

CovRelex: A COVID-19 Retrieval System with Relation Extraction

Vu Tran¹ Van-Hien Tran² Phuong Minh Nguyen¹ Chau Minh Nguyen¹
Ken Satoh³ Yuji Matsumoto⁴ Minh Le Nguyen¹

¹Japan Advanced Institute of Science and Technology
{vu.tran, phuongnm, chau.nguyen, nguyenml}@jaist.ac.jp

²Nara Institute of Science and Technology, Japan
tran.van.hien.ts1@is.naist.jp

³National Institute of Informatics, Japan
ksatoh@nii.ac.jp

⁴RIKEN Center for Advanced Intelligence Project (AIP), Japan
yuji.matsumoto@riken.jp

Abstract

This paper presents CovRelex, a scientific paper retrieval system targeting entities and relations via relation extraction on COVID-19 scientific papers. This work aims at building a system supporting users efficiently in acquiring knowledge across a huge number of COVID-19 scientific papers published rapidly. Our system can be accessed via <https://www.jaist.ac.jp/is/labs/nguyen-lab/systems/covrelex/>.

Keywords: COVID-19, biomedical domain, scientific paper analysis, relation extraction, entity recognition, document retrieval.

1 Introduction

This work aims at facilitating knowledge acquisition from a huge number of COVID-19 scientific papers. Due to the COVID-19 outbreak, researchers have been focusing on studying the virus and publishing a huge number of papers rapidly. According to the estimation of Silva et al. (2020), 23,634 unique documents were published in just 6 months between January 1st and June 30th, 2020. In the records of the COVID-19 Open Research Dataset (CORD-19) Challenge¹, the number of collected papers about COVID-19, SARS-Cov-2 and related coronaviruses is more than 400K by January 9th, 2021. The rapid speed of new publication and the huge number of related papers challenges specialists to seek knowledge by connecting findings across papers efficiently and timely.

¹<https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>

When focusing on knowledge acquisition of biomedical entities, several questions can be asked regarding the entities and their relations:

- Which papers mention entity E_1 ?
- Which papers mention the relation R between entity E_1 and entity E_2 ?
- Which papers mention the relation R_1 between entity E_1 and entity E_2 , and the relation R_2 between entity E_2 and entity E_3 ?
- What relations R_x exist between entity E_1 and entity E_2 and in which papers?
- What entity E_x has relation R with entity E_1 and in which papers?

Such questions can be answered by our system.

2 Related Work

FACTA+ (Tsuruoka et al., 2011, 2008) was presented as a text search engine that helps users discover and visualize indirect associations between biomedical concepts from MEDLINE abstracts. Liu et al. (2015) introduced an online text-mining system (PolySearch2) for identifying relationships between biomedical entities over 43 million articles covering MEDLINE abstracts, PubMed Central full-text articles, Wikipedia full-text articles, US Patent abstracts, open access textbooks from NCBI and MedlinePlus articles. More recently, LitVar (Allot et al., 2018), a semantic search engine, utilized advanced text mining techniques to compute and extract relationships between genome

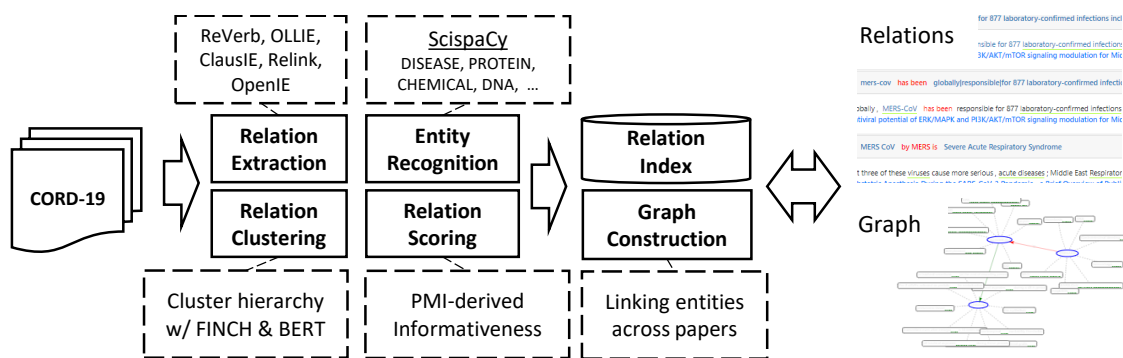


Figure 1: System Overview.

arg_1	rel	arg_2
[MERS-CoV] _{GGP}	include	[fever] _{DISEASE} , [chills/rigors] _{DISEASE} , [headache] _{DISEASE} , non-productive [cough] _{DISEASE}
[MERS-CoV] _{GGP}	is responsible for causing	lower [respiratory infections] _{DISEASE} with [fever] _{DISEASE} and [cough] _{DISEASE}

Figure 2: An example of relations extracted from COVID-19 papers.

variants and other associated entities such as diseases and chemicals/drugs. Wei et al. (2019) presented a web service PubTator Central (PTC) that provides automated bioconcept annotations in full text biomedical articles, in which bioconcepts are extracted from state-of-the-art text mining systems.

Due to the COVID-19 outbreak, it is essential to grasp valuable knowledge from a huge number of COVID-19-related papers for dealing with the pandemic effectively. Sohrab et al. (2020) introduced the BENNERD system that detects named entities in biomedical text and links them to the unified medical language system (UMLS) to facilitate the COVID-19 research. Hope et al. (2020) created a dataset annotated for mechanism relations and trained an information extraction model on this data. Then, they used the model to extract a Knowledge Base (KB) of mechanism and effect relations from papers relating to COVID-19. Zhang et al. (2020) built Covidex, a search infrastructure that provides information access to the COVID-19 Open Research Dataset such as answering questions. Esteva et al. (2020) also presented Co-Search, a retriever-ranker semantic search engine designed to handle complex queries over the COVID-19 literature. Wang et al. (2020) created the EvidenceMiner web-based system. Given a query as a natural language statement, EvidenceMiner automatically retrieves sentence-level textual evidence from the CORD-19 corpus.

Clearly, previous works made a great effort to

acquire useful knowledge from the COVID-19 literature, such as recognizing biomedical entities (Sohrab et al., 2020), extracting mechanism relations between entities (Hope et al., 2020), or retrieving relevant text segments based on the user query (Zhang et al., 2020; Wang et al., 2020). However, there is still a lack of a system that has the ability to automatically detect both entities with various types and their diverse relations through papers, especially when COVID-19 papers are published rapidly. This motivates us to build the CovRelex system, which aims to exploit such information.

3 Method

3.1 Overview

The core of our system is built from extracting an enormous number of relations from COVID-19 related scientific papers (in CORD-19 corpus) by several open domain relation extraction methods. The extracted relations are represented not only by their original form from the extraction methods but also by the contained biomedical entities. Furthermore, the relations are clustered and scored for their informativeness over the corpus (Fig. 1).

A **relation** is a triplet in the form (arg_1, rel, arg_2) , where arg_1 , and arg_2 are noun phrases which may contain biomedical entities, and rel is an expression describing the directed relation from arg_1 to arg_2 (shown in Fig. 2).

Table 1: SciSpacy models used in our system.

Name	Training Data	Entity Types
en_ner_craft_md	CRAFT	GGP, SO, TAXON, CHEBI, GO, CL
en_ner_jnlpba_md	JNLPBA	DNA, CELL_TYPE, CELL_LINE, RNA, PROTEIN
en_ner_bc5cdr_md	BC5CDR	DISEASE, CHEMICAL
en_ner_bionlp13cg_md	BIONLP13CG	AMINO_ACID, ANATOMICAL_SYSTEM, CANCER, CELL, CELLULAR_COMPONENT, DEVELOPING_ANATOMICAL_STRUCTURE, GENE_OR_GENE_PRODUCT, IMMATERIAL_ANATOMICAL_ENTITY, MULTI-TISSUE_STRUCTURE, ORGAN, ORGANISM, ORGANISM_SUBDIVISION, ORGANISM_SUBSTANCE, PATHOLOGICAL_FORMATION, SIMPLE_CHEMICAL, TISSUE

3.2 Relation Extraction

With the objective of extracting as many relations as possible, we employ several relation extraction methods. Each method has their own characteristics, thus, may extract different kinds of relations. By combining several methods, we can obtain higher extraction coverage. The methods are briefly described as follows.

- ReVerb (Fader et al., 2011) tackles the problems of incoherent and uninformative extractions by introducing constraints on binary, verb-based relation phrases.
- OLLIE (Mausam et al., 2012) addresses the problems that Open IE systems such as ReVerb only extract relations that are mediated by verbs. Not only by verbs, OLLIE extracts relations mediated also by nouns, adjectives, and more.
- ClausIE (Del Corro and Gemulla, 2013) is a clause-based approach to open information extraction. It separates the detection of clauses and clause types from the actual generation of propositions.
- Relink (Tran and Nguyen, 2020) is a method partly inherited from ReVerb, extracts relations from the connected phrases, not for identifying clause type like ClauseIE.
- OpenIE (Angeli et al., 2015) extracts relations by breaking a long sentence into short, coherent clauses, and then finds the maximally simple relations.

The extracted relations are also tagged with biomedical entities recognized by using entity recognition models presented in the next subsection.

3.3 Entity Recognition

We use biomedical entity recognition models specialized for predicting entity type and provided by SciSpacy (Neumann et al., 2019) (Table 1). Each of the models is trained on a different annotated corpus, thus, covers a different set of biomedical entities. By using multiple entity systems, we can obtain various specialized entity information: chemicals and diseases with BCD5CDR (Li et al., 2016), cell types, chemicals, proteins, and genes with CRAFT (Bada et al., 2012), cell lines, cell types, DNAs, RNAs, and proteins with JNLPBA (Collier and Kim, 2004), and cancer genetics with BioNLP13CG (Pyysalo et al., 2015).

3.4 Relation Clustering

We build a cluster hierarchy on a subset of the extracted relations (this subset contains all relations in which both arg_1 and arg_2 are biomedical entities), so users can quickly find their interested relation expressions or they can choose some clusters which may contain their interested relation expressions.

We utilize FINCH (Sarfranz et al., 2019), hierarchical clustering method, and BERT (Devlin et al., 2019) for this task. First, BERT-Base model is used to encode each relation as a simple sentence “ arg_1 *rel* arg_2 ” into a 768-dimensional vector. Then, FINCH is used to build the cluster hierarchy. For each cluster, representative expressions of the cluster are selected from its *rels* from top informative relations scored by the formula presented in the next subsection. The result cluster hierarchy is illustrated in Fig. 3.

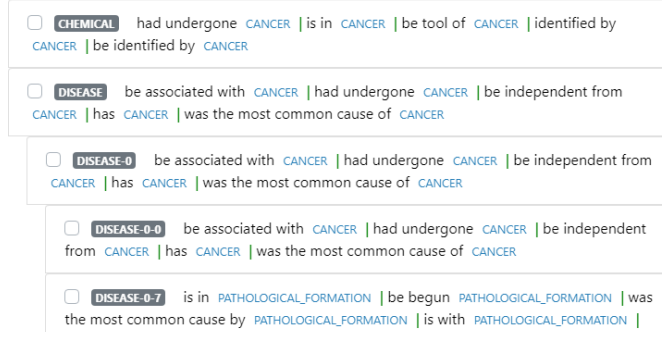


Figure 3: Illustration of cluster hierarchy. “DISEASE-0-7”: the type of an entity contained in the arg_1 is DISEASE, the id of the level 0 (root) cluster is 0, the id of the level 1 cluster is 7. An expression has the form of ENTITY TYPE (in arg_1 , omitted) relation/verb phrase ENTITY TYPE (in arg_2). Expressions are separated by |.

Figure 4 shows a search interface for a single-relation query. The search bar contains "mers-cov". The "ENTITY TYPES" dropdown is set to "DISEASE". The "Relation" field is empty, and the "Argument 2" field is set to "DISEASE". The "Matching" slider is positioned towards the "Informative" end. Below the search bar, a result is displayed: "MERS-CoV include fever, chills/rigors, headache, non-productive cough".

Figure 4: An example of **Single-Relation Query** for (mers-cov, any-relation, DISEASE).

3.5 Relation Scoring

Relations are scored for informativeness based from Pointwise Mutual Information (PMI) (Church and Hanks, 1990), the association ratio for measuring word association norms, based on the information-theoretic concept of mutual information. The informativeness of a relation (arg_1, rel, arg_2) can be regarded as PMI (Eq. 1) of two points: arg -pair $args = (arg_1, arg_2)$ and its relation expression rel through occurrence $p(\cdot)$.

$$PMI(args, rel) = \log_2 \frac{p(args, rel)}{p(args) p(rel)} \quad (1)$$

It is difficult to apply Eq. 1, which computes the occurrence by exact matching, for our system because of the variation and noise in the contents of the extracted relations. To mitigate the difficulty of using exact match, we propose to use cosine similarity with Tf-idf vectorization (Sparck Jones, 1988). While exact match counting of occurrence indicates the presence of an instance ($args$ or rel) in the relation set, our use of cosine similarity indicates the presence of the *contents* of the instance in the relation set, thus can adapt to the variation and noise in the contents of the relations.

With our approach, the relation’s informativeness $InfoScore(args, rel)$ is computed following Eq. 2.

$$InfoScore(args, rel) = \log_2 \frac{S(args, rel)}{S(args)S(rel)} \quad (2)$$

$$S(args, rel) =$$

$$\sum_{(args', rel')} \cos(v(args, rel), v(args', rel'))$$

$$S(args) = \sum_{args'} \cos(v(args), v(args'))$$

$$S(rel) = \sum_{rel'} \cos(v(rel), v(rel'))$$

where $(args', rel')$ are all relations other than $(args, rel)$, $args'$ are arg -pairs in all relations other than $(args, rel)$, rel' are expressions in all relations other than $(args, rel)$, and $v(t_1, t_2, \dots, t_n)$ is the vectorization function which concatenates the input texts t_1, t_2, \dots, t_n and converts the concatenated text into a single Tf-idf vector.

3.6 Retrieval System

The retrieval system provides two kinds of queries: **Single-Relation Query** and **Graph Query**. While **Single-Relation Query** provides simple way to

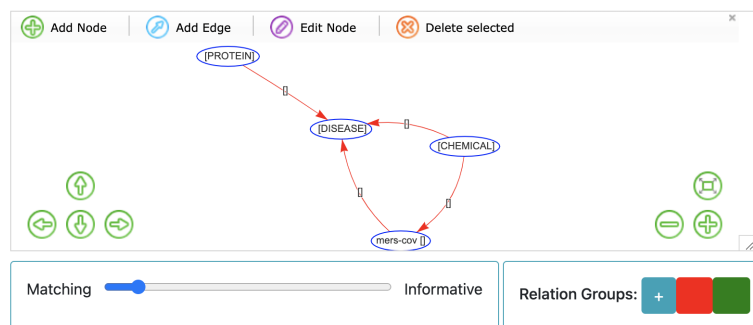


Figure 5: Graph Query: searching for a paper containing relations matching the query graph.

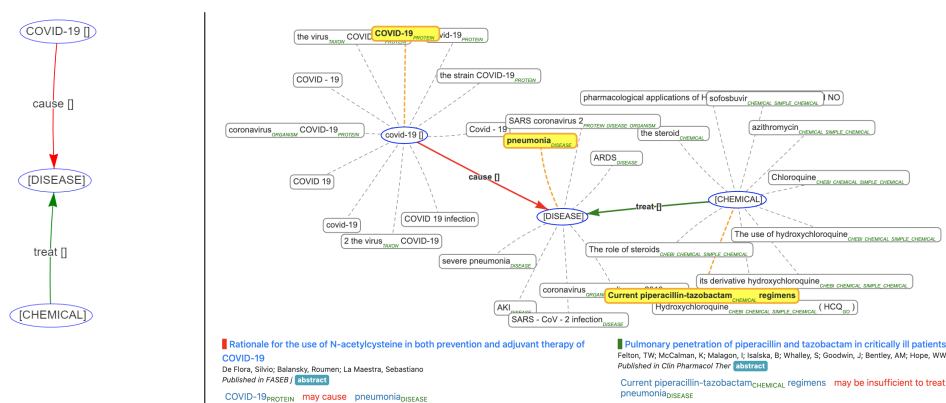


Figure 6: Example of **Multi-Paper Graph Query**. Left-hand side graph is the query. The right-hand side graph is the summary of the results showing candidate entities. The highlighted nodes of the summary graph show entities related to each other and mentioned in the two papers at the bottom.

search for specific relations, **Graph Query** provides a sophisticated way to search for papers containing entities connected in a complex relation graph.

3.6.1 Single-Relation Query

A query consists of partial information of a relation which can contain keywords about arg_1 , arg_2 , and rel , types of entities possibly included in the arg_1 or arg_2 , or clusters which the relation belongs to. The retrieved results are relevant relations with their corresponding papers. An example of **Single-Relation Query** is illustrated in Fig. 4. The query relation is (mers-cov, any-relation, DISEASE). The results are best matched relations, for instance, (MERS-CoV, include, “fever, chills/rigors, headache, non-productive cough”).

The candidate relations are retrieved based on the keyword matching score by BM25 (Schütze et al., 2008) and InfoScore (Eq. 2), then filtered by the entity types and the clusters. Keyword matching score and InfoScore can be weighed for the need of searching candidates that have high lexical matching with the query or candidates that are

highly informative.

3.6.2 Graph Query

This extends **Single-Relation Query** by enabling more sophisticated paper search covering a complex graph describing relations among entities. An example of **Graph Query** is illustrated in Fig. 5 with a query consists of 4 relations: (mers-cov, cause, DISEASE), (CHEMICAL, any-relation, mers-cov), (CHEMICAL, any-relation, DISEASE), and (PROTEIN, any-relation, DISEASE). The result graph is built from linking entities and relations obtained from each paper, which matches the query graph. The entity linking is done through lexical matching and type matching. This approach faces the challenges from entities with synonyms and performance of entity recognition.

One special feature of **Graph Query** is **Multi-Paper Graph Query** which supports searching relations across multiple papers. The important use case is that interested relations are not described in one single paper, i.e., one entity is mentioned in different papers and thus engaged in different relations. For example, if users want to “find some

Table 2: Evaluation results on relation extraction. **Correct I, II**, and **I&II**: evaluated as correct relations (can be entailed from the corresponding sentences) by the first, the second, and both the evaluators, respectively. **Overall**: evaluation on the unique relations per sentence from all methods. **Kappa**: Cohen’s kappa coefficient.

Method	Total	Correct I	Correct II	Correct I&II	Kappa
ReVerb	255	183 (72%)	224 (88%)	181 (71%)	0.47
OLLIE	398	304 (76%)	303 (76%)	275 (69%)	0.60
ClausIE	1,061	880 (83%)	760 (72%)	720 (68%)	0.47
Relink	302	210 (70%)	193 (64%)	173 (57%)	0.58
OpenIE	1,609	1,042 (65%)	901 (56%)	700 (44%)	0.30
Overall	3,477	2,479 (71%)	2,242 (64%)	1,913 (55%)	0.41

Table 3: Statistics of extracted relations.

Method	Non-uniq. /corpus	Uniq. /corpus	Uniq. /abstract
ReVerb	2.3M	1.7M	8
OLLIE	4.7M	3.6M	16
ClausIE	9.0M	6.9M	31
Relink	5.5M	4.1M	19
OpenIE	24.4M	18.6M	84
Overall	45.9M	33.3M	150

Table 4: Statistics of recognized entities.

Model	/corpus	/abstract
en_ner_craft_md	1.8M	6
en_ner_jnlpba_md	3.1M	11
en_ner_bc5cdr_md	1.8M	6
en_ner_bionlp13cg_md	1.4M	5
Total	6.4M	22

CHEMICAL that can treat some DISEASE caused by COVID-19”, they will look for two relations: (COVID-19, cause, DISEASE), and (CHEMICAL, treat, DISEASE). In that case, the two relations may be retrieved from two different papers. Therefore, aggregating information scattering over multiple papers is necessary for building a more comprehensive understanding. It is done through relation grouping allowing users to segment the query graph into several segments each belonging to different papers. With the above example, users can define a query graph (the left-hand side of Fig. 6) and our system could find that “pneumonia” is a DISEASE caused by COVID-19 and is treated with “Current [piperacillin-tazobactam]_{CHEMICAL} regimens” (the right-hand side of Fig. 6) from two separate papers, and more.

4 Results

4.1 Corpus

We performed relation extraction and entity recognition from the CORD19 corpus provided in the COVID-19 Open Research Dataset Challenge updated by January 3rd, 2021. The corpus contains \approx 400K entries to COVID-19 related papers. Relation extraction and entity recognition were performed on the abstracts of the papers.

4.2 Relation Extraction

As shown in Table 3, we extracted 40.5 million relations including 29.8 million unique relations. Among the relation extraction methods, OpenIE outputs the largest number. The other three relation extraction methods tend to output long and composite relations while OpenIE tends to break down and output shorter and simpler relations. However, OpenIE also outputs small variations of similar relations.

For assessing the quality of relation extraction, we conduct an evaluation on a small data sample consisting of 100 papers selected from the corpus. The evaluation was conducted by two human evaluators with the criteria to answer *whether the relation can be entailed from the sentence*.

The results (Table 2) show that the evaluation is a difficult task. The evaluation agreement between the two evaluators is 0.41 in term of Cohen’s kappa coefficient (McHugh, 2012). It’s considered fair agreement (Fleiss et al., 2003). Among the relation extraction methods, OLLIE yields the best kappa coefficient of 0.60 (good agreement), OpenIE yields the worst coefficient of 0.30 (poor agreement), and the others yield the coefficients of 0.47 to 0.58 (fair to good agreement). One of the possible reasons is the complexity of biomedical texts: sentences with 31 tokens in average and up

to 167 tokens in the evaluated sample, and common use of conjunctions and nested clauses.

4.3 Entity Recognition

As shown in Table 4, a total of 6.4M entities were recognized from the corpus with the four entity recognition models. For each abstract of a COVID-19 related paper, an average of 22 entities were recognized. Among the four models, **en_ner_jnlpba_md** outputs the largest number of entities, about 1.7 to 2.2 times more than the other models, where this model’s specialized entity types are cell lines, cell types, DNAs, RNAs, and proteins.

5 Conclusion

We have presented our COVID-19 scientific paper retrieval system which focuses on analysing entities and their relations. The system is empowered with several relation extraction and entity recognition methods. The system supports users in acquiring knowledge efficiently across a huge number of COVID-19 scientific papers published rapidly. There, however, exist extremely challenging problems to tackle for making the system more practical: dealing with the newly created and unknown data, solving the performance gap when utilizing present methods, and do these in the nick of time of fighting with pandemics.

Acknowledgment

This work was supported by JST CREST Grant Number JPMJCR1513, Japan.

References

Alexis Allot, Yifan Peng, Chih-Hsuan Wei, Kyubum Lee, Lon Phan, and Zhiyong Lu. 2018. [LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC](#). *Nucleic Acids Research*, 46(W1):W530–W536.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. [Leveraging linguistic structure for open domain information extraction](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor,

Judith A Blake, and Lawrence E Hunter. 2012. [Concept annotation in the CRAFT corpus](#). *BMC Bioinformatics*, 13(1).

Kenneth Ward Church and Patrick Hanks. 1990. [Word association norms, mutual information, and lexicography](#). *Computational Linguistics*, 16(1):22–29.

Nigel Collier and Jin-Dong Kim. 2004. [Introduction to the bio-entity recognition task at JNLPA](#). In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Luciano Del Corro and Rainer Gemulla. 2013. [Clausic: clause-based open information extraction](#). In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. 2020. [Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization](#).

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. [Identifying relations for open information extraction](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 2003. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, Inc.

Tom Hope, Aida Amini, David Wadden, Madeleine van Zuylén, E. Horvitz, Roy Schwartz, and Hannaneh Hajishirzi. 2020. [Extracting a Knowledge Base of Mechanisms from COVID-19 Papers](#).

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wieggers, and Zhiyong Lu. 2016. [BioCreative v CDR task corpus: a resource for chemical disease relation extraction](#). *Database*, 2016:baw068.

Yifeng Liu, Yongjie Liang, and David Wishart. 2015. [PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more](#). *Nucleic Acids Research*, 43(W1):W535–W542.

- Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. [Open language learning for information extraction](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, Jeju Island, Korea. Association for Computational Linguistics.
- Marry L. McHugh. 2012. [Interrater reliability: the kappa statistic](#). *Biochemia Medica*, pages 276–282.
- Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. [ScispaCy: Fast and robust models for biomedical natural language processing](#). In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.
- Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun’ichi Tsujii, and Sophia Ananiadou. 2015. Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013. *BMC bioinformatics*, 16(S10):S2.
- Saqib Sarfraz, Vivek Sharma, and Rainer Stiefelhaugen. 2019. Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.
- Jaime A Teixeira da Silva, Panagiotis Tsigaris, and Mohammadamin Erfanmanesh. 2020. Publishing volumes in major databases related to covid-19. *Scientometrics*, pages 1–12.
- Mohammad Golam Sohrab, Khoa Duong, Makoto Miwa, Goran Topić, Ikeda Masami, and Hiroya Takamura. 2020. [Bennerd: A neural named entity linking system for covid-19](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 182–188, Online. Association for Computational Linguistics.
- Karen Sparck Jones. 1988. A statistical interpretation of term specificity and its application in retrieval. In *Document retrieval systems*, pages 132–142. Taylor Graham Publishing.
- Xuan-Chien Tran and Le-Minh Nguyen. 2020. [ReLink: Open information extraction by linking phrases and its applications](#). In *Distributed Computing and Internet Technology*, pages 44–62. Springer International Publishing.
- Yoshimasa Tsuruoka, Makoto Miwa, Kaisei Hamamoto, Jun’ichi Tsujii, and Sophia Ananiadou. 2011. [Discovering and visualizing indirect associations between biomedical concepts](#). *Bioinformatics*, 27(13):i111–i119.
- Yoshimasa Tsuruoka, Jun’ichi Tsujii, and Sophia Ananiadou. 2008. [FACTA: a text search engine for finding associated biomedical concepts](#). *Bioinformatics*, 24(21):2559–2560.
- Xuan Wang, Weili Liu, Aabhas Chauhan, Yingjun Guan, and Jiawei Han. 2020. Automatic textual evidence mining in covid-19 literature. *arXiv preprint arXiv:2004.12563*.
- Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. [PubTator central: automated concept annotation for biomedical full text articles](#). *Nucleic Acids Research*, 47(W1):W587–W593.
- Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang, and Jimmy Lin. 2020. [Covidex: Neural ranking models and keyword search infrastructure for the COVID-19 open research dataset](#). In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 31–41, Online. Association for Computational Linguistics.