Combining Neural and Pattern-Based Similarity Search

Domain experts often face the need to efficiently search for specific kind of information in a large collection of documents. In this work, we propose a method to do so that combines exact-match search over symbolic structures, with the ability of modern neural models to provide rich semantic representations which generalize over surface forms.

Consider, for example, an epidemiologist who is looking for research findings on the relation between human-to-human transmission of the COVID19 virus and its persistence on surfaces. The recently proposed SPIKE system (Shlain et al., 2020) allows querying for such information directly, using a query interface that minimizes the need for background in IE and linguistics. SPIKE contains several search modules among them an implicit search-by-example over syntactic structures.

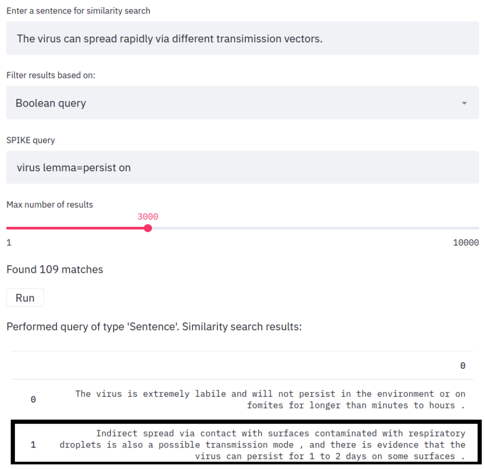
Yet, the syntactic-search approach has several limitations. The search over syntactic structure is only a proxy of the ideal objective of *semantic* search: syntax-based search is sensitive to often irrelevant structural alternations such as passive vs. active, which need to be accounted for specifically and manually in rule-based systems Moreover, syntax is sparse, and it is unlikely to witness the very same structure twice. This sparsity limits the complexity of the syntactic query. Finally, users do not necessarily know in advance which kind of example-queries is suitable for their goal. Neural similarity search, on the other hand, suffers from other limitations. while powerful, it is unstructured and "underdetrmined": in example-based search, the need often arises to specify exactly what aspect of similarity the query should focus on. Doing it with distributed neural representations is nontrivial.

We propose to combine the simple interface and useful functionalities of SPIKE search with the powerful distributed representations of neural models. Specifically, we make use of SciBERT (Beltagy et al., 2019), a transformer-based model which was trained on scientific texts. We collected the representations of the model over the CORD dataset (Wang et al., 2020), a collection of tens of thousands of COVID-related research papers, containing millions of sentences. We provide the following functionalities:

- Sentence-based search, where we search for sentences that are represented similarly to the user-inputted sentence.
- A neural re-ranking of syntactic search results, where the user inputs both a sentence and a SPIKE syntactic query. We retrieve the results of the syntactic query, and sort them by their cosine similarity to the representation of the input sentence.
- Syntactic-search-assisted query augmentation, where we first retrieve the results of the user-provided SPIKE query, and then look for sentences with similar SciBERT representation.
- K-means clustering of the results, which can aid in unsupervised extraction of semantically-meaningful topical clusters.

Those functionalities allow combining between the versatile and powerful semantic neural representations, which are often insensitive to lexical choices or surface form, and the ability to perform an exact-match search over symbolic syntactic structures. The SPIKE-based search can help a user to automatically find *alternative* surface forms which express the same topical focus as of their initial query, and can thus help to improve recall, and the spike-filtration of sentence-based search results can allow controlling for the "underdetermined" nature of neural similarity search, by limiting the results to those that express a specific symbolic structure. In Figure 1 we demonstrate those advantages by a set of queries that aim to attain the goal of finding information on the relation between persistence and spread. The natural-language query sentence "The virus can spread rapidly via different transmission vectors" provides the desired topical focus, while the boolean SPIKE query "virus lemma=persist on" limits the results to those that mention persistence. The 2nd result in this case is a relevant sentence discussing transmission via droplets that persist on surfaces.

In Figure 2 we present one the clusters found by K-means clustering of the sentences closest in the representation to the same sentence query, "The virus can spread rapidly via different transmission vectors". This cluster is relatively topically-coherent, and deals with animal hosts of the virus. We note the lexical and syntactic diversity of the sentences that describe essentially the same phenomenon: "...are accidental hosts of the virus"; "the reservoir of the virus is..."; "...are the principal reservoir for the virus"; "the natural hosts of the virus are..."; "...are lifelong carriers and shedders of the virus"; and more. Designing exact queries that would capture this diversity while maintaining semantic relevance is a nontrivial task.



Enter a sentence for similarity search

The virus can spread rapidly via different transimission vectors.

Filter results based on:

Boolean query

SPIKE query

virus lemma=persist on

Max number of results

3000

1

10000

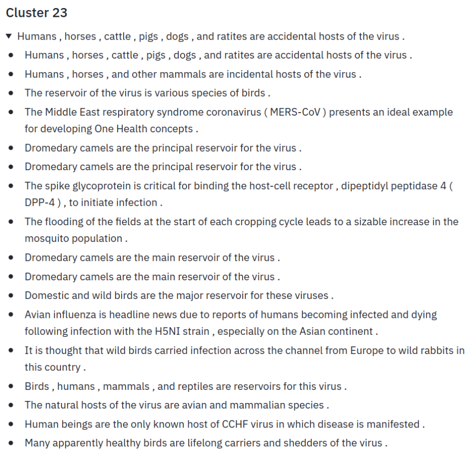
Found 109 matches

Run

Performed query of type 'Sentence'. Similarity search results:

	0
0	The virus is extremely labile and will not persist in the environment or on fomites for longer than minutes to hours .
1	Indirect spread via contact with surfaces contaminated with respiratory droplets is also a possible transmission mode , and there is evidence that the virus can persist for 1 to 2 days on some surfaces .

Figure 1: Filtration of neural similarity search with a SPIKE boolean query.



Cluster 23

▾ Humans , horses , cattle , pigs , dogs , and ratites are accidental hosts of the virus .

- Humans , horses , cattle , pigs , dogs , and ratites are accidental hosts of the virus .
- Humans , horses , and other mammals are incidental hosts of the virus .
- The reservoir of the virus is various species of birds .
- The Middle East respiratory syndrome coronavirus (MERS-CoV) presents an ideal example for developing One Health concepts .
- Dromedary camels are the principal reservoir for the virus .
- Dromedary camels are the principal reservoir for the virus .
- The spike glycoprotein is critical for binding the host-cell receptor , dipeptidyl peptidase 4 (DPP-4) , to initiate infection .
- The flooding of the fields at the start of each cropping cycle leads to a sizable increase in the mosquito population .
- Dromedary camels are the main reservoir of the virus .
- Dromedary camels are the main reservoir of the virus .
- Domestic and wild birds are the major reservoir for these viruses .
- Avian influenza is headline news due to reports of humans becoming infected and dying following infection with the H5NI strain , especially on the Asian continent .
- It is thought that wild birds carried infection across the channel from Europe to wild rabbits in this country .
- Birds , humans , mammals , and reptiles are reservoirs for this virus .
- The natural hosts of the virus are avian and mammalian species .
- Human beings are the only known host of CCHF virus in which disease is manifested .
- Many apparently healthy birds are lifelong carriers and shedders of the virus .

Figure 2: One of the resulting clusters from K-means clustering of the sentences most similar to the input sentence.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.

Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. 2020. Syntactic search by example.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: the covid-19 open research dataset. *CoRR*, abs/2004.10706.