A smart literature exploration environment for COVID-19 literature.

http://covid19.ccg.unam.mx:82/

The Problem.

Historically the vast amount of knowledge that experts publish has been increasing in such a pace that keeping up to date and having a full perspective, even in particular topics, has become quite challenging.

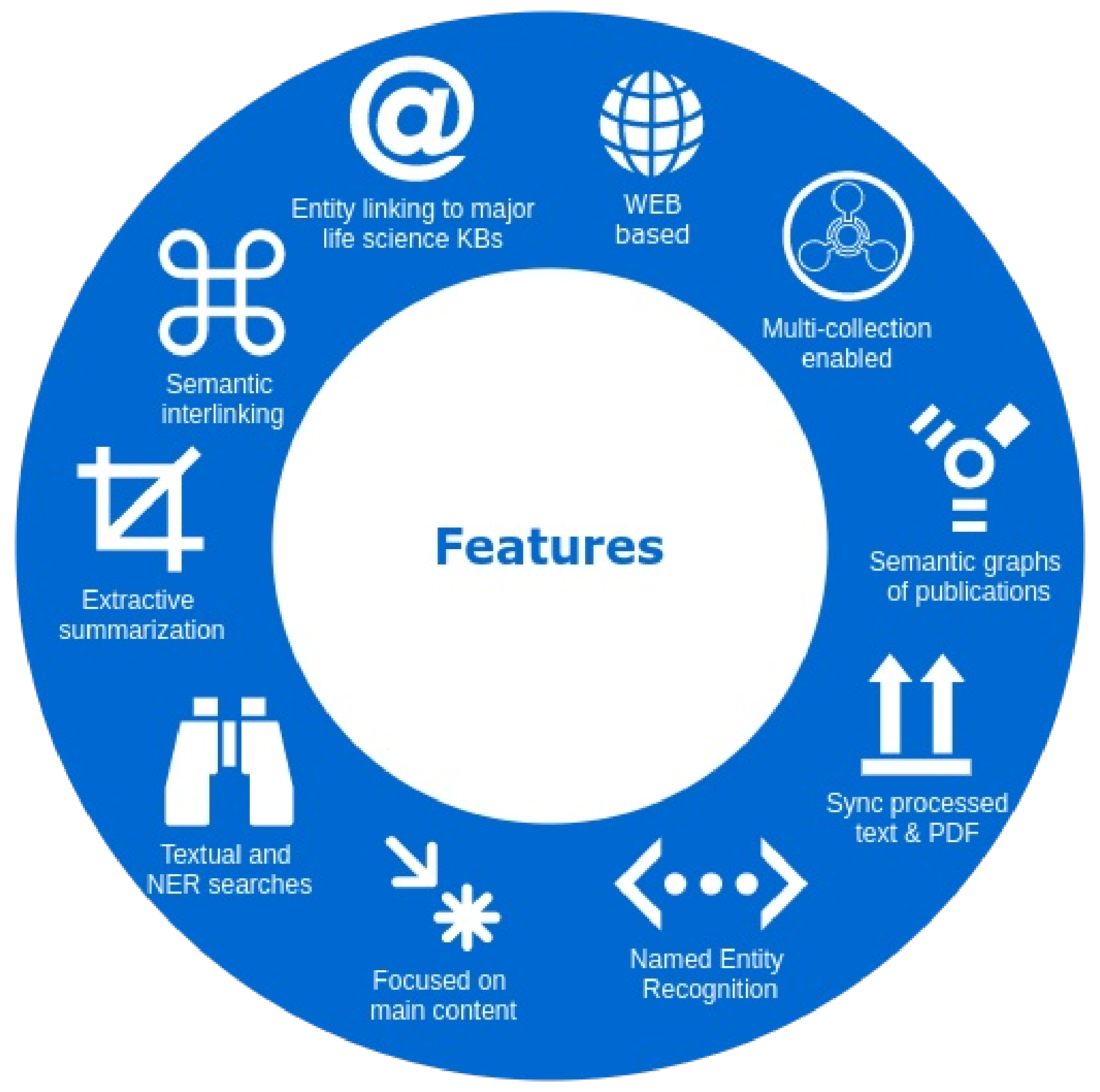
Such is the case of the current COVID-19 pandemic were there are so many clinical notes, experiments, expert observations around the world that doctors, researchers, and public authorities struggle to explore pieces of related but not explicitly connected knowledge concerning to their respective duties.

HOW WE TACKLE IT

A EXPLORATION ENVIRONMENT. *We propose a smart literature analysis environment, which includes several NLP-powered components to enable a more efficient reading process. The following two strategies are the core of our environment.*

TRANSVERSAL READING. *We propose a semantically-guided transversal reading. We believe that this type of reading can significantly benefit the process of grasping the prominent opinion and state-of-the-art of a particular aspect. Our strategy to provide this feature was to interlink all semantically related sentences by semantic-textual-similarity (STS).*

SEMANTIC ENRICHMENT. *We enrich the literature with named-entity recognition and disambiguation (NERD), using the major life science databases as entity sources, enable named-entity searches, provide network-graphs of the most interconnected publications and, an interactive tool to highlight the most central statements within an article.*



COOPERATING GROUPS

THE SWISS AI LAB IDSIA

CENTER OF GENOMIC SCIENCES, UNAM, M'EXICO

NATIONAL INSTITUTE OF RESPIRATORY DISEASES, MEXICO

ABOUT US

ALEJANDRA LOPEZ-FUENTES. *CCG, UNAM. alejandra.lof@gmail.com*

OSCAR LITHGOW-SERRANO. *IDSIA, SUPSI. oscarwilliam.lithgow@idsia.ch*

FABIO RINALDI. *IDSIA, SUPSI. fabio.rinaldi@idsia.ch*

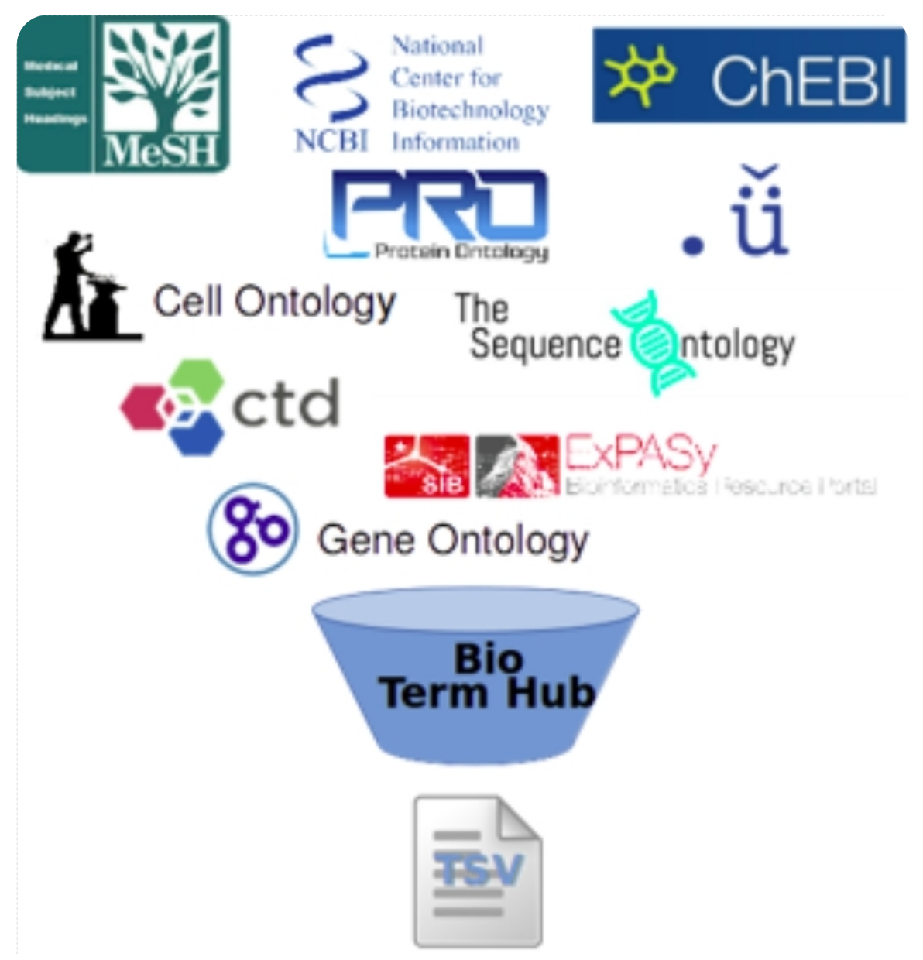
YALBI BALDERAS-MARTÍNEZ. *INER. yalbibalderas@gmail.com*

JULIO COLLADO-VIDES. *CCG, UNAM. collado@ccg.unam.mx*

Methodology.

NAMED ENTITY RECOGNITION AND DISAMBIGUATION (NERD). *These capabilities are provided by OGER, a state-of-the-art biomedical NER annotator which in turn depends on the Bio Term Hub (BTH). BTH is a combined terminological resource created by dynamically sourcing entity names and their identifiers from reference databases.*

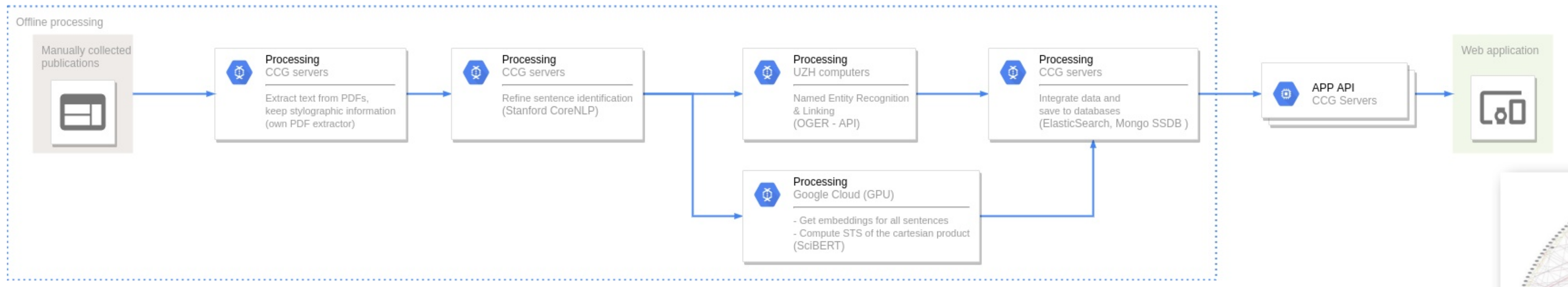
The OntoGene's Biomedical Entity Recogniser (OGER) is a RESTful web service implemented on top of the BTH which allows a remote user to batch annotate a collection of documents.



SEMANTIC TEXTUAL SIMILARITY (STS). *Our approach to measure STS is representing the sentences as embeddings and then use the cosine between two embeddings as their semantic similarity.*

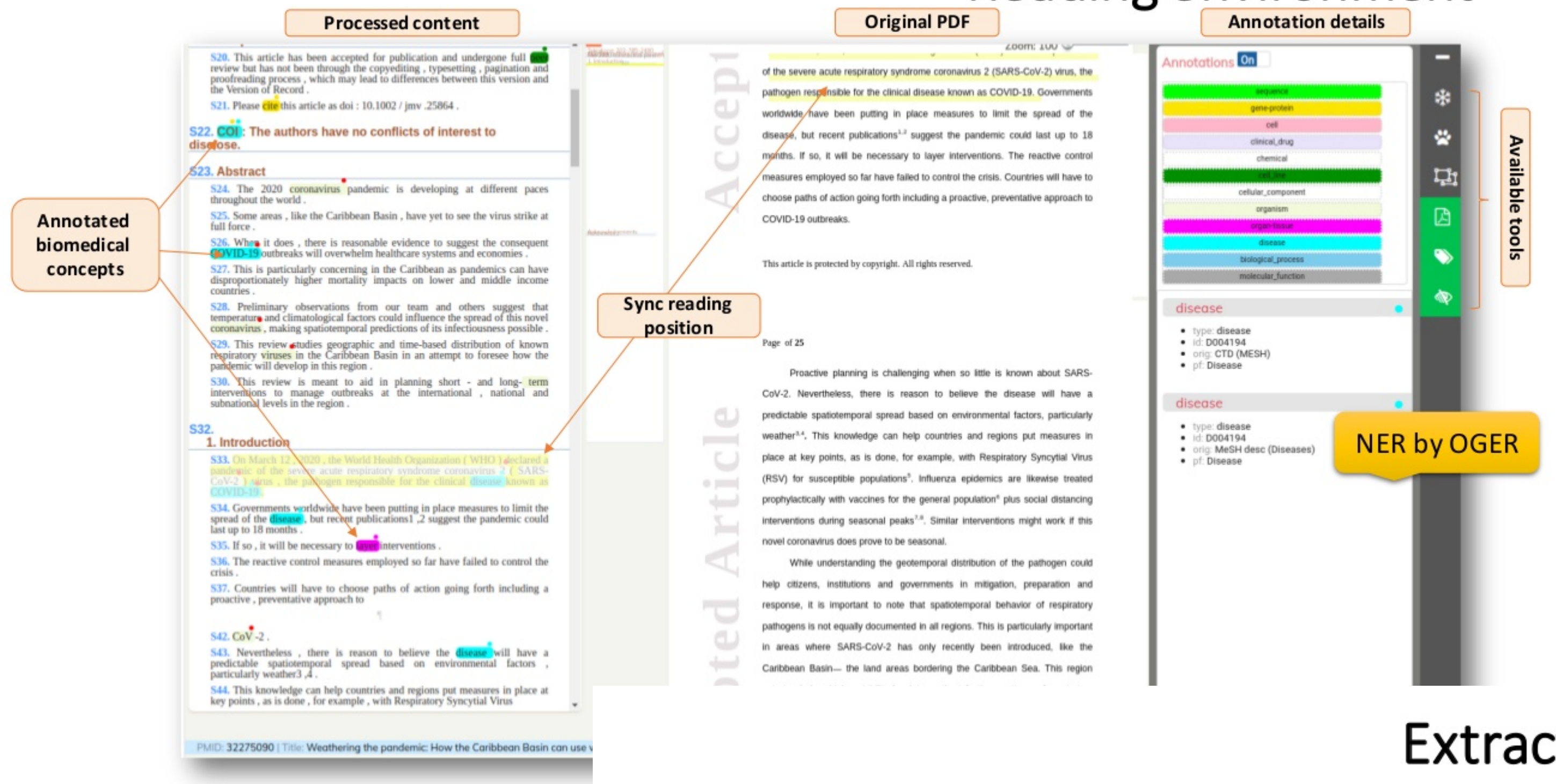
To compute the embeddings we used SciBERT, an unsupervised transformer language model pre-trained in the scientific literature. First, we map tokens to embeddings and then apply mean pooling to get fixed-sized sentence vectors.

Due to the lack of STS corpora specific to the COVID-19 literature we did not apply any fine-tuning..

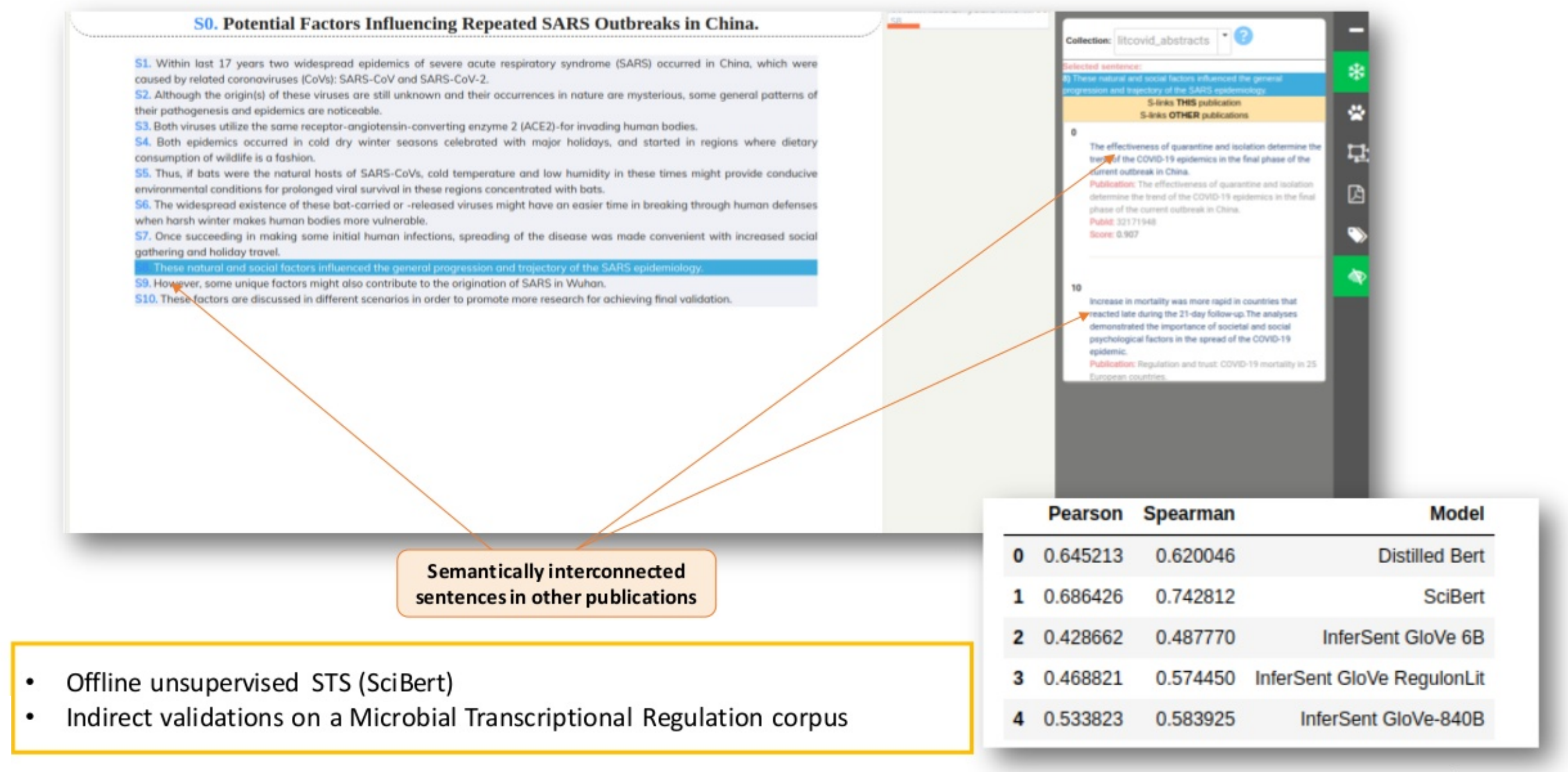

Offline processing
Manually collected publications
Processing
CCG servers
Extract text from PDFs, keep stylographic information (own PDF extractor)
Processing
CCG servers
Refine sentence identification (Stanford CoreNLP)
Processing
UZH computers
Named Entity Recognition & Linking (OGER - API)
Processing
CCG servers
Integrate data and save to databases (ElasticSearch, Mongo SSDB)
Processing
Google Cloud (GPU)
- Get embeddings for all sentences
- Compute STS of the cartesian product (SciBERT)
APP API
CCG Servers
Web application

Results.

Reading environment



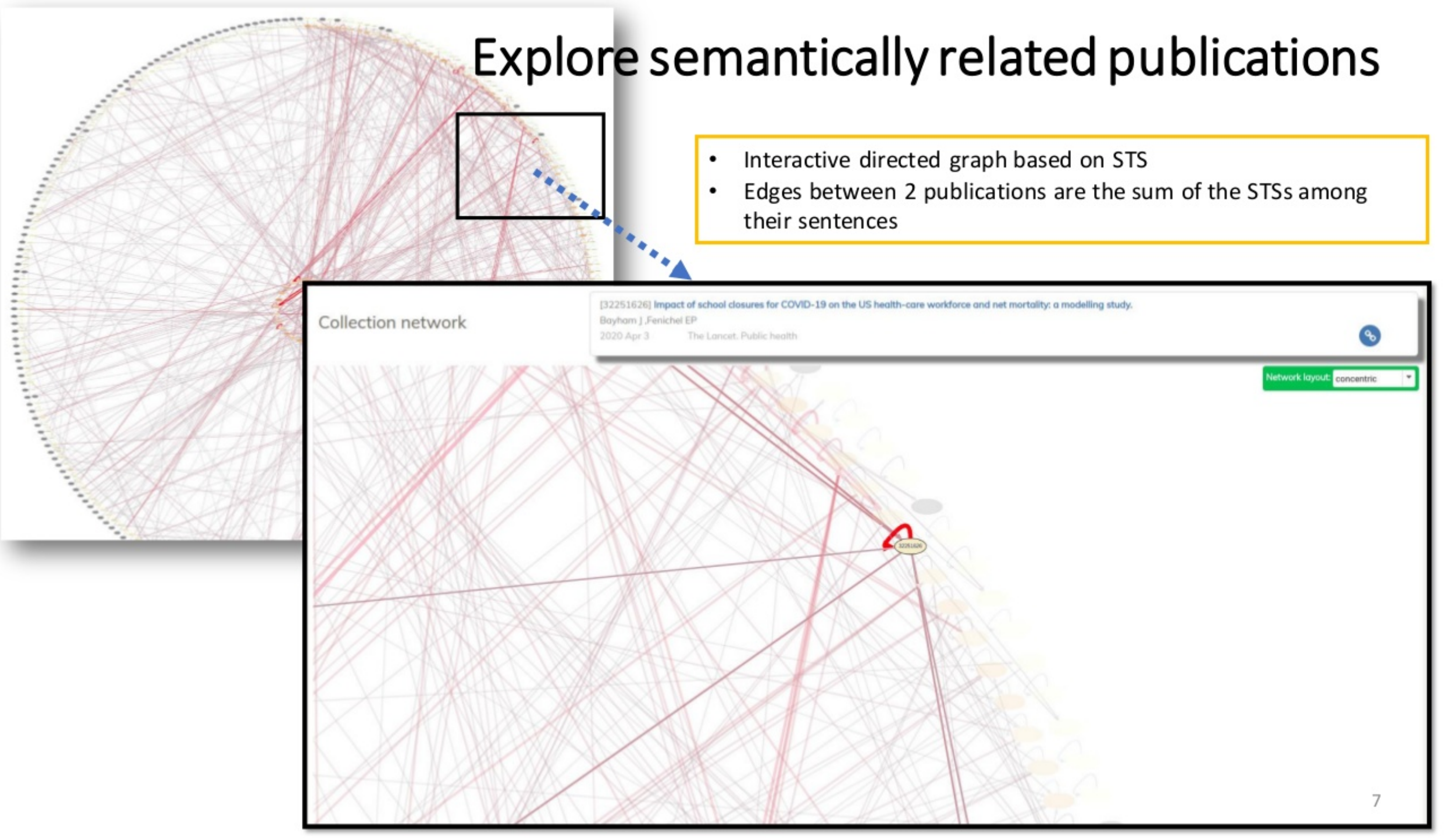
Browse articles through semantically similar sentences


S0. Potential Factors Influencing Repeated SARS Outbreaks in China.
Semantically interconnected sentences in other publications

	Pearson	Spearman	Model
0	0.645213	0.620046	Distilled Bert
1	0.686426	0.742812	SciBert
2	0.428662	0.487770	InferSent GloVe 6B
3	0.468821	0.574450	InferSent GloVe RegulonLit
4	0.533823	0.583925	InferSent GloVe-840B

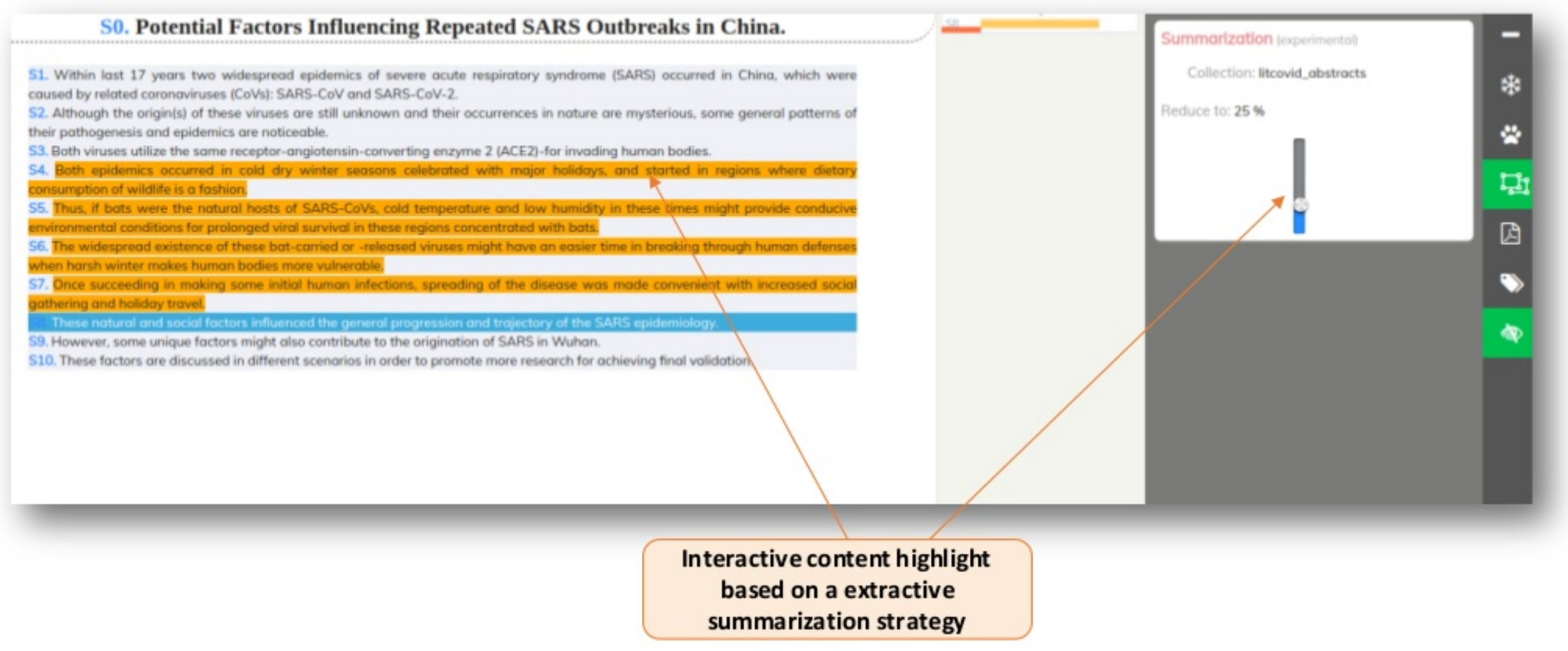
- Offline unsupervised STS (SciBert)
- Indirect validations on a Microbial Transcriptional Regulation corpus

Explore semantically related publications


Collection network
[32251626] Impact of school closures for COVID-19 on the US health-care workforce and net mortality: a modelling study.
Bayham J ,Fenichel EP
2020 Apr 3
The Lancet. Public health
Network layout: concentric
7

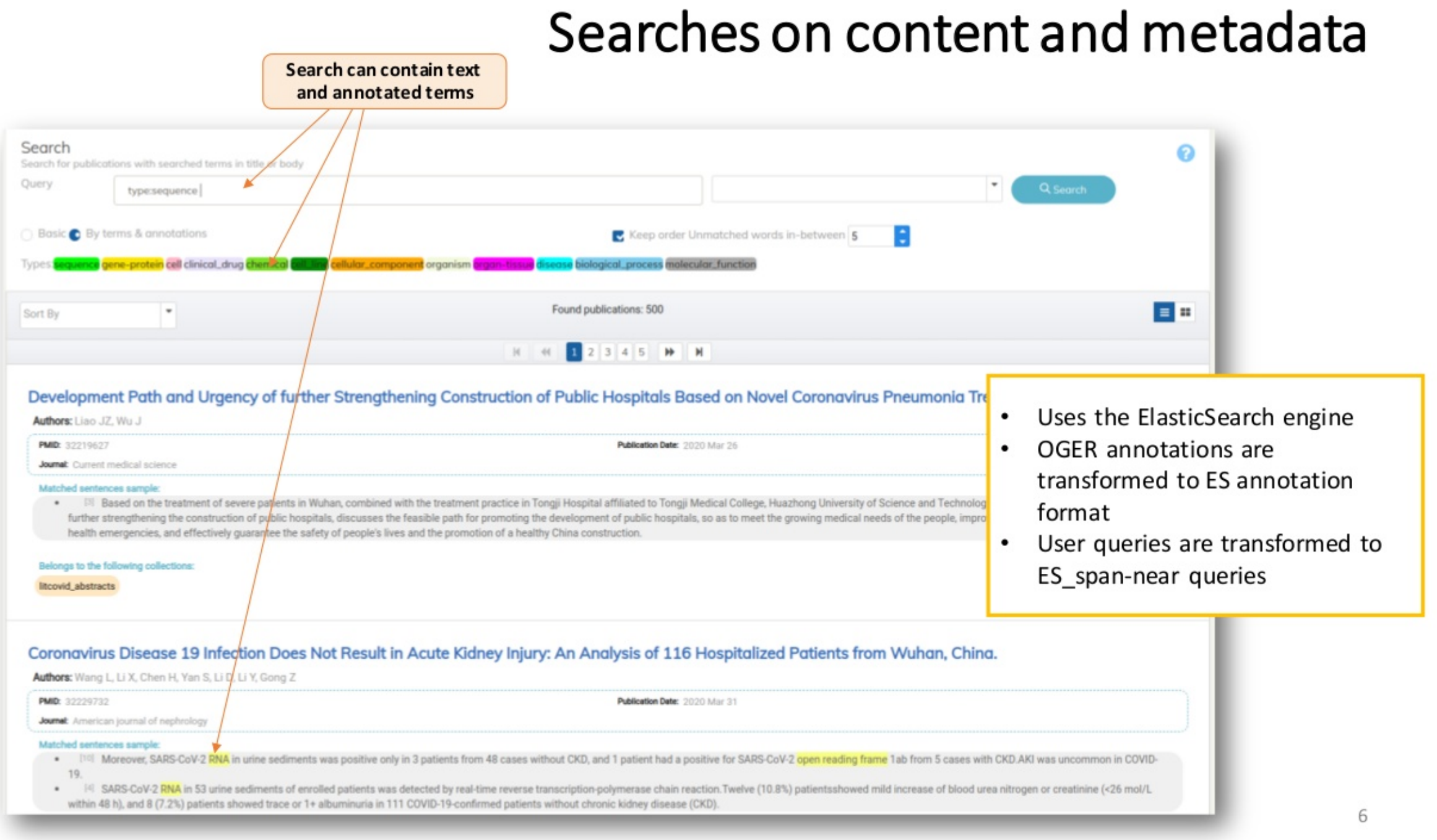
- Interactive directed graph based on STS
- Edges between 2 publications are the sum of the STSs among their sentences

Extractive summarization

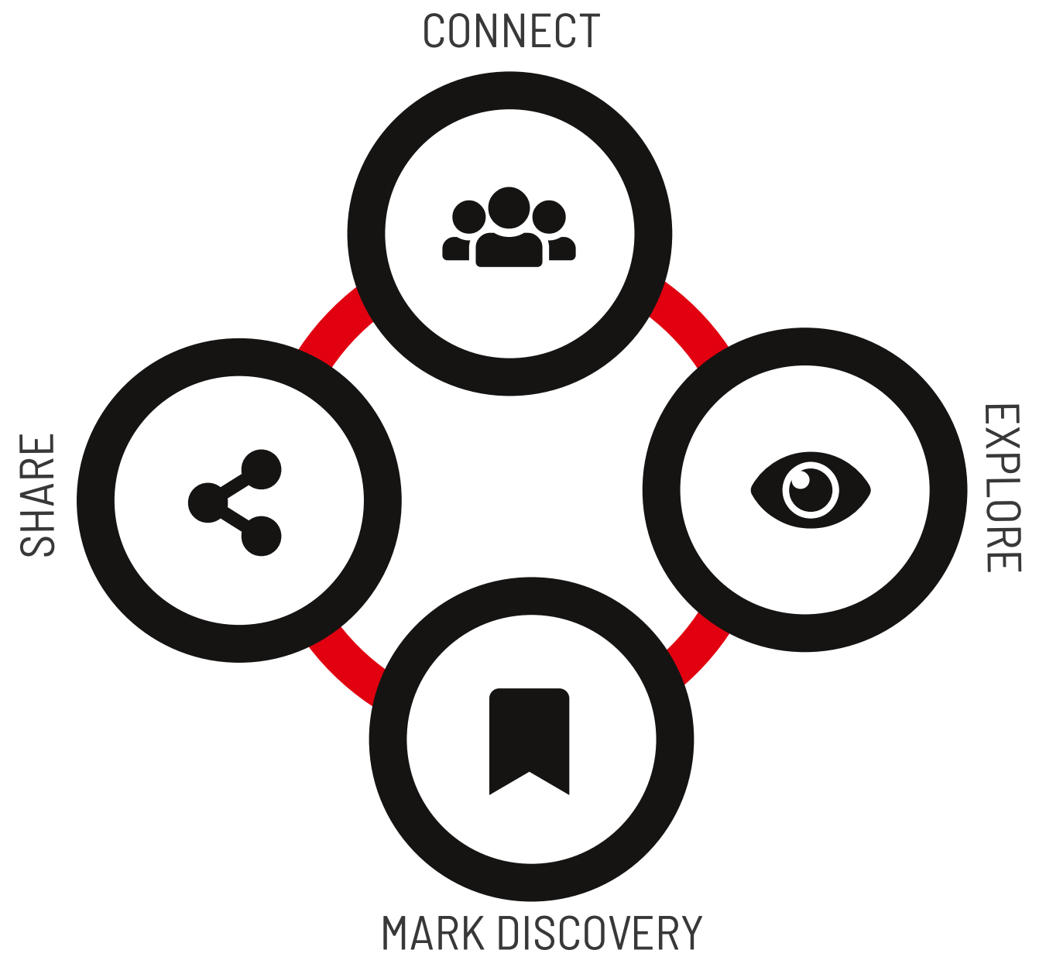

S0. Potential Factors Influencing Repeated SARS Outbreaks in China.
Interactive content highlight based on a extractive summarization strategy

- Select the sentences based in their centrality within the publication content
- Strategy inspired on the Textual Energy[1] of the sentence interactions

Searches on content and metadata


Search can contain text and annotated terms
Search
Search for publications with searched terms in title or body
Search
Development Path and Urgency of further Strengthening Construction of Public Hospitals Based on Novel Coronavirus Pneumonia Tre
Coronavirus Disease 19 Infection Does Not Result in Acute Kidney Injury: An Analysis of 116 Hospitalized Patients from Wuhan, China.
6

- Uses the ElasticSearch engine
- OGER annotations are transformed to ES annotation format
- User queries are transformed to ES_span-near queries



DISCOVERY PATHS