COVID-SEE: The Scientific Evidence Explorer for COVID-19 Related Research

Karin Verspoor
RMIT University
Melbourne, Australia
karin.verspoor@rmit.edu.au

Yulia Otmakhova
University of Melbourne
Melbourne, Australia
yotmakhova@student.unimelb.edu.au

Simon Šuster
University of Melbourne
Melbourne, Australia
simon.suster@unimelb.edu.au

Jey Han Lau
University of Melbourne
Melbourne, Australia
jeyhan.lau@unimelb.edu.au

Timothy Baldwin
University of Melbourne
Melbourne, Australia
tbaldwin@unimelb.edu.au

Antonio Jimeno Yepes
RMIT University
Melbourne, Australia
antonio.jose.jimeno.yepes@rmit.edu.au

David Martinez Iraola
Doctor Evidence, LLC
Melbourne, Australia
jibmaird@gmail.com

We present COVID-SEE[1], a system for medical literature discovery, which augments search through a structured visual overview of a collection enabling evidence exploration. While many search and question answering tools emerged in response to the COVID-19 outbreak, relatively few leverage domain knowledge to organise and present information found within the literature [5]. To fill this gap, we developed a web application that combines a search engine for COVID-19 medical literature with summary visualisations of document content, such as concepts, relations, and topics.

A typical usage scenario in COVID-SEE begins with a textual query over the COVID-19 literature, supporting semantic search and providing: (i) a list of *retrieved documents* together with their metadata, and (ii) a *visualisation dashboard* with three distinct interactive views which highlight the relations between entities and concepts detected in the documents. As a user reviews and interacts with the information in these views, documents of interest can be saved into a *collection* for later export or targeted visualisation. Our objective is to combine learning and investigation with direct retrieval to support the known health information seeking behaviour of alternating between focused and exploratory search [3]. We facilitate exploration by providing views of document content that provide a user with deeper insight into retrieved articles.

The first view is a **relational concept view** – a Sankey diagram frame, in which we organise the medical concepts found in the retrieved articles according to key entities for clinical queries, known as PICO [4] (Population, Intervention, Comparator, Outcome). In this view, more salient relations – based on the number of supporting abstracts – carry more weight, and once a relation is clicked, the corresponding articles are revealed (Fig. 1). To detect PICO entities, we train a BiLSTM-CRF model on the EBM-NLP dataset [1] containing reports of randomised clinical trials annotated with textual spans that describe the PICO elements. We then recognise medical terms from Medical Subject Headings (MeSH), using the MetaMap tool[2]. In the Sankey diagram, we display pairwise relations based on article co-occurrence of Population–Intervention and Intervention–Outcome concepts.

The second, **topic view** (Fig. 2), is thematic and shows representative topics for the current collection. We train a Latent Dirichlet allocation (LDA) topic model over the whole dataset, and display the topics in the retrieved subset of articles visually[3]. To represent the documents for topic modelling, we use Unified Medical Language System (UMLS) concepts extracted using MetaMap rather than tokens, as it helps to preserve multi-word concepts such as *intensive care unit* and map different variants of a given term into a single concept, thus reducing noise, and highlighting important keywords [2].

Our third component is a **concept cloud view** (Fig. 2, Inset), showing the 20 most representative concepts for each active document. Concepts here again correspond to UMLS terms extracted using MetaMap. To select discriminative concepts, concept distributions in the selected article are compared to those in the collection using the log-likehood test.

In conclusion, COVID-SEE facilitates more interactive exploration of the COVID-19 literature, through integration of sub-collection thematic analysis, document-level salient concept summaries, and PICO-structured concept relations.

References

[1] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access, 2018.

[2] Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, and Simon Suster. Improved topic representations of medical documents to assist covid-19 literature exploration. 2020.

[3] Patrick Cheong-Iao Pang, Karin Verspoor, Shanton Chang, and Jon Pearce. Conceptualising health information seeking behaviours and exploratory search: result of a qualitative study. *Health and Technology*, 5(1):45–55, Jun 2015.

[1] http://covid-see.com

[2] http://metamap.nlm.nih.gov

[3] https://pyldavis.readthedocs.io/en/latest/readme.html

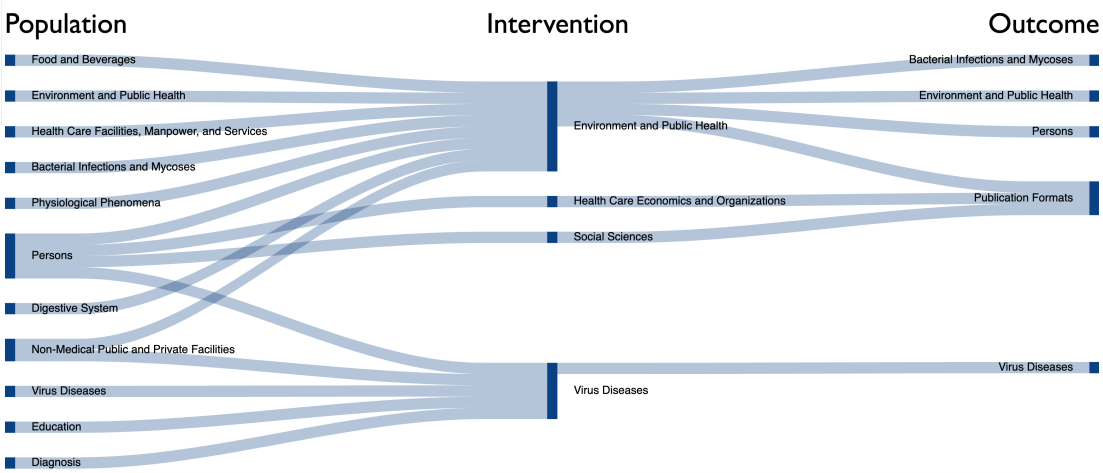

Fig. 1. Visualisation of PICO concepts and relations in articles retrieved for query *incubation period of COVID-19*. Links between concepts can be selected to reveal papers with those relations.

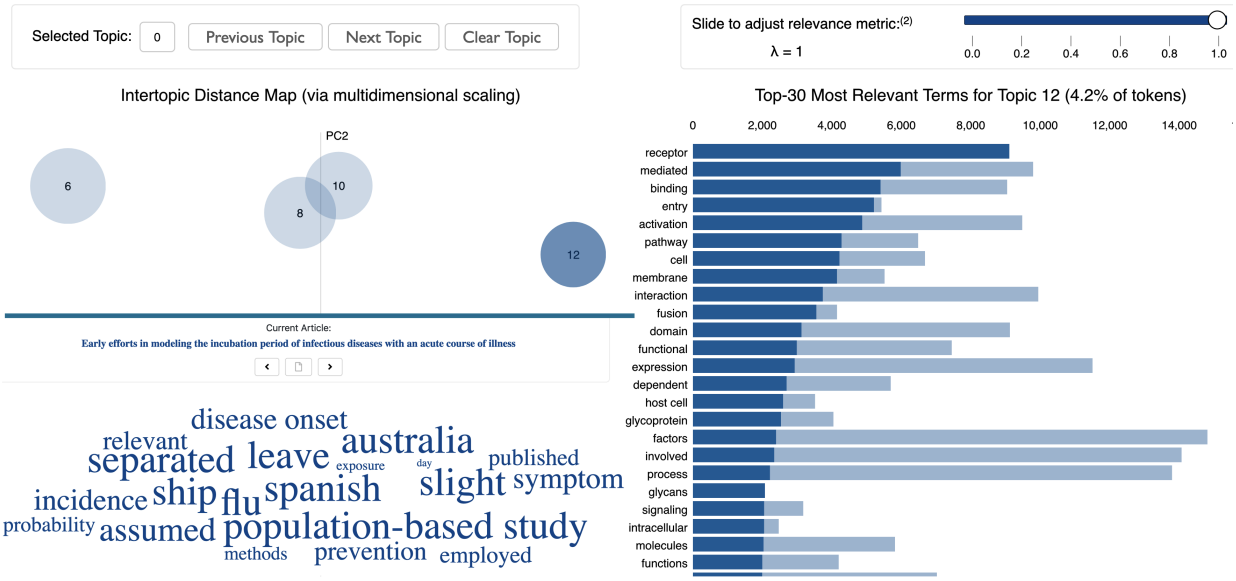

Fig. 2. Topic visualisation for articles retrieved for query incubation period of COVID-19. *Inset:* Word cloud view of an individual document showing 20 key concepts, including multi-word terms.

[4] W Scott Richardson, Mark C Wilson, Jim Nishikawa, Robert S Hayward, et al. The well-built clinical question: a key to evidence-based decisions. *Acp j club*, 123(3):A12–3, 1995.

[5] Lucy Lu Wang and Kyle Lo. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Briefings in Bioinformatics*, 12 2020. bbaa296.