

CL Scholar: The ACL Anthology Knowledge Graph Miner

Mayank Singh, Pradeep Dogga*, Sohan Patro*, Dhiraj Barnwal*, Ritam Dutt*,
Rajarshi Haldar, Pawan Goyal and Animesh Mukherjee

Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur, WB, India
mayank.singh@cse.iitkgp.ernet.in

Abstract

We present *CL Scholar*, the ACL Anthology knowledge graph miner to facilitate high-quality search and exploration of current research progress in the computational linguistics community. In contrast to previous works, periodically crawling, indexing and processing of new incoming articles is completely automated in the current system. *CL Scholar* utilizes both textual and network information for knowledge graph construction. As an additional novel initiative, *CL Scholar* supports more than 1200 scholarly natural language queries along with standard keyword-based search on constructed knowledge graph. It answers *binary*, *statistical* and *list* based natural language queries. The current system is deployed at <http://cnerg.iitkgp.ac.in/aclakg>. We also provide REST API support along with bulk download facility. Our code and data are available at <https://github.com/CLSCholar>.

1 Introduction

ACL Anthology¹ is one of the popular initiatives of the Association for Computational Linguistics (ACL) to curate all publications related to computational linguistics and natural language processing at one common place. At present, it hosts more than 44,000 papers and is actively updated and maintained by Min Yen Kan. Since its inception, ACL Anthology functions as a repository with the collection of papers from ACL and related organizations in computational linguistics. However, it does not provide any additional statistics about authors, papers, venues, and topics. Also, it lacks advance search features such as article ranking by factoring in popularity or relevance, natural language query support, author profiles, topical search etc.

*These authors contributed equally to the study.

¹<https://aclweb.org/anthology/>

1.1 Previous systems built on ACL anthology

Owing to above limitations, ACL anthology remained an archival repository for quite a long time. Bird et al. (2008) developed the *ACL Anthology Reference Corpus (ACL ARC)* as a collaborative attempt to provide a standardized testbed reference corpus based on the ACL Anthology. Later, Radev et al. (2009) have invested humongous manual efforts to construct *The ACL Anthology Network Corpus (AAN)*. AAN consists of a manually curated database of citations, collaborations, and summaries and statistics about the network. They have utilized two OCR processing tools PDFBox² and ParsCit (Councill et al., 2008) for curation. AAN was continuously updated till 2013 (Radev et al., 2013). Recently, this project has been moved to Yale University as part of the new LILY group³.

1.2 The computational linguistic knowledge graph

As a similar initiative, in this paper, we demonstrate the development of *CL Scholar* which automatically mines ACL anthology and constructs computational linguistic knowledge graph (hereafter '*CLKG*'). The current framework automatically crawls new articles, processes, indexes, constructs knowledge graph and generates searchable statistics without involving tedious manual annotations. We leverage state-of-the-art scientific article processing tool OCR++ (Singh et al., 2016) for robust and automatic information extraction from scientific articles. OCR++ is an open-source framework that can extract from scholarly articles the metadata, the structure and the bibliography.

The constructed *CLKG* is modeled as a heterogeneous graph (Sun et al., 2009) consisting of four

²<https://pdfbox.apache.org/>

³<http://tangra.cs.yale.edu/newaan/>

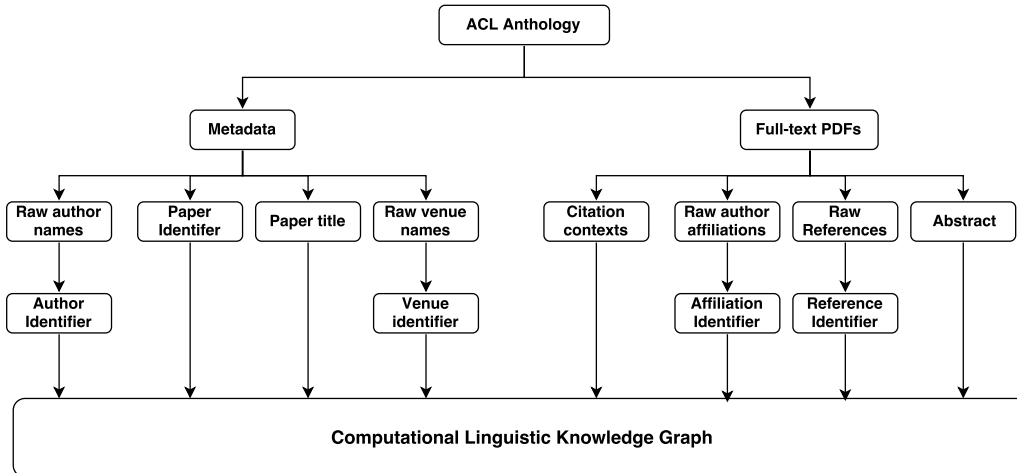


Figure 1: Data processing flow diagram.

entities: author, paper, venue, and field. We utilize metapaths (Sun and Han, 2012) to implement the query retrieval framework.

1.3 Natural language queries

In the first-of-its-kind initiative, we extend the functionalities of *CL Scholar* to answer natural language queries (hereafter ‘*NLQ*’) along with standard keyword-based queries. Currently, it answers *binary*, *statistical* and *list* based *NLQ*. Overall, we handle more than 1200 variations of *NLQ*.

Outline: The rest of the paper is organized as follows. Section 2 describes the ACL Anthology dataset. Section 3 details step by step extraction procedure for *CLKG* construction. In section 4, we describe *CLKG*. We describe our framework in section 5. We conclude in section 6 and identify future work.

Table 2: General statistics about the ACL Anthology dataset.

Number of papers	42,069
Year range	1965–2017
Total authors	37,752
Total unique authors	33,372
Total unified venues	33

2 Dataset

CL Scholar uses metadata and full-text PDF research articles crawled from ACL Anthology. ACL Anthology consists of more than 40,000 research articles published in more than 33 computational linguistic events (venues) including conferences, workshops, and journals. Table 2 presents general statistics of the crawled dataset.

We crawl both metadata information (unique article identifier, article title, authors’ names, and venue) as well as full-text PDF articles. Next, we describe in detail several pre-processing steps and knowledge graph construction methodology.

3 Pre-processing and knowledge graph construction

We process full-text PDFs using state-of-the-art extraction tool *OCR++* (Singh et al., 2016). We extract references, citation contexts, author affiliations and URLs from full-text. *OCR++* also provides reference to citation contexts mapping. Raw information with several variations like author names, venue names and affiliations are assigned unique identifiers using standard indexing approaches. We only consider those reference papers that are present in ACL anthology. This rich textual, as well as citation relationship information, is utilized in the construction of *CLKG*. Figure 1 presents the *CLKG* construction from metadata and full-text PDF files crawled from ACL anthology.

4 Computational linguistic knowledge graph

Computational linguistic knowledge graph (*CLKG*) is a *heterogeneous graph* (Sun et al., 2009) consisting of four entities: author (*A*), paper (*P*), venue (*V*) and field (*F*) as nodes. Each entity is associated with few properties, for example, properties of *P* are publication year, title, abstract, etc. Similarly, properties of *A* are name, publication trend, affiliation etc. We utilize *metapaths* (Sun and Han, 2012) between entities to express semantic relations. For

example, simple metapaths like $A \rightarrow P$ and $V \rightarrow P$ represent “author of” and “published at” relations respectively, whereas complex metapaths like $V \rightarrow A \rightarrow P$ and $F \rightarrow A \rightarrow P$ represent “authors of papers published at” and “authors of papers in” relations respectively. We leverage metapaths to develop *CL Scholar* (described in the next section).

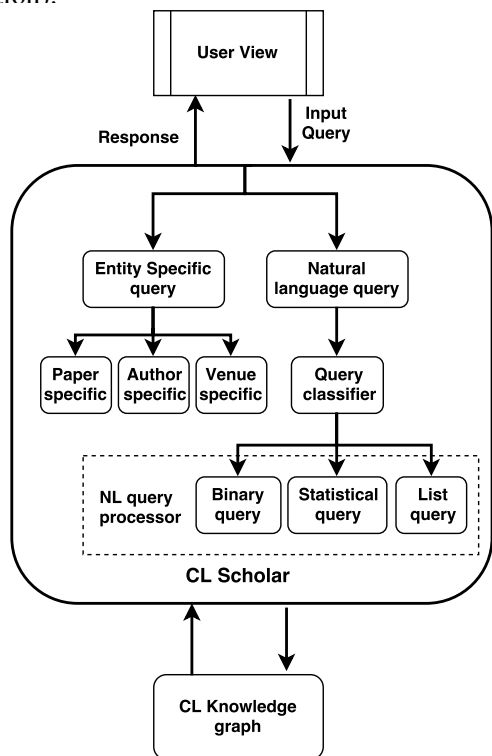


Figure 3: CL Scholar framework.

5 CL Scholar

CL scholar fetches information from *CLKG* as per the input query from the user. The current framework is divided into two modules – 1) natural language based query retrieval, and 2) entity specific query retrieval. Figure 3 shows *CL Scholar* framework.

5.1 Natural language query retrieval

The first module answers natural language queries (*NLQ*). It consists of two sub-modules, 1) the query classifier, and 2) the NL query processor. *Query classifier* classifies user input into one of the three basic types of *NLQ* using regular expression patterns. *NL query processor* processes query based on its type determined by query classifier. Given an input natural language query, we utilize longest subsequence match to identify entity instances. The three types of *NLQ* are:

1. **Binary queries:** These represent a set of queries for which user demands a ‘yes’ or

‘no’ type answer. Table 4 lists few interesting binary queries.

2. **Statistical queries:** These represent set of queries which the knowledge base returns with some statistics. Currently, we support three types of statistics – 1) temporal, 2) cumulative, and 3) comparison. Temporal represents year-wise statistics, cumulative represents overall statistics and comparison represents comparative statistics between two or more instances of the same entity type. Table 4 lists few representative statistical queries.
3. **List queries:** These represent set of queries for which the knowledge base returns a list of papers, authors or venues. Table 4 also enumerates few representative list queries.

5.2 Entity specific query retrieval

CL scholar also supports entity specific retrieval. As described in section 4, *CLKG* consists of four entities: paper, author, venue, and field. Currently, our system supports three⁴ entity specific retrieval schemes handled by three sub-modules:

1. **Paper specific:** This sub-module returns paper specific information. Currently, we retrieve and display author names and affiliations, abstract, publication year and venue, cumulative and year-wise citations, list of references, citer papers, co-cited papers present in ACL anthology and list of URLs present in the paper text. We also show average sentiment score received by the queried paper by utilizing incoming citation contexts. Table 5 shows three representative paper specific queries.
2. **Author specific:** This sub-module handles author specific queries. Given an author name, the system shows its cumulative and year-wise publication and citation count, collaborator list with an average number of collaborations, current and temporal H-index and temporal topic distribution. We also list author’s publications in ACL anthology. Table 5 lists three author specific queries with first name, last name and full name respectively.
3. **Venue specific:** We also answer venue specific queries. For each venue specific query, the system shows cumulative and year-wise

⁴The fourth sub-module is still under development.

Table 4: Representative queries from three natural language query classes. A represents author, P represents paper, V represents venue and F represents field. The list of supported queries is available online at *CL Scholar* portal.

Binary queries	Statistical queries	List queries
1. Is V accepting papers from F	How many F papers are published in V over the years	List the papers from F accepted in V
2. Has A published any paper	How many papers are published by A	List the papers published by A
3. Does A publish papers on F	How many papers are published by A in F	List the papers published by A on F
4. Are there any papers published by A_1 and A_2 together	How many papers are published by A_1 and A_2 together	List the papers published by A_1 and A_2 together
5. Are there any papers of A with positive sentiment	How many papers are there of A with positive sentiment	List of papers with positive sentiment of A

publication and citation count, 2-year impact factor, recently held year and list of collaborating venues. Table 5 shows three representative venue specific queries.

Table 5: Representative entity specific queries.

Paper specific	Author specific	Venue specific
OCR	Chris	NAACL
Deep learning	Singh	SIGDAT
Word embeddings	Aravind Joshi	ACL

5.3 Additional insights

We provide two additional insights by analyzing incoming citation contexts. First, we present a summary generated from incoming the citation contexts (Qazvinian and Radev, 2008). Currently, we show five summary sentences against each paper. Second, we also compute sentiment score of each citation context by leveraging a standard sentiment analyzer (Athar and Teufel, 2012). We aggregate by averaging over the sentiment score of all the incoming citation contexts.

5.4 Ranking

Currently, we employ popularity based ranking of retrieved results. We utilize current citation count as a measure of popularity. In future, we plan to deploy other ranking schemes like recency, impact, sentiment, relevance, etc.

5.5 Deployment

CL Scholar is developed using ReactJS framework. The system also supports REST API requests which are powered by a NodeJS server with data being served using MongoDB. It is currently

accessible at our research group page⁵. More information about API usage is available at API support page⁶. In addition, the entire knowledge graph can also be easily downloaded in a plain text format. Figure 6 shows a snapshot of the *CL Scholar* landing page.

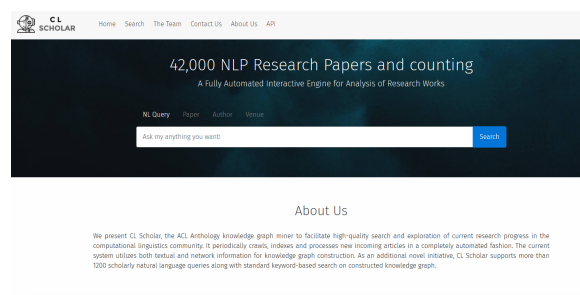


Figure 6: Snapshot of *CL Scholar* landing page.

The current system is still under development. Currently, we assume that spellings are correct for NLQ. We do not support instant query search. We also do not support query recommendations.

6 Conclusion

In this paper, we propose a fully automatic approach for the development of computational linguistic knowledge graph from full-text PDF articles available in ACL Anthology. We also develop first-of-its-kind academic natural language query retrieval system. Currently, our system can answer three different types of natural language queries. In future, we plan to extend the query set. We also plan to append structural information within knowledge graphs such as section labeling of citations, figure and table captions etc. We also plan to conduct extensive evaluation to compare *CL Scholar* with state-of-the-art systems.

⁵<http://cnerg.iitkgp.ac.in/aclakg>

⁶<http://cnerg.iitkgp.ac.in/aclakg/api>

References

- Awais Athar and Simone Teufel. 2012. **Context-enhanced citation sentiment detection**. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12*, pages 597–601, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Steven Bird, Robert Dale, Bonnie J Dorr, Bryan Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, and Yee Fan Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics.
- Isaac G Council, C Lee Giles, and Min-Yen Kan. 2008. Parscit: an open-source crf reference string parsing package. In *LREC*, volume 8, pages 661–667.
- Vahed Qazvinian and Dragomir R. Radev. 2008. **Scientific paper summarization using citation summary networks**. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1, COLING '08*, pages 689–696, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Dragomir R Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The acl anthology network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 54–61. Association for Computational Linguistics.
- Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. **The acl anthology network corpus**. *Language Resources and Evaluation*, pages 1–26.
- Mayank Singh, Barnopriyo Barua, Priyank Palod, Manvi Garg, Sidhartha Satapathy, Samuel Bushi, Kumar Ayush, Krishna Sai Rohith, Tulasi Gamidi, Pawan Goyal, and Animesh Mukherjee. 2016. **Ocr++: A robust framework for information extraction from scholarly articles**. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3390–3400, Osaka, Japan. The COLING 2016 Organizing Committee.
- Yizhou Sun and Jiawei Han. 2012. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2):1–159.
- Yizhou Sun, Yintao Yu, and Jiawei Han. 2009. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 797–806. ACM.