filter sentences that are not in English, mostly numeric/mathematical, or that are very short-/long (often due to PDF parsing issues), resulting in 2894 sentences and their surrounding contexts, from 1786 papers.

Annotator agreement: 60% of the sentences were labeled by all annotators[7], with high average pairwise agreement. Following common practice we measure micro-F1 and macro-F1, treating labels from one annotator as ground-truth and the other as predicted, obtaining 85% for challenges and 88% for directions for micro-F1, and 84% and 82% for macro-F1. Positive label proportions are 39.66% and 22.74% for challenges/directions, respectively. We create a train/dev/test stratified split of 40%/10%/50%(Table §1), splitting by distinct *papers*. We opt for a large, diverse test set for model evaluation (Card et al. 2020). The sampled sentences originate from papers published in 1108 journals.

A note on crowdsourcing. We also attempted crowdsourcing to scale the collection process.[8] However, despite multiple trials and strict quality assurance, the nuanced nature of the task was found to be difficult for crowd workers, especially due to false negatives.

[4] Annotation guidelines are available in our code repository.
[5] We use a snapshot of CORD-19 from 08-02-2021.
[6] Our list of keywords is available in our code repository.
[7] Final labels selected by majority vote, with ties (fewer than 100 cases) adjudicated by a member of the research team.
[8] Using the Appen platform https://appen.com/.

Labels	Train	Dev	Test	All
Not Challenge, Not Direction	602	146	745	1493
Not Challenge, Direction	106	25	122	253
Challenge, Not Direction	288	73	382	743
Challenge, Direction	155	40	210	405

Table 1: Distribution of labels across data splits. Splits are stratified with no overlap in papers.

3.2 Baseline Models

The classification task at hand is a multi-label sentence classification problem, with the goal of predicting whether a sentence mentions a challenge, a research direction, both, or neither. The definitions of the challenge and direction categories are as described in §2. We evaluate a range of baseline models we examine for our novel task.

- **Keyword-based**: A simple heuristic based on the lexicon we curated for data collection (§3.1) — sentences with a challenge keyword are labeled as challenge, and similarly for direction.
- **Sentiment**: Challenge statements potentially have a negative tone, and directions are potentially more positive. We score the sentiment of each sentence using an existing tool (Loria 2018) and classify negative sentiment sentences as challenges and positive ones as directions.
- **Zero-shot inference**: In zero-shot classification, models predict labels they were not trained on (Yin, Hay, and Roth 2019). This could be particularly relevant in emerging domains such as COVID-19, where collecting large amounts of labeled data could be prohibitive. We use a language model trained for natural language inference (NLI), letting the model infer whether the input text *entails* the label name. See Appendix A.2 for full details.
- **Scientific language models**: We also experiment with fine-tuning language models that were pre-trained on scientific papers. We report results for PubMedBERT-abstract-fulltext (Gu et al. 2020) which was pre-trained on PubMed paper abstracts & full texts, and for SciBERT (Beltagy, Lo, and Cohan 2019), trained on a corpus of biomedical and computer science papers. In addition, we also experiment with a non-scientific language model, RoBERTa-large, which has been shown to obtain excellent results when fine-tuned on scientific texts (Gururangan et al. 2020). We also experimented with other language models, with very similar results. We fine-tune all language models and perform basic hyperparameter search on the development set. See Technical Appendix A.2 for full reproducibility details.

3.3 Context Modelling Variants

We also experiment with models motivated by examination of baseline errors (see Technical Appendix A.4). Specifically, we find that adding context helps in certain cases: For example, in the sentence *"... the patient had an extreme elevation of procalcitonin without signs of bacterial infection."* which was misclassified as a non-challenge, adding context helped identify the unexplained elevation as problematic. However, context can also introduce noise (see Table 2). We explore different ways in which the context can affect predictions — during training, and during inference. In addition to simply fine-tuning PubMedBERT with full context, we explore two main customized approaches.

Hierarchical Attention Network (HAN) (Yang et al. 2016) Recall Section §2, where candidate sentences are denoted by $\mathcal{X}$ and their context by $\mathcal{X}_{\text{previous}}, \mathcal{X}_{\text{next}}$. Denote by $\mathcal{X}_{\text{context}}$ the concatenation: `[CLS]` $\mathcal{X}_{\text{previous}}$ `[SEP]` $\mathcal{X}$ `[SEP]` $\mathcal{X}_{\text{next}}$ `[SEP]`. We compute a weighted average of `[CLS]` and the first two `[SEP]` tokens using attention weights, and use this average embedding for final classification. The weights are learned as part of end-to-end training.[9] While this model can potentially learn to re-weight the context, it encodes the full $\mathcal{X}_{\text{context}}$ jointly before this weighting takes place, which can lead to noise propagating early on. We thus also test a different approach we design.

Context Slice + Combine Let $f_{\mathcal{X}}(\mathbf{x})$ denote the logits from the classification layer of the PubMedBERT model fine-tuned on $\mathcal{X}$ only, for input text $\mathbf{x}$. Denote by $f_{\mathcal{X}_{\text{context}}}(\mathbf{x})$ the logits from PubMedBERT fine-tuned using the *full context*. At inference time, we obtain outputs using the following "slices" of f and $\mathbf{x}$: (1) $l_1 = f_{\mathcal{X}}(\mathcal{X})$, (2) $l_2 = f_{\mathcal{X}_{\text{context}}}(\mathcal{X}_{\text{context}})$, (3) $l_3 = f_{\mathcal{X}}(\mathcal{X}_{\text{context}})$, and (4) $l_4 = f_{\mathcal{X}_{\text{context}}}(\mathcal{X})$. We then average ("combine") all four, yielding a final pair of logits used for prediction. (1) and (2) are just the models reported in Table 2 – feeding $\mathcal{X}$ as input to PubMedBERT fine-tuned on $\mathcal{X}$, and similarly for $\mathcal{X}_{\text{context}}$. (3) and (4) switch between training and inference inputs: in (3) $f_{\mathcal{X}}$ takes $\mathcal{X}_{\text{context}}$ as input during inference, and in (4) $\mathcal{X}$ is fed as input into $f_{\mathcal{X}_{\text{context}}}$. The reason we include these is to tease apart different ways in which the context may introduce noise or signal, during training using context (3) and during inference (4). We empirically find all four are in agreement in roughly 70% / 83% of the cases for challenges / directions; 3 out of 4 agree in 20% / 11%, and the rest are tied. This suggests each variant may capture complementary information.

3.4 Results

Classification Results. As seen in Table 2, fine-tuned scientific language models outperform the Zero-Shot model, which still does well considering it had no supervision and was pre-trained on non-scientific texts. The sentiment analysis and keyword-based classifiers, both based on large lists of "positive/negative" keywords, have good recall but poor precision. The best individual classifier by F1 is PubMedBERT with a binary-F1 of 0.770 and 0.766 on the challenge and direction labels, respectively. The HAN approach was able to increase recall substantially for problems, but at the cost of reduction in precision, leading to overall inferior F1 on par with PubMedBERT+context.

The Slice-Combine approach leads to an improvement of about one F1 point for both labels over the best individual model (standard error of 1.05×10^{-4}). In an ablation experiment we compute the averaged logits of l_1, l_2 and l_3, l_4 *sep-*

[9] See Yang et al. (2016) for details about the general framework.

	Challenge			Direction		
Model	P	R	F1	P	R	F1
Keyword	0.535	0.760	0.628	0.455	0.792	0.578
Sentiment	0.405	0.966	0.571	0.239	0.837	0.371
NLI-Zeroshot	0.659	0.693	0.675	0.401	0.825	0.540
RoBERTa-large	0.723 (0.042)	0.824 (0.046)	0.769 (0.004)	0.697 (0.065)	0.825 (0.06)	0.754 (0.004)
SciBERT	0.729 (0.023)	0.799 (0.03)	0.761 (0.007)	0.719 (0.044)	0.783 (0.043)	0.749 (0.01)
PubMedBERT	0.738 (0.018)	0.804 (0.017)	**0.770** (0.006)	0.755 (0.017)	0.778 (0.015)	**0.766** (0.006)
+context	0.716 (0.048)	0.809 (0.047)	0.758 (0.007)	0.701 (0.038)	0.771 (0.026)	0.733 (0.01)
PubMedBERT-HAN	0.671 (0.02)	0.863 (0.03)	0.759 (0.01)	0.674 (0.04)	0.804 (0.04)	0.734 (0.001)
Slice-Combine	0.742 (0.011)	0.829 (0.012)	**0.783** (0.004)	0.732 (0.02)	0.82 (0.03)	**0.773** (0.005)

Table 2: Model Results. The PubMedBERT model fine-tuned on our multi-label classification task performs best. For the neural models we present the average over 5 training seeds where the number in parentheses is the standard deviation.

arately, and also simply ensemble four model runs of fine-tuned PubMedBERT, both leading to inferior results (see Technical Appendix A.3). Finally, an oracle that selects the best logit l_1-l_4 for each input based on ground truth labels has F1 of 0.907 and 0.896 for challenges/directions, suggesting much room for future work on adaptive use of context during training and inference. See in-depth analysis of additional model errors in Technical Appendix A.4.

Precision@Recall Our primary focus is a novel search engine application (§4). For such applications, it is often more important to have high precision for top retrieved results. We examine precision for a range of values of recall, shown in Figure 2. We observe that for 20% recall we obtain well over 90% precision, and for 40% recall about 90% precision.

Evaluating predictions across CORD-19 To further ensure quality, we run the PubMedBERT model across all sentences in CORD-19. Out of all sentences indexed in our search engine as either a challenge or a direction, we sample roughly 350 sentences (see Appendix A.3 for details). These sentences are labeled by an expert annotator following the same criteria used to annotate our dataset (§3.1). As shown in Figure 3, we obtain very high mean average precision (MAP) of 98% and area under the precision-recall curve (AUC) of over 97% for directions, and 97% / 96% for challenges. We conclude that for high-confidence challenge and direction sentences indexed in our search engine, accuracy is expected to be overall considerably high. Our test set consists of considerably harder examples, explaining the gap in performance (see discussion in A.4).

Zero-shot generalization to biomedicine and AI domains. We explore whether a model trained on our dataset can, with no additional training, generalize to identify challenges and directions in *general* biomedical papers, which we sample from S2ORC (Lo et al. 2020), and also AI papers (Jain et al. 2020). In total, we sample about 1000 sentences, following the same procedure as described above for CORD-19 sentences (see Appendix A.3 for more details). CORD-19 papers are highly interdisciplinary (Hope et al. 2021, 2020), raising the possibility of using our dataset to train models that can be applied to new domains without additional

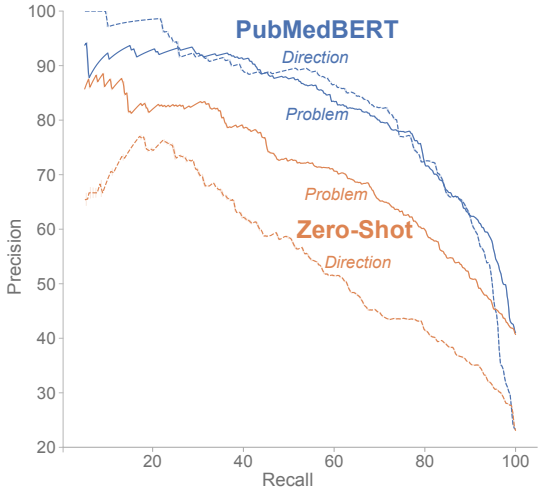

Figure 2: Precision/Recall results for the PubMedBERT model, and the zero-shot model. Precision for PubMedBERT is high for reasonably large values of recall.

data collection. As seen in Figure 3, for directions we obtain MAP and AUC of around 96% for biomedicine, and around 95% for AI. For challenges, MAP and AUC reach around 97-98% for biomedicine and around 96% for AI. These preliminary results could be explored further in future work.

Training data size. Finally, we also tried training our model on only 10% of the training set. We obtained average F1 of 0.72/0.71 for challenges/directions. This suggests that a low-resource effort can obtain decent results — potentially important in emerging scenarios where time is limited.

4 Search Engine User Studies

We now explore user studies designed to evaluate our framework's utility. First, we explore whether our system can be helpful for quick discovery of challenges/directions. Second, we conduct a study with nine medical researchers working on COVID-19 treatment and research. In total our studies include 19 researchers and over 70 distinct search queries.

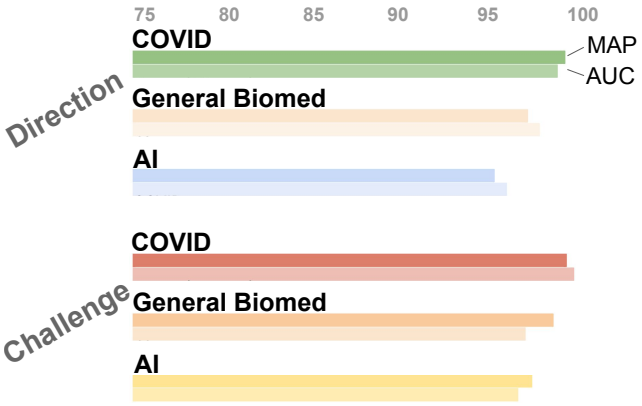

Figure 3: Evaluating predictions beyond our test set. We use a model trained on our data to identify challenges and directions across CORD-19 (denoted by *COVID*), S2ORC (general biomedical papers, denoted *general biomed*) and SciRex (full-text AI papers, denoted *AI*). Accuracy is considerably high. Zero-shot generalization over non-COVID papers, even non-biomedical papers, is encouragingly high, indicating the utility of our resource beyond COVID-19.

4.1 Search Engine

Challenge and direction indexing We build a search engine that indexes challenges and directions across the entire CORD-19 corpus up to and including August 2021. To build the search engine (see Figure 5 in the Appendix for a screen capture), we first apply PubMedBERT to 550K papers, totalling 29M sentences. 180K of the papers are with full text, the rest are abstracts. We then clean poorly tokenized sentences, non-English sentences, very short sentences or texts with latex code. We classify the remaining sentences leaving 2.2M sentences — about 950K sentences with high-confidence predictions for at least one of challenge/direction and their surrounding context sentences. We select high-confidence sentences by using a threshold of 0.99 for both challenges and directions, using a thresholds leading to well over 90% precision at top-10% on our test set.

Entity-based indexing For each sentence in our set of 2.2M, we add another layer of indexing, by extracting entities and linking them to knowledge base entries. This allows us to partially group together all challenges or directions into "topics" referring to a specific fine-grained combination of concepts (e.g., *AI* + *diagnosis* + *pneumonia*), and facilitate entity-centric faceted search which is known to be useful in scientific exploratory search (Hope et al. 2020, 2021).We extract a range of biomedical entities and link them to a biomedical KB of MeSH (Medical Subject Headings) entities (Lipscomb 2000). See Appendix A.5 for full details.

In the experiments that follow, we compare our system with a strong real-world system — PubMed biomedical search engine[10], a leading search site that clinicians and researchers regularly peruse as their go-to tool. While PubMed was not designed to find challenges and directions, no existing tool is; PubMed allows users to search for entities such as MeSH terms, is supported by a KB of biomedical entities used for automatic query expansion, and has many other functions — and as such is a strong real-world baseline.

[10]https://pubmed.ncbi.nlm.nih.gov/

Metric	Chal./Dir. Search	PubMed
Search	**90%**	48%
Utility	**94%**	57%
Interface	**91%**	68%
Overall	**92%**	59%

Table 3: Nine medical researchers expressed much higher satisfaction with our system (Chal./Dir.) than PubMed.

4.2 Challenge/Direction Exploration

We recruited ten participants with education and experience in medicine, microbiology, public health, molecular, cellular, and developmental biology, biochemistry, chemical & biological engineering, environmental science, and mathematics. Participants are paid $50 per hour of work, comparing query results from our system and PubMed. Participants were given guidelines, which include definitions for research challenges and directions with simple examples.[11]

Each participant was given twenty queries, split into two sections for challenges and directions, respectively. For each query, participants were asked to find as many research challenges as possible in no more than 3 minutes. The total number of unique queries among the participants is 65. Some examples of queries used for the challenges section include "antibodies" and "inflammation, lung", with the paired entities being searched jointly; example queries for the directions section include "telemedicine" and "vaccines, technology". All queries were curated by a domain expert.

As seen in Figure 4, our system yielded a greater number of challenges and directions, on average, than the PubMed tool. Users found roughly 4.46 challenges and 6.43 directions per query using our system compared to the 2.24 challenges and 2.03 directions per query found using PubMed (p-value of .00192 for challenges and .000529 for directions using a paired t-test). For each participant we included 5 challenges and 5 directions that were overlapping across all participants, in order to control and compare between results for the same queries. We find that on average across users, 70.0% of the query results using our system led to a strictly larger number of challenges discovered than the respective query results using PubMed, and 22% were ties. For directions, we find a larger gap between the two systems, with 96.0% of the query results using our system yielding strictly more directions than PubMed, and 2% yielding ties.

4.3 Evaluation with Medical Researchers

We now report on an evaluation of our search engine performed with nine medical researchers at a large hospital.[12]

[11]Full annotation guidelines are included in our code repository.

[12]See Appendix A.7 for a more detailed example scenario where medical researchers need to search for challenges and directions.

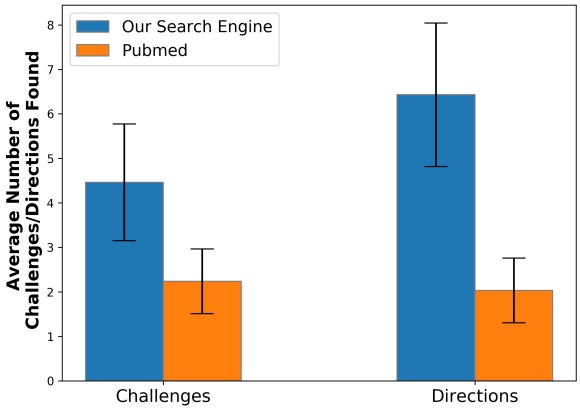

Figure 4: Study participants using our search engine were able to find substantially more challenges and directions they considered useful than with PubMed. Error bars represent 90% confidence intervals.

Study. We recruited nine expert MDs with a wide range of specialization including cardiology, pulmonary and critical care medicine, gastroenterology and general medicine who are actively involved in clinical research both for COVID-19 and specialty areas, and each have over 1000 citations. Each expert completed randomly ordered search tasks, in the form of challenge/direction queries curated by an expert medical researcher, using both PubMed and our system. For example, experts performed the following search tasks:

- **Find problems/limitations** related to COVID-19 and each of (1) *hospital infections*, (2) *diagnosis*, (3) *vaccines for children*, (4) *probiotics and the gastrointestinal tract*.
- **Find directions/hypotheses** related to COVID-19 and each of (1) *mechanical ventilators*, (2) *liver*, (3) *artificial intelligence*, (4) *drug repositioning*.

Experts using our UI viewed sentences and their contexts (previous/next sentences). In addition we also displayed metadata such as paper title, date, url. After all search tasks were completed for both systems, experts were given seven-point Likert-scale questions to judge system utility, interface, and search quality. Following (Hope et al. 2021), we use a standardized Post Study System Usability Questionnaire (PSSUQ) (Lewis 2002), widely used in system quality research, and added questions designed to evaluate search and exploration utility: *overall search accuracy*, *results that are not only relevant but interesting or new*, *finding papers interesting to read*, and *ability to understand and judge each individual result quickly without additional context*. Each question is asked twice, once for PubMed and once for our system, leading to $15 \times 2 \times 6 = 180$ responses.

Results. Table 3 shows the average Likert scores (normalized to [0%,100%]). We group questions by three types for brevity. The medical experts strongly prefer our search engine to PubMed (overall average of 92% vs. 59%, non-normalized scores of 6.42 vs. 4.14). Across all questions, the majority of the nine MDs assigned our system a higher score than PubMed, at an average rate of 85% per question. When considering ties, the rate is 92%. Our system significantly outperformed PubMed across all questions (Wilcoxon signed rank test p-value of 5.409×10^{-6}).

5 Related Work

Scientific information extraction and text classification. The goal in this line of work is to extract structured information from literature, such as sentence-level classification into categories including objectives/methods/findings (Dernoncourt and Lee 2017) or extracting entities and relations (Li et al. 2016; Wallace et al. 2016; Kim, Wang, and Yasunori 2013). Unlike previous work, our labelling schema encapsulates underexplored facets and covers diverse variants of challenges/directions and can help generalize across the interdisciplinary COVID-19 literature (Hope et al. 2021). Some previous schemas focus on subsets of categories that are subsumed by ours, which is broader in scope and captures aspects missed by other work (e.g., societal hardships)(Teufel, Siddharthan, and Batchelor 2009; Liakata, Q, and Soldatova 2009; Fisas, Saggion, and Ronzano 2015; Fisas, Ronzano, and Saggion 2016). Most previous work was also multi-*class* rather than multi-*label*, which excludes important cases of statements that are both challenges and directions. Importantly, most other datasets are much narrower and limited in size, with the largest relevant corpus consisting of about 10X fewer papers than our own.

COVID-19 IE and search tools. Recent work includes visualizing COVID-19 concepts and relations (Hope et al. 2020), a syntactic search engine (Shlain et al. 2020), and a search engine for causal and functional relations (Hope et al. 2021). Ours system is focused on challenges and directions, not captured by existing tools. Recent work (Huang et al. 2020) has used crowd workers to annotate *abstracts* (not full-texts as in this paper) for Background, Purpose, Method, Finding/Contribution. As discussed in §3.1, we find that crowd workers fail on our task, even though recruited with high quality assurance standards.

6 Conclusion

We presented methods for extracting scientific challenges and directions from scholarly papers. We collected 3K expert-labeled sentences and their contexts from COVID-19 papers, and used the dataset to fine-tune scientific language models on our multi-label sentence classification task. Our model identifies challenges and directions with high precision, and achieves high zero-shot generalization on general biomedical and AI papers. We used the model to index 950K sentences and build a novel search engine that allows researchers to search for biomedical entities and retrieve sentences mentioning difficulties, limitations, hypotheses and directions. Researchers using our system, including those working on COVID-19, found that our system provided better support than PubMed in terms of utility and relevance. In future work, we hope to build more tools to explore and visualize challenges and directions across science.

Acknowledgments

Lahav is partially supported by the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. Weld's work at the University of Washington is funded by ONR grant N00014-18-1-2193, NSF RAPID grant 2040196, the WRF/Cable Professorship, and AI2.

References

Bada, M.; Eckert, M.; Evans, D.; Garcia, K.; Shipley, K.; Sitnikov, D.; Baumgartner, W. A.; Cohen, K. B.; Verspoor, K.; Blake, J. A.; et al. 2012. Concept annotation in the CRAFT corpus. *BMC bioinformatics*.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3606–3611.

Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*.

Card, D.; Henderson, P.; Khandelwal, U.; Jia, R.; Mahowald, K.; and Jurafsky, D. 2020. With Little Power Comes Great Responsibility. In *EMNLP*.

Chalmers, I.; Bracken, M. B.; Djulbegovic, B.; Garattini, S.; Grant, J.; Gülmezoglu, A. M.; Howells, D. W.; Ioannidis, J. P. A.; and Oliver, S. 2014. How to increase value and reduce waste when research priorities are set. *The Lancet*, 383(9912): 156–165.

Clarke, M.; and Hopewell, S. 2013. Many reports of randomised trials still don't begin or end with a systematic review of the relevant evidence. *Journal of the Bahrain medical society*, 24: 145–148.

Cooper, N.; Jones, D. R.; and Sutton, A. 2005. The use of systematic reviews when designing studies. *Clinical Trials*, 2: 260 – 264.

Dernoncourt, F.; and Lee, J. Y. 2017. PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 308–313.

Fisas, B.; Ronzano, F.; and Saggion, H. 2016. A Multi-Layered Annotated Corpus of Scientific Papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3081–3088. Portorož, Slovenia: European Language Resources Association (ELRA).

Fisas, B.; Saggion, H.; and Ronzano, F. 2015. On the Discoursive Structure of Computer Graphics Research Papers. In *Proceedings of The 9th Linguistic Annotation Workshop*, 42–51. Denver, Colorado, USA: Association for Computational Linguistics.

Gilbert, R.; Salanti, G.; Harden, M.; and See, S. 2005. Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. *International journal of epidemiology*, 34 4: 874–87.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *ACL*.

Hope, T.; Amini, A.; Wadden, D.; van Zuylen, M.; Parasa, S.; Horvitz, E.; Weld, D.; Schwartz, R.; and Hajishirzi, H. 2021. Extracting a Knowledge Base of Mechanisms from COVID-19 Papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4489–4503.

Hope, T.; Chan, J.; Kittur, A.; and Shahaf, D. 2017. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 235–243.

Hope, T.; Portenoy, J.; Vasan, K.; Borchardt, J.; Horvitz, E.; Weld, D. S.; Hearst, M. A.; and West, J. 2020. SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 135–143.

Huang, T.-H.; Huang, C.-Y.; Ding, C.-K. C.; Hsu, Y.-C.; and Giles, C. L. 2020. CODA-19: Using a Non-Expert Crowd to Annotate Research Aspects on 10,000+ Abstracts in the COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Jain, S.; van Zuylen, M.; Hajishirzi, H.; and Beltagy, I. 2020. SciREX: A challenge dataset for document-level information extraction. *arXiv preprint arXiv:2005.00512*.

Ker, K.; Edwards, P.; Perel, P.; Shakur, H.; and Roberts, I. 2012. Effect of tranexamic acid on surgical bleeding: systematic review and cumulative meta-analysis. *BMJ*, 344.

Kim, J.-D.; Ohta, T.; Tsuruoka, Y.; Tateisi, Y.; and Collier, N. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*.

Kim, J.-D.; Wang, Y.; and Yasunori, Y. 2013. The genia event extraction shared task, 2013 edition-overview. In *BioNLP Shared Task Workshop*.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Lewis, J. R. 2002. Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, J.; Sun, Y.; Johnson, R. J.; Sciaky, D.; Wei, C.-H.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Wiegers, T. C.; and

Lu, Z. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*.

Liakata, M.; Q, C.; and Soldatova, L. N. 2009. Semantic Annotation of Papers: Interface & Enrichment Tool (SAPIENT). In *Proceedings of the BioNLP 2009 Workshop*, 193–200. Boulder, Colorado: Association for Computational Linguistics.

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

Lo, K.; Wang, L. L.; Neumann, M.; Kinney, R.; and Weld, D. S. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of ACL*.

Longhurst, C. A.; Harrington, R. A.; and Shah, N. H. 2014. A 'green button'for using aggregate patient data at the point of care. *Health affairs*, 33(7): 1229–1235.

Loria, S. 2018. textblob Documentation. *Release 0.15*, 2.

Mohan, S.; and Li, D. 2018. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. In *Automated Knowledge Base Construction (AKBC)*.

Neumann, M.; King, D.; Beltagy, I.; and Ammar, W. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Portenoy, J.; Radensky, M.; West, J.; Horvitz, E.; Weld, D.; and Hope, T. 2021. Bridger: Toward Bursting Scientific Filter Bubbles and Boosting Innovation via Novel Author Discovery. *arXiv preprint arXiv:2108.05669*.

Robinson, K.; and Goodman, S. 2011. A Systematic Examination of the Citation of Prior Research in Reports of Randomized, Controlled Trials. *Annals of Internal Medicine*, 154: 50 – 55.

Shlain, M.; Taub-Tabib, H.; Sadde, S.; and Goldberg, Y. 2020. Syntactic Search by Example. In *ACL*.

Sinclair, J. 1995. Meta-analysis of randomized controlled trials of antenatal corticosteroid for the prevention of respiratory distress syndrome: discussion. *American journal of obstetrics and gynecology*, 173 1: 335–44.

Taub-Tabib, H.; Shlain, M.; Sadde, S.; Lahav, D.; Eyal, M.; Cohen, Y.; and Goldberg, Y. 2020. Interactive Extractive Search over Biomedical Corpora. arXiv:2006.04148.

Teufel, S.; Siddharthan, A.; and Batchelor, C. 2009. Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1493–1502. Singapore: Association for Computational Linguistics.

Wallace, B. C.; Kuiper, J.; Sharma, A.; Zhu, M.; and Marshall, I. J. 2016. Extracting PICO Sentences from Clinical Trial Reports Using Supervised Distant Supervision. *J. Mach. Learn. Res.*, 17(1): 4572–4596.

Wang, L. L.; Lo, K.; Chandrasekhar, Y.; Reas, R.; Yang, J.; Eide, D.; Funk, K.; Kinney, R.; Liu, Z.; Merrill, W.; et al. 2020. CORD-19: The Covid-19 Open Research Dataset. *arXiv preprint arXiv:2004.10706*.

Wolf, T.; Chaumond, J.; Debut, L.; Sanh, V.; Delangue, C.; Moi, A.; Cistac, P.; Funtowicz, M.; Davison, J.; Shleifer, S.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489.

Yin, W.; Hay, J.; and Roth, D. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3905–3914.