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

Karin Verspoor[1,2(✉)], Simon Šuster[2], Yulia Otmakhova[2,3], Shevon Mendis[2], Zenan Zhai[2], Biaoyan Fang[2], Jey Han Lau[2], Timothy Baldwin[2], Antonio Jimeno Yepes[2,3], and David Martinez[2,3]

[1] RMIT University, Melbourne, Australia
karin.verspoor@rmit.edu.au
[2] The University of Melbourne, Melbourne, Australia
[3] IBM Research Australia, Carlton, Australia

Abstract. We present COVID-SEE, a system for medical literature discovery based on the concept of information exploration, which builds on several distinct text analysis and natural language processing methods to structure and organise information in publications, and augments search through a visual overview of a collection enabling exploration to identify key articles of interest. We developed this system over COVID-19 literature to help medical professionals and researchers explore the literature evidence, and improve findability of relevant information. COVID-SEE is available at http://covid-see.com.

1 Introduction

The outbreak of COVID-19 led to a rapid and proactive response from research communities worldwide. In information retrieval and natural language processing, efforts have concentrated on building tools for efficiently managing the growing literature on COVID-19 [21]. While many tools emerged for article retrieval and question answering, relatively few systems go beyond returning a list of (relevant) documents, or leverage domain knowledge to organise and present information found within the literature [35]. Building on observations about the importance of *exploratory search* [25], with **COVID-SEE** (Scientific Evidence Explorer), we aim 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. Our work is the first comprehensive system incorporating semantic search with visualisation of concepts, relations, and topics [34], extending the capabilities of systems such as SciSight [12,20] and SemViz [13,33] which provide more narrowly-scoped views of the literature (see summary in Table 2).

A typical usage scenario in COVID-SEE begins with a textual query over the COVID-19 literature, providing: (i) a list of *retrieved documents*, and (ii) a *visualisation dashboard*. As a user reviews and interacts with the information in these views, documents of interest can be selected and saved into a *collection* for later export or targeted visualisation. Our objective is to combine learning and investigation with direct

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12657, pp. 559–564, 2021.
https://doi.org/10.1007/978-3-030-72240-1_65

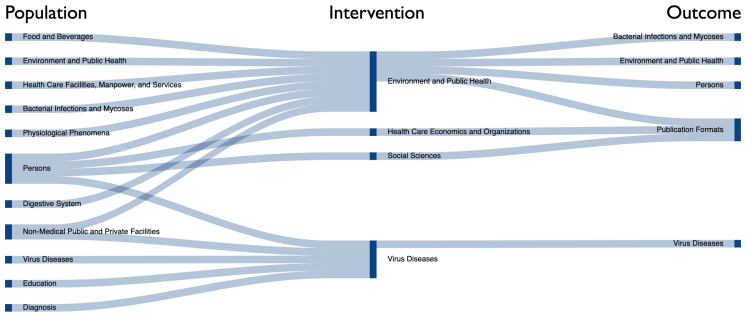

Fig. 1. Visualisation of PICO concepts and relations in articles retrieved for query *incubation period of COVID-19*. Links between concepts can be selected to reveal papers with those relations.

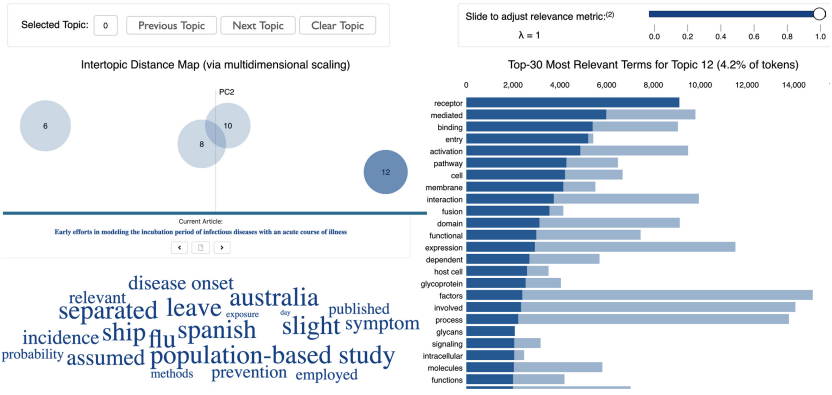

Fig. 2. Topic visualisation for articles retrieved for query *incubation period of COVID-19*. *Inset:* Word cloud view of an individual document showing 20 key concepts, including multi-word terms.

retrieval to support the known health information seeking behaviour of alternating between focused and exploratory search [28]. We facilitate exploration by providing views of document content – in terms of key concepts, relevant themes, and relations of medical interest observed in the articles – that provide a user with deeper insight into retrieved articles.

2 System Overview

The system adopts several well-established techniques, integrating them in a novel manner. After standard information retrieval based on query analysis, the dashboard represents the current active collection with three distinct *interactive views*. This draws on insights from research in information visualisation that demonstrate the value of multiple coordinated views of documents, with a specific emphasis on visually illustrating connections between entities [18, 32].

Table 1. Examples of extracted PICO textual spans and MeSH terms found in them. The PICO concepts we use are the PICO-typed MeSH terms (e.g. Vaccines+Intervention).

PICO snippet	PICO cat.	MeSH terms
Patients presenting with RTI	Population	Patients; Respiratory Tract Infections
Mass vaccination campaigns with parenteral vaccines	Intervention	Immunization Programs; Vaccines; Parenteral Nutrition
Cumulative COVID-19-related hospitalization and death rates	Outcome	Hospitalization; Mortality

Data: CORD-19 is currently the most extensive coronavirus literature corpus publicly available [36]. The dataset contains all COVID-19 and coronavirus-related research (e.g. SARS, MERS, etc.) from sources including PubMed Central full text articles, and bioRxiv and medRxiv pre-prints. As of 6 June it consisted of more than 130k documents.

Information Retrieval: Article retrieval is powered by an existing search engine developed for the CORD-19 dataset, COVIDEX [37]. After submitting a query, a list of retrieved documents is shown. Each document entry can be expanded to display its abstract as well as metadata (authors, journal, source, year, license). The user can also filter by criteria such as year and source. Articles can be selected and added to a collection, the set of documents a user wishes to keep track of, which can be visualised, versioned, and exported.

We also support *semantic search*, where search criteria can be defined in terms of the typed medical concepts we also use for our relational concept view (see below); boolean matching is used in this retrieval approach.

2.1 Visual Overviews

The first view is a **relational concept view** in which we organise the medical concepts found in the articles according to key categories of evidence-based medicine, known as PICO [30] (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. We use an example based on the query *incubation period of COVID-19* to illustrate this functionality (Fig. 1). This view is a Sankey diagram frame, which shows which medical concepts are identified within PICO statements in the articles and illustrates how they co-occur in specific documents in the retrieved results.

To detect PICO statements, we train a BiLSTM-CRF model [22] on the EBM-NLP dataset [26] containing reports of randomised clinical trials annotated with textual spans that describe the PICO elements. As pretrained word representations for the model, we use 200-dimensional word2vec embeddings induced on PubMed abstracts and MEDLINE articles [19], obtaining comparable results to published figures. We then recognise medical terms from Medical Subject Headings (MeSH), a structured vocabulary maintained by the US National Library of Medicine, using the MetaMap tool [6,15]. Examples of extracted PICO concepts are shown in Table 1. In the Sankey diagram, we display pairwise relations based on article co-occurrence of Population–Intervention and Intervention–Outcome concepts.

Table 2. Comparison of COVID-SEE (1*) with related systems. (2) SciSight [20], (3) DOC Search [4], (4) COVID-19 Navigator [2], (5) LitCOVID [17], (6) SemViz [33], (7) WellAI [14], (8) COVID Intelligent Search [3], (9) Le Bras et al. [23], (10) COVID-19 LOVE [1], (11) Trialstreamer [27].

		(1*)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Search	NL/IR	✓	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓
	Concepts	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	PICO	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓
Visualisation	Concepts	✓	?	?	✗	✗	✓	✗	✗	✗	✗	?
	Relations	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
	Topics	✓	✗	✗	✗	✗	✗	✗	✗	?	✗	✗

The second, **topic view** (Fig. 2), is thematic and shows representative topics for the current collection. We trained a global topic model on medical concepts extracted from CORD-19. using Latent Dirichlet allocation (LDA) [16] to learn topics over the whole dataset, and display the topics in the retrieved subset of articles visually [9]. LDA represents each document as a mixture of topics, and each topic as a mixture of words. We chose 20 topics as optimal based on the C_v topic coherence measure [31].

Our third component is a **concept cloud view** (Fig. 2, Inset), showing the 20 most representative concepts for each active document in a wordcloud [11]. Concepts here correspond to pre-identified medical terms from the Unified Medical Language System (UMLS [24]), extracted using MetaMap [6]. To select discriminative concepts, concept distributions of articles in the collection are compared to those in the data set as a whole using the log-likehood test [29]. Analysis is done over concepts rather than words, thereby capturing multi-word terms such as *intensive care unit*.

2.2 Technical Details

All data is stored in the graph database neo4j [7]. The front-end of our web application accesses it via the Cypher language and the py2neo library [8]. The website was built with React [10] and Flask [5], and topic visualisations are supported by pyLDAVis [9].

A screencast of the system can be viewed at https://youtu.be/vL_tXuTz-LU.

3 Conclusions and Future Work

COVID-SEE is designed to facilitate more interactive exploration of the COVID-19 literature, through integration of sub-collection thematic analysis, document-level visual concept summaries, and PICO-structured concept relations. Documents retrieved for a query are visually summarised through the relational and topic views, and the salient concepts in individual documents are highlighted through the word cloud views. Our system goes beyond other systems by coupling the relational structure of medical literature with collection-level visual summaries.

In future work, a recommendation system for articles which have similar topic distributions could be added. For visual representations, we will experiment with expanding beyond the MeSH term vocabulary to include more specific terminology, and more effective use of the hierarchical relationships that exist between terms. Finally, we are planning a user study with medical professionals to evaluate the potential of COVID-SEE as a knowledge discovery tool.

Acknowledgements. This research was conducted by the Australian Research Council Training Centre in Cognitive Computing for Medical Technologies (project number ICI70200030) and funded by the Australian Government.

References

1. COVID-19 Living OVerview of Evidence. https://app.iloveevidence.com/loves/5e6fdb9669c00e4ac072701d?utm=aile. Accessed 30 Oct 2020
2. COVID-19 Navigator. https://covid-19-navigator.mybluemix.net/. Accessed 30 Oct 2020
3. COVID Intelligent Search. https://covidsearch.sinequa.com. Accessed 30 Oct 2020
4. DOC Search. https://covid-search.doctorevidence.com. Accessed 30 Oct 2020
5. Flask application development framework. https://flask.palletsprojects.com/. Accessed 30 Oct 2020
6. MetaMap. http://metamap.nlm.nih.gov. Accessed 30 Oct 2020
7. neo4j graph database. http://neo4j.com. Accessed 30 Oct 2020
8. py2neo Python library. http://py2neo.org. Accessed 30 Oct 2020
9. pyLDAvis package. https://github.com/bmabey/pyLDAvis. Accessed 30 Oct 2020
10. React Javascript framework. https://reactjs.org/. Accessed 30 Oct 2020
11. React wordcloud package. https://github.com/chrisrzhou/react-wordcloud. Accessed 30 Oct 2020
12. SciSight. https://scisight.apps.allenai.org. Accessed 30 Oct 2020
13. SemViz. https://www.semviz.org/. Accessed 30 Oct 2020
14. WellAI. https://wellai.health/covid/. Accessed 30 Oct 2020
15. Aronson, A.R., Lang, F.M.: An overview of metamap: historical perspective and recent advances. J. Am. Med. Inform. Assoc. **17**(3), 229–236 (2010)
16. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Machine Learn. Res. **3**, 993–1022 (2003)
17. Chen, Q., Allot, A., Lu, Z.: Keep up with the latest coronavirus research. Nature **579**(7798), 193 (2020). https://doi.org/10.1038/d41586-020-00694-1, https://www.ncbi.nlm.nih.gov/pubmed/32157233
18. Görg, C., et al.: Visualization and language processing for supporting analysis across the biomedical literature. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pp. 420–429 (2010)
19. Hakala, K., Kaewphan, S., Salakoski, T., Ginter, F.: Syntactic analyses and named entity recognition for PubMed and PubMed central – up-to-the-minute. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pp. 102–107. Association for Computational Linguistics, Berlin, Germany (2016). https://doi.org/10.18653/v1/W16-2913, https://www.aclweb.org/anthology/W16-2913
20. Hope, T., et al.: SciSight: combining faceted navigation and research group detection for COVID-19 exploratory scientific search. bioRxiv (2020)
21. Hutson, M.: Artificial-intelligence tools aim to tame the coronavirus literature. Nature (2020). https://doi.org/10.1038/d41586-020-01733-7

22. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260–270. Association for Computational Linguistics, San Diego, California (2016). https://doi.org/10.18653/v1/N16-1030, https://www.aclweb.org/anthology/N16-1030
23. Le Bras, P., Gharavi, A., Robb, D.A., Vidal, A.F., Padilla, S., Chantler, M.J.: Visualising COVID-19 Research. arXiv:2005.06380 (2020)
24. Lindberg, D.A., Humphreys, B.L., McCray, A.T.: The unified medical language system. Yearbook Med. Inform. **2**(01), 41–51 (1993)
25. Marchionini, G.: Exploratory search: From finding to understanding. Commun. ACM **49**(4), 41–46 (2006). https://doi.org/10.1145/1121949.1121979
26. Nye, B., et al.: A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers), pp. 197–207. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1019, https://www.aclweb.org/anthology/P18-1019
27. Nye, B., Nenkova, A., Marshall, I., Wallace, B.C.: Trialstreamer: mapping and browsing medical evidence in real-time. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 63–69. Association for Computational Linguistics, Online (2020). https://www.aclweb.org/anthology/2020.acl-demos.9
28. Pang, P.C.I., Verspoor, K., Chang, S., Pearce, J.: Conceptualising healthinformation seeking behaviours and exploratory search: result of aqualitative study. Health Technol. **5**(1), 45–55 (2015).https://doi.org/10.1007/s12553-015-0096-0, https://doi.org/10.1007/s12553-015-0096-0
29. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. The Workshop on Comparing Corpora, pp. 1–6 (2000)
30. Richardson, W.S., Wilson, M.C., Nishikawa, J., Hayward, R.S., et al.: The well-built clinical question: a key to evidence-based decisions. Acp J Club **123**(3), A12–3 (1995)
31. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408 (2015)
32. Stasko, J., Görg, C., Liu, Z.: Jigsaw: supporting investigative analysis through interactive visualization. Inf. Visual. **7**(2), 118–132 (2008). https://doi.org/10.1057/palgrave.ivs.9500180
33. Tu, J., Verhagen, M., Cochran, B., Pustejovsky, J.: Exploration and Discovery of the COVID-19 Literature through Semantic Visualization. arXiv:2007.01800 (2020)
34. Verspoor, K., et al.: Covid-see: Scientific evidence explorer for covid-19 related research (2020)
35. Wang, L.L., Lo, K.: Text mining approaches for dealing with the rapidly expanding literature on COVID-19. Brief. Bioinform. (2020). https://doi.org/10.1093/bib/bbaa296, https://doi.org/10.1093/bib/bbaa296
36. Wang, L.L., et al.: CORD-19: the COVID-19 open research dataset. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. Association for Computational Linguistics, Online (2020). https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1
37. Zhang, E., Gupta, N., Nogueira, R., Cho, K., Lin, J.: Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset: Preliminary Thoughts and Lessons Learned. arXiv:2004.05125 (2020)