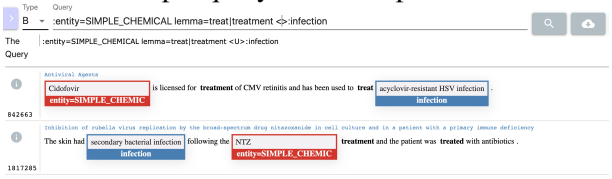
Interactive Extractive Search over Biomedical Corpora

Hillel Taub-Tabib[1] **Micah Shlain**[1,2] **Shoval Sadde**[1] **Dan Lahav**[3]
Matan Eyal[1] **Yaara Cohen**[1] **Yoav Goldberg**[1,2]
[1] Allen Institute for AI, Tel Aviv, Israel
[2] Bar Ilan University, Ramat-Gan, Israel
[3] Tel Aviv University, Tel-Aviv, Israel

We present a search system that works in a paradigm we call Extractive Search, and which allows rapid information seeking queries that are aimed at extracting facts, rather than documents. Versions of the system which facilitates queries over the CORD-19 corpus (Wang et al., 2020) as well as over a corpus of all PUBMED abstracts is publicly available at `https://allenai.github.io/spike`. The system is based on the following components.

Expressive Query Languages Our system combines three query modes: boolean, sequential and syntactic, targeting different stages of the analysis process, and different extraction scenarios. Boolean queries are the most standard, and look for the existence of search terms, or groups of search terms, in a sentence, regardless of their order. These are very powerful for finding relevant sentences, and for co-occurrence searches. Sequential queries focus on the order and distance between terms. They are intuitive to specify and are very effective where the text includes "anchor-words" near the entity of interest. Lastly, syntactic queries focus on the linguistic constructions that connect the query words to each other. Syntactic queries are powerful, and can work also where the concept to be extracted does not have clear linear anchors. To simplify their syntax and use we make use of the specification-by-example interface introduced in (Shlain et al., 2020).

Linguistic Information, Captures, and Expansions. Each of the three query types are linguistically informed, and the user can condition not only on the word forms, but also on their lemmas, parts-of-speech tags, and identified entity types. The user can also request to *capture* some of the search terms, and to *expand* them to a linguistic context. For example, in a boolean search query looking for a sentence that contains the lemmas "treat" and "treatment" ('lemma=treat|treatment'), a chemical name ('entity=SIMPLE_CHEMICAL') and the word "infection" ('infection'), a user can mark the chemical name and the word "infection" as *captures*. This will yield a list of chemical/infection pairs, together with the sentence from which they originated, all of which contain the words relating to treatments. Capturing the word "infection" is not very useful on its own: all matches result in the exact same word. But, by *expanding* the captured word to its surrounding linguistic environment, the captures list will contain terms such as "PEDV infection", "acyclovir-resistant HSV infection" and "secondary bacterial infection". The snippet below shows such an example query, and a couple of its results:



Running this query over PubMed allows us to create a large and relatively focused list in just a few seconds.

Sentence Focus, Contextual Restrictions. As our system is intended for extraction of information, it works at the sentence level. However, each sentence is situated in a context, and we allow secondary queries to condition on that context, for example by looking for sentences that appear in paragraphs that contain certain words, or which appear in papers with certain words in their titles, in papers with specific MeSH terms, in papers whose abstracts include specific terms, etc.

Interactive Speed. Central to the approach is an indexed solution, based on (Valenzuela-Escárcega et al., 2020), that allows to perform all types of queries efficiently over very large corpora, while getting results almost immediately. This allows users to interactively refine their queries and improve them based on the feedback from the results.

References

Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. 2020. Syntactic search by example. In *Proceedings of ACL 2020, System Demonstrations*.

Marco A. Valenzuela-Escárcega, Gus Hahn-Powell, and Dane Bell. 2020. Odinson: A fast rule-based information extraction framework. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France. European Language Resources Association (ELRA).

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*, abs/2004.10706.