Analysis and Evaluation of Open Source Scientific Entity Datasets

Qinya Xu[1,2], **Qianqian Yu**[1,2,3†], **Li Qian**[1,2,3†]

[1]National Science Library, Chinese Academy of Sciences, Beijing 100190, China
[2]Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China
[3]Key Laboratory of New Publishing and Knowledge Services for Scholarly Journals, Beijing 100190, China

Keywords: Scientific literature; Dataset; Evaluation; Research entity; Open source

Citation: Xu Q.Y., Yu Q.Q., Qian L.: Analysis and Evaluation of Open Source Scientific Entity Datasets. Data Intelligence, Vol. 7, Art. No.: 2025r24, pp. 667–691, 2025. DOI: https://doi.org/10.3724/2096-7004.di.2024.0009

ABSTRACT

Focusing on the construction and application of research entity datasets based on scientific literature, it emphasizes the crucial role of data quality in enhancing natural language processing applications in scientific literature mining, addressing the gaps in detailed descriptions and evaluations of datasets in existing research, and promoting advancements in scientific literature entity recognition technology. Through extensive online research and literature review, 22 open-source research entity datasets were analyzed, focusing on the dataset lifecycle's collection, annotation, release, and application stages. A set of quality assessment methods for research entity datasets is proposed, along with a discussion of the challenges in dataset construction and strategies for efficient use. The study emphasizes the importance of high-quality datasets for natural language processing applications, proposing evaluation methods and sharing strategies intended to serve as a reference for constructing, selecting, and using scientific literature entity datasets.

1. INTRODUCTION

Scientific and technological literature, as an essential carrier for disseminating scientific knowledge, plays a crucial role in promoting knowledge acquisition and research progress. With the rapid development of science and technology, the quantity of scientific and technological literature has increased dramatically, and the complexity of content has also significantly increased. Traditional literature retrieval and analysis techniques gradually show limitations. At the same time, breakthroughs in artificial intelligence and

[†] Corresponding author: Qianqian Yu (Email: yuqianqian@mail.las.ac.cn); Li Qian (Email: qianl@mail.las.ac.cn).

natural language processing technologies provide new methods for in-depth analysis of scientific and technological literature content. Researchers use these cutting-edge technologies to mine literature content deeply, achieving knowledge organization and management from macro to micro levels [1]. However, the performance of these automated technologies mostly depends on the quality of the applied datasets [2]. High-quality scientific entity datasets provide the foundation for entity recognition and classification and are also critical for evaluating the performance of new methods and algorithms. Existing research mainly focuses on improving technical methods or algorithm models, with insufficient emphasis on detailed description and analysis of datasets.

Through extensive online surveys and literature reviews, this study conducts an in-depth analysis of 22 open-source scientific research entity datasets based on scientific literature, exploring the characteristics of these datasets and their applications in scientific literature mining. This study focuses on the key stages of the dataset lifecycle: collection, annotation, publication, and application, proposing a set of methods for evaluating the quality of scientific research entity datasets. Additionally, this paper explores the challenges of constructing full-text datasets and efficiently utilizing datasets and proposes strategies to enhance the efficiency and sustainability of dataset sharing. This study aims to conduct an in-depth analysis and comprehensive evaluation of scientific research entity datasets in scientific literature, aiming to promote the application and development of natural language processing technology in scientific literature content mining, address the shortcomings in existing research on dataset analysis, provide insights into the development of scientific literature entity recognition technology from a data perspective, and offer guidance and reference for researchers in related fields.

2. RELATED WORK

2.1 Research Status of Scientific Entity Dataset Construction

Scientific research entity datasets have become essential resources in natural language processing, receiving extensive attention and application from the academic community. For instance, datasets like ScienceIE, FTD, and SCIERC significantly advance core NLP tasks such as automated text classification, precise identification of key entities, and exploration of complex relationships between entities. However, the construction of scientific literature datasets faces various challenges, such as the lagging update speed of datasets compared to the rapid evolution of machine learning models [3], which could limit the potential of models in practical applications. Taking ACL RD-TEC [4] as an example, as a golden standard dataset in the field of computational linguistics, although it covers 10922 ACL conference papers published between 1965 and 2006, its depth of content and annotation is limited. In comparison, ACL RD-TEC 2.0 [5] has done more detailed work on the sub-classification of terms, such as methods, tools, and language resources. The lack of standards in dataset construction is also an important issue. When annotating scientific literature texts at a fine granularity, it is crucial to precisely define entity categories and boundaries, ensure the generality of annotations, and clarify the annotation process. Research indicates that documents such as lists, guidelines, and data cards released alongside datasets enhance the transparency and reusability of datasets [6]. These documents provide an overview of the structure and content of datasets, including information on the motivation, purpose, collection process, and expected uses of dataset construction, helping researchers to utilize these datasets for scientific research and model development more effectively.

The academic community is actively exploring solutions to these issues and challenges. By developing automated tools and adopting crowdsourcing methods to accelerate dataset maintenance and updates [7], datasets' maintenance efficiency and quality are improved. Researchers explore methods for automatically detecting and reweighting incorrect labels to reduce the problem of dataset annotation noise in named entity recognition tasks [8]. More and more machine learning and natural language processing conferences are adopting reproducible checklists trained by machine learning models to enhance the transparency and quality of research. Constructing multimodal datasets such as text and images provides a richer perspective for training deep learning models [9]. With the release of large language models like ChatGPT, researchers are attempting to leverage the comprehension capabilities of these models to determine whether they can perform human-like labeling tasks. Zhu et al. found that [10] ChatGPT has the potential to handle these data annotation tasks with an average accuracy of 0.609, but there are significant differences in performance between different labels. Fabrizio Gilardi et al. found that [11] ChatGPT outperforms crowdsourcing in tasks such as relevance, stance, topic, and frame detection annotation. Although ChatGPT shows excellent potential in various text annotation and classification tasks, Reiss et al. argue that [12] ChatGPT labeling is non-deterministic, meaning even minor changes in prompts or repeated inputs may lead to different outputs, requiring manual verification.

Previous research has made significant progress in the construction of scientific literature entity datasets, particularly in supporting core tasks in natural language processing. However, issues such as outdated datasets and lack of standardized practices continue to limit the effectiveness of models in practical applications. While existing studies have primarily focused on optimizing technical methods and algorithmic models, there remains a clear gap in the systematic exploration and establishment of fine-grained annotation standards during the dataset construction process. Therefore, this paper conducts an in-depth analysis of 22 open-source datasets, aiming to explore the importance of standardization in the construction of scientific literature entity datasets. The study seeks to further clarify the critical role that standardization plays in the extraction of scientific literature entities, thereby providing more effective support for scientific research and knowledge discovery.

2.2 Research Status of Dataset Quality Evaluation

In natural language processing and scientific literature analysis, dataset quality is critical to ensuring model accuracy. Recent studies have revealed various quality issues in datasets, such as label errors and entity omissions [13]. These issues may lead to overestimating model performance in tasks such as semantic understanding and natural language inference, affecting their generalization ability [14]. To address dataset quality issues, Picard et al. [15] proposed seven quality dimensions, including accuracy, accessibility, consistency, relevance, timeliness, traceability, and usability. The dataset quality dimensions proposed by Chug et al. [16] include source, dataset characteristics, consistency, metadata coupling, statistics, and relevance. Dong et al. [17] constructed a dataset quality assessment framework covering three dimensions: credibility, difficulty, and validity, and developed corresponding evaluation metrics.

In the data-driven research paradigm, the importance of datasets is continuously increasing, emphasizing the improvement of dataset quality through automatic or semi-automatic means to enhance model performance. Andrew et al. [18] first proposed a research paradigm focused on optimizing data rather than models, aiming to study data-centric approaches to enhance the performance of machine learning models. Chen et al. [19] proposed a framework for evaluating complex data quality, focusing on data such as knowledge graphs. DataCLUE proposed by Xu et al. [20], as the first data-centric benchmark test in the field of NLP, emphasizes the importance of improving dataset quality.

In current research, dataset quality is a critical factor influencing the performance of natural language processing models. However, most studies to date have primarily focused on issues related to datasets during the model training phase, often neglecting a systematic evaluation of quality indicators across the entire dataset lifecycle. While previous research has identified issues such as annotation errors and entity omissions, there has been a lack of comprehensive analysis covering the entire process from data collection to application. This paper develops and assesses a quality evaluation framework for scientific literature entity datasets, approached from the perspective of the dataset lifecycle. By addressing gaps in existing research, this framework offers new methodologies and practical guidance for enhancing dataset quality.

3. RESEARCH METHODOLOGY AND DATASET SELECTION

This study aims to delve into and comprehend open-source scientific research entity datasets based on scientific literature, employing systematic literature review and data collection methods. In collecting data, major academic databases such as Web of Science, Google Scholar, CNKI, etc., were searched to ensure the comprehensiveness and accuracy of the collected information. The search strategy involved using precise keyword combinations and Boolean logical operators, such as employing keywords like “scientific literature”, “scientific papers”, “academic papers”, “entity recognition”, “datasets” and “information extraction”. In addition to database searches, reference lists of relevant datasets and academic conference materials were consulted to ensure the relevance and timeliness of the datasets.

Clear criteria were established during the dataset selection process to ensure the quality and relevance of the datasets. The selected datasets must be publicly available and open-source, covering entity types such as research questions, method models, theoretical principles, software systems, measurement metrics, instruments, data, and concept definitions. The detailed information of the final 22 datasets is listed in Table 1 within the document. The diversity and richness of these datasets provide a solid foundation for this study, enabling comprehensive analysis and evaluation of the application and impact of open-source datasets in scientific literature analysis.

Table 1. 22 Datasets Selected for this Study.

Dataset	Domain	Time	Number	Annotation Level	Entity	Language	Access
FTD [21]	CL	2011	474 abstracts	abstracts	Domain, Focus, Technique	English	https://nlp.stanford.edu/pubs/FTDDataset_v1.txt
ACL RD-TEC2.0	CL	2016	300 abstracts	abstracts	Language Resource, Language Resource Product, Measures and Measurements, Models, Other, Technology and Method, Tool and Library	English	http://pars.ie/lr/acl_rd-tec
SciERC [22]	CL	2018	500 abstracts	abstracts	Generic, Task, Material, Other ScientificTerm, Method, Metric	English	http://nlp.cs.washington.edu/sciIE
OA-STM [23]	10 STEM	2020	110 abstracts	abstracts	Process, Method, Material, Data	English	https://github.com/elsevierlabs/OA-STM-Corpus
ScienceIE [24]	CS, MS, Phy	2017	500 paragraphs	sentences	Material, Process, Task	English	https://alt.qcri.org/semeval2017/task10/
SCIREX [25]	ML	2020	438 papers	full text	Dataset, Method, Metric, Task	English	https://github.com/allenai/SciREX
TDM Tagged Corpus [26]	NLP	2021	1500 sentences	sentences	Task, Metric, Dataset	English	https://github.com/Ishani-Mondal/SciKG
NLP-TDMS [27]	CL	2019	153 papers	full text	Dataset, Metric, Score, Task	English	https://github.com/IBM/science-result-extractor
CL-Title [28]	CL	2021	50237 titles	titles	Research problem, Solution, Resource, Language, Tool, Method	English	https://github.com/jd-coderepos/cl-titles-parser

Table 1. *Continued.*

Dataset	Domain	Time	Number	Annotation Level	Entity	Language	Access
ACL [29]	CL	2022	31044 titles	titles	Language, Method, Research problem, Resource, Dataset, Solution, Tool	English	https://github.com/jd-coderepos/contributions-ner-cs/tree/main/acl
PwC [29]	AI	2022	12271 titles、abstracts	titles、abstracts	Research problem, Method	English	https://github.com/jd-coderepos/contributions-ner-cs/tree/main/pwc
STEM-ECR [30]	10 STEM	2020	110 abstracts	abstracts	Data, Material, Method, Process	English	https://github.com/elsevierlabs/OA-STM-Corpus
STEM-60k [31]	10 STEM	2022	59984 abstracts	abstracts	Data, Material, Method, Process	English	https://github.com/jd-coderepos/stem-ner-60k/
ORKG-TDM [32]	AI	2021	5361 papers	full text	Task, Dataset, Metric	English	https://github.com/Kabongosalomon/task-dataset-metric-nli-extraction/
scholarly-entity-usage-detection [33]	CL	2021	1000 sentences	sentences	Method, Dataset	English	https://github.com/michaelfaerber/scholarly-entity-usage-detection
SoMeSci [34]	Bio	2021	1367 papers	full text	Software	English	https://data.gesis.org/somesci/ https://zenodo.org/record/4701764
CZ Software Mentions [35]	Bio	2022	2433010 papers	full text	Software	English	https://github.com/chanzuckerberg/software-mention-extraction
ner_dataset_recognition [36]	CS	2021	6000 sentences	sentences	Dataset	English	https://github.com/xjaeh/ner_dataset_recognition

Table 1. *Continued.*

Dataset	Domain	Time	Number	Annotation Level	Entity	Language	Access
DMDD [37]	CL	2023	31219 papers	full text	Dataset	English	https://www.kaggle.com/datasets/panhuitong/dmdd-corpus
TDMSci [38]	CL	2021	2000 sentences	sentences	Task, Dataset, Metric	English	https://github.com/IBM/science-result-extractor
bioNerDS [39]	Bio	2013	60 papers	full text	Database, Ontology, Classification, Software, Programs, Tools, Network services	English	https://bionerds.sourceforge.net/
Softcite [40]	Bio	2021	4971 papers	sentences	Software name, Version, publisher, URL	English	https://github.com/softcite/software-mentions

Domain Acronyms: CL - Computational Linguistics; CS - Computer Science; MS - Material Science; Phy - Physics; AI - Artificial Intelligence; 10 STEM - Agriculture, Astronomy, Biology, Chemistry, Computer Science, Earth Science, Engineering, Material Science, Mathematics, and Medicine.; ML - Machine Learning; Bio - Biomedical sciences; NLP - Natural language processing; CV - Computer Vision.

4. DETAILED ANALYSIS OF DATASETS

Knowledge entity recognition plays a central role in scientific literature content mining. With the development of fields such as knowledge graphs and entity metrics, their challenges and importance are increasing. For instance, the ScienceIE semantic evaluation competition 2017 required the extraction of entities such as tasks, processes, and materials from scientific literature, significantly advancing research in knowledge entity recognition. With technological advancements and methodological innovations, the application scope of knowledge entity recognition has expanded from mere titles to abstracts and even full texts. This reflects the scientific research field's pursuit of more comprehensive and in-depth information, demonstrating the powerful capabilities of research methods in handling complex texts. Early research, limited by technology and corpora, primarily focused on structurally clear title texts. However, current research can identify and extract richer vocabulary and entity types at the abstract and full-text levels.

4.1 Title Datasets

The titles of scientific literature are the most information-dense part, typically containing the literature's topic, research methods, and key concepts. Therefore, datasets focusing on titles are significant for rapidly identifying the literature's research areas and core topics. In natural language processing and information retrieval, such datasets are widely used to train algorithms to enhance the automatic identification of literature topics and research trends.

The CL-Titles dataset, released by TIB Leibniz Information Centre for Science and Technology in 2021, aims to automatically extract key scientific entities from academic article titles. After analyzing the titles of 50237 articles in the ACL literature database, it successfully extracted 19799 research questions, 18111 solutions, 20033 resources, 1059 languages, 6878 tools, and 21687 methods. CL-Titles uses a rule-based automatic extraction method, which analyzes lexical-syntactic patterns in article titles to identify scientific entity types, providing valuable resources for constructing scientific knowledge graphs and research in natural language processing.

The ACL dataset focuses on the semantic entity analysis of academic article titles in computational linguistics, further developed based on the CL-Titles parser. This dataset covers all article titles in the ACL literature database, providing a more comprehensive analysis of scientific entities. Compared to the CL-Titles's automated annotation method, the ACL dataset has been expanded through a more detailed manual review to include key scientific entity types such as "datasets." The ACL dataset contains seven core entity types: language, method, research question, resource, dataset, solution, and tool, providing a more comprehensive capture of the core elements of computational linguistics research. After manually correcting entity annotation from 31041 titles, a blind test showed an annotator consistency of 71.52%, demonstrating the reliability of dataset annotations.

4.2 Sentence Datasets

Sentence datasets are dedicated to in-depth analysis of individual sentences or groups of sentences in literature, aiming to reveal their underlying fine-grained knowledge structure. These datasets analyze sentences' linguistic structure and semantic relationships to gain insight into entity relationships, concept definitions, and argumentation methods. In text mining and knowledge extraction, sentence datasets play a key role in analyzing the content structure of scientific literature from a micro perspective.

The TDMSci dataset is a high-precision corpus specifically designed for annotating tasks, datasets, and metrics in NLP papers. These concepts are central to the analysis and understanding of experimental science papers. These concepts are central to the analysis and understanding of experimental science papers. To ensure comprehensive coverage, the sentences in the TDMSci corpus are primarily selected from entire papers, not limited to abstracts. The corpus contains 2000 sentences involving 2937 mentions of tasks, datasets, and metrics. The TDMSci corpus provides a vital annotated resource for the NLP field, aiding in the construction and application of automated NLP TDM taxonomies.

ScienceIE is a release task of SemEval 2017 task10, aiming to automatically extract key phrases from scientific publications. The corpus for this task is directly obtained from ScienceDirect, and 500 paragraphs from journal articles in computer science, materials science, and physics are selected for annotation, including three types of key entities: processes, tasks, and materials. Scientific research and practice are based on acquiring, maintaining, and understanding existing scientific work in specific domains related to these basic objects. Entity recognition in scientific research can provide a more comprehensive perspective on understanding scientific papers' experimental design and research results.

4.3 Abstract Datasets

Abstract datasets focus on the abstract section of scientific literature, which typically condenses the core content of the document. Abstracts not only summarize the document's theme documents theme but also include fundamental research background, methods, results, and conclusions. Abstract datasets are essential in rapidly extracting and evaluating core information from the literature. Additionally, these datasets are valuable in developing automatic summarization and key information extraction algorithms.

The FTD dataset aims to analyze the titles and abstracts of computational linguistics articles to characterize the focus, application areas, and techniques used in the research. It focuses on extracting the main contributions, methods or tools used, and application areas of articles through semantic pattern matching of abstract dependency trees. The dataset covers the titles and abstracts of 15016 computational linguistics articles from 1965 to 2009, and its use can validate the effectiveness of information extraction systems and demonstrate the development dynamics and overall impact of the computational linguistics subfield.

The SciERC dataset focuses on entity recognition, relation classification, and coreference cluster processing in scientific article abstracts. SciERC covers scientific abstracts from 500 conferences in artificial intelligence, annotated with scientific entities, their relationships, and coreference clusters. The dataset expands the entity and relationship types of traditional scientific information extraction datasets, adds annotations for cross-sentence relationships, and provides a more comprehensive perspective on understanding scientific papers' experimental design and research results.

4.4 Full-text Datasets

Full-text datasets provide a solid foundation for in-depth analysis of scientific literature, covering the entire content from introduction to conclusion. Such datasets help reveal detailed information and complex arguments in the literature. They are of great value in supporting deep learning algorithms to process long texts, build knowledge graphs, and optimize recommendation systems and information retrieval tools.

The ORKG-TDM dataset aims to automatically construct a ranking list of scientific papers. The dataset extracts information such as tasks, datasets, and metrics from the full text to build a document-level scientific knowledge graph. ORKG-TDM combines remote supervision and PDF text preprocessing techniques to effectively reduce the need for manual labeling. The dataset contains over 4500 academic articles and their related triple information, providing rich empirical data for text mining and knowledge graph construction.

The SCIREX dataset, developed in 2020, focuses on document-level information extraction from scientific articles, covering tasks such as significant entity recognition and multi-relational extraction. SCIREX adopts a combined approach of automation and manual annotation, reducing the labeling workload while improving data quality. The dataset demonstrates innovative value in understanding the full content of scientific articles and constructing knowledge graphs, significantly driving research progress in natural language processing.

5. DATASET QUALITY AND EVALUATION

This study conducts a demand analysis of knowledge entity units based on scientific and technological literature datasets. We propose a quality evaluation framework for scientific and technological literature datasets focusing on the dataset lifecycle (Table 2). This framework covers four key stages of dataset collection, annotation, publication, and application. Evaluation metrics are designed for each stage to monitor and enhance the dataset's quality throughout. In this study, the maximization of the common set of entity categories in the dimensions of methods, tools, and tasks, as well as the careful consideration of annotation levels, are used as filtering criteria. Six datasets (ACL, ScienceIE, FTD, SciERC, ORKG-TDM, and SCIREX) were selected for in-depth analysis. To address the unique characteristics of scientific and technological literature entity datasets, this study integrates these factors into the proposed evaluation framework. This approach ensures that the evaluation and assessment processes are specifically tailored to align with the distinct features and application contexts of these datasets. This study aims to provide a new understanding framework for the construction and quality evaluation of scientific and technological literature datasets.

5.1 Data Collection

In constructing datasets for scientific and technological literature, controlling the quality of raw data is a key factor in ensuring the effectiveness and reliability of the dataset. This study delves into the quality control mechanisms during the data collection process, especially evaluating from the two dimensions of the authority of data sources and the adequacy of the data scale. First, the authority of data sources plays a decisive role in ensuring the basic quality of data. When selecting data sources, priority should be given to highly recognized and authoritative databases. For example, data from core journals are generally more credible and valuable than those from general journals, and formally published academic papers are usually more trustworthy and valuable than literature on preprint platforms (such as arXiv). This selection mechanism ensures that the dataset starts from a relatively high-quality point at the beginning of construction. Secondly, the scale of data is also an essential criterion for measuring dataset quality. High-quality datasets should focus on the accuracy and relevance of data and ensure an adequate volume of data to support complex data analysis and model training requirements. Table 3 reveals the common emphasis on the authority of sources and the volume of data during the dataset collection stage while also demonstrating differences in data depth and availability. In particular, the

ACL dataset highlights its large number of titles. At the same time, the full-text data of SCIREX provides depth information, indicating that the evaluation of data volume in the construction of scientific and technological literature entity datasets needs to consider the depth of text content and its support for complex analysis and model training. In addition, copyright issues for full text are factors that need to be carefully considered when selecting data sources, affecting the construction and usage scope of datasets.

Table 2. Evaluation Indicators for Dataset Quality.

Lifecycle Stage	Metric	Significance
Data Collection	Authority of Source	Assessment of the credibility and authority of data sources
	Volume of Data	Indicators of the coverage and depth of datasets
Data Annotation	Annotation Normativity	Standardization and thoroughness of the annotation process
	Professional Level	Professional background and skill level of the annotation team
	Consistency of Results	Consistency rate of results among different annotation sources
Data Publication	Metadata Integrity	Comprehensiveness of dataset description information
	Granularity of Labels	Completeness and balance of label granularity
	Novelty	Recency and timeliness of dataset publication time
	Accessibility	Convenience and universality of dataset access
	Interoperability	Compatibility of dataset formats, annotation schemes, and data exchange capability
Data Application	Impact	Academic and practical application impact of datasets

Table 3. Dataset Collection Stage.

Dataset	Authority of the source	Data volume
ACL	ACL paper collection	31044 titles
ScienceIE	ScienceDirect publications	500 paragraphs
FTD	ACL paper collection	474 abstracts
SCIERC	AI conference, workshop paper collection	500 paragraphs
ORKG-TDM	papers with code	4500 papers
SCIREX	-	438 papers

5.2. Data Annotation

Data annotation is a key process in machine learning tasks involving transforming raw data from scientific and technological literature texts into machine-readable information (Table 4). The quality of annotation is an essential factor affecting the performance of machine learning models. When annotating scientific and technological literature texts at a fine granularity, the key is to precisely define entity categories and boundaries, ensure the generality of annotations, and clarify the annotation process. Therefore, the data annotation stage evaluation focuses on the existence and thoroughness of annotation standards, the professional level of annotators, and the consistency of annotation results. These indicators collectively ensure the accuracy and reliability of the data annotation process. High annotation consistency means that the reliability and consistency of data annotation are high, which is particularly important for training machine learning models.

Table 4. Dataset Annotation Stage.

Dataset	Annotation standardization	Professional-level of annotators	Consistency of results
ACL	Dataset construction process	Annotation by rule-based parsers, validation by postdoctoral researchers	0.7152
ScienceIE	Introduction to the corpus and annotation process	Undergraduates	0.45-0.85
FTD	Dataset construction process	Matching phrase trees, manual annotation of test sets	-
SCIERC	The dataset construction process, the appendix of annotation guidelines	Experts in the field	0.769
ORKG-TDM	Dataset construction process	Remote supervision	-
SCIREX	The dataset construction process, annotation statistics, the appendix of standard guidelines	Remote supervision and validation by doctoral students	0.95

Taking the SCIREX dataset as an example, implementing remote supervision labeling and manual quality control effectively improves the consistency of annotations. The SciERC dataset proves its annotation consistency through meticulous annotations by domain experts and a high Kappa coefficient (76.9%). These examples indicate that carefully designed annotation processes and strict quality control can significantly improve the consistency of dataset annotations. Therefore, in the future construction of datasets, attention should be paid to the transparency and rigor of the annotation process, strengthening cross-annotation and expert review to ensure the broad applicability and accuracy of datasets across various research backgrounds. Additionally, using statistical tools such as the Kappa coefficient to quantitatively assess annotation consistency is an important way to improve dataset quality. In this way, it promotes the in-depth research of academia and the development of technological innovation, providing more solid and reliable data support for accumulating and applying scientific knowledge.

5.3. Data Publication

In the data publishing phase, the evaluation dimensions cover the completeness of metadata, the granularity balance of labels, the novelty of the dataset, and the accessibility of the publishing channels.

5.3.1 Metadata Completeness

Metadata, as data about data, provides a structured description of datasets. It helps users discover, identify, and access data and serves management, validation, and data quality assurance functions. It is crucial for efficiently utilizing datasets and providing high-quality information services. Dataset metadata fields include dataset name, field, description, source, proposal time, data size, annotation level, annotation method, entity object, sample distribution, language, and access link. The number of non-empty fields in metadata measures metadata completeness. The evaluation method is as follows:

$$X = \frac{A \cap B}{B} \tag{1}$$

A represents the number of occurrences of defined dataset metadata fields, and B represents the total number of metadata fields. $X \in [0, 1]$, the higher the value of X, the higher the metadata completeness of the dataset.

As shown in Figure 1, the dataset analyzed demonstrates excellent metadata completeness, providing crucial information to support academic research and practical applications. Detailed descriptions, creators, and source information enhance the transparency and credibility of datasets. In contrast, detailed explanations of release time, annotation methods, entity objects, and sample distributions improve the applicability of datasets. Comprehensive metadata evaluation results indicate that these datasets have a high level of completeness, which is crucial for ensuring dataset quality and advancing academic research and technological applications.

5.3.2 Fine-grained label balance

In developing and evaluating scientific literature datasets, the balance of fine-grained labels is a key consideration. Analyzing the distribution of different category labels in the dataset, the balance of categories is measured, affecting the effectiveness of subsequent model training and bias issues. The proportion of fine-grained labels in the dataset in this study is shown in Figure 2, which exhibits phenomena such as long-tail distribution and sparsity of labels. The balance of fine-grained labels is calculated using normalized entropy with the formula:

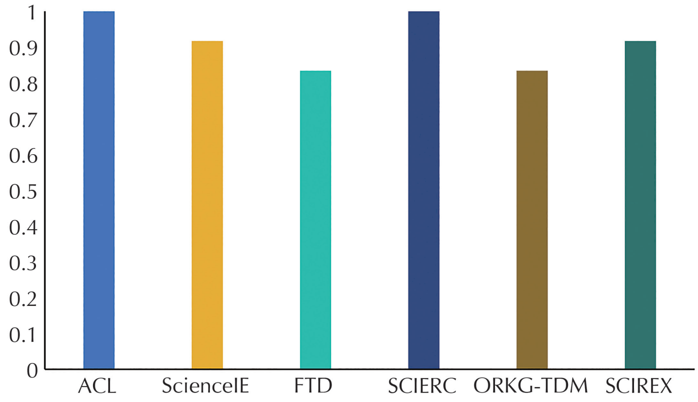
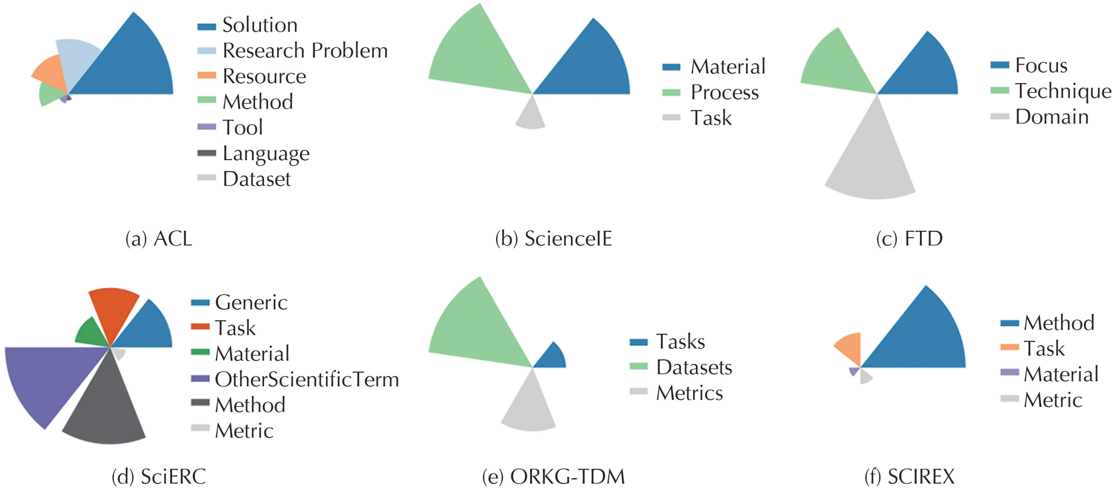


Figure 1. Metadata Completeness for Datasets.



(a) ACL (b) ScienceIE (c) FTD

(d) SciERC (e) ORKG-TDM (f) SCIREX

Figure 2. Distribution of Fine-Grained Label Proportions for Datasets.

$$\mathrm{E} = -\frac{\sum_{i=1}^{k} p_i \log_2(p_i)}{\log_2(k)} \tag{2}$$

Where k is the total number of labels, p_i is the relative frequency of the ith class label in the dataset, that is, the probability of occurrence of that class label. $\mathrm{E} \in [0, 1]$, the larger the E value, the more balanced the dataset is regarding category label distribution.

As shown in Figure 3, the normalized entropy of the SciERC and ScienceIE datasets are 0.9225 and 0.9177, respectively, indicating good category balance. This helps reduce category bias during training and improves the model's generalization ability and accuracy. Even with a high normalized entropy value, the sample size of specific categories may still be insufficient, which requires special attention during dataset design and collection. For example, the normalized entropy value of the SCIREX dataset is 0.7416, indicating good category balance, but the actual distribution of data labels still needs further analysis.

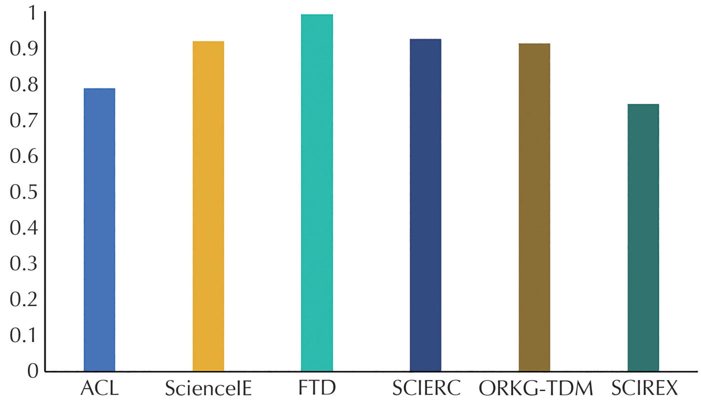


Figure 3. Balance of Fine-Grained Labels for Datasets.

Through the calculation and analysis of normalized entropy, measures are recommended to ensure the balanced distribution of fine-grained labels when constructing scientific literature datasets, such as improving the imbalance of label distribution through data augmentation or resampling techniques. At the same time, attention should be paid to the sample size of each category in the dataset to ensure the representativeness and practicality of each category. When using datasets to train machine learning models, balanced label distribution is essential for avoiding overfitting and improving model generalization ability.

5.3.3 Dataset Novelty

In this study, the novelty of the dataset is evaluated based on its release time. The six datasets selected for evaluation were all released after 2017, indicating that these datasets have high timeliness and novelty. The novelty of the dataset reflects the advantages of the latest research results and dynamics in science and technology, ensuring the dataset's cutting-edge nature and practical value.

5.3.4 Dataset Accessibility

This study selected six open-source datasets, ensuring that each dataset provides download links and is publicly accessible. This selection process was designed to ensure research transparency and reproducibility, enabling a broader community of researchers to utilize these datasets for exploration and application development. However, the accessibility of these datasets is influenced not only by their open-source nature but also by factors such as storage location, data format, and ease of access. These datasets are hosted on different platforms, which may affect data retrieval speed and reliability, especially for large datasets or in regions with slower internet connections. For instance, the FTD dataset is stored on Stanford University's servers, while the SciERC and ScienceIE datasets are hosted by academic institutions. In contrast, datasets such as SCIREX, ORKG-TDM, and ACL are stored on GitHub, which offers version control and more convenient access to updates. By accounting for these factors, researchers can better anticipate and address challenges related to data storage and access, thereby enhancing the usability of these datasets across various research environments.

5.3.5 Data Interoperability

In scientific literature entity datasets, interoperability specifically refers to the compatibility of dataset formats, annotation schemes, and data exchange capabilities. In interoperability assessments, the SciERC and SCIREX datasets, which utilize standardized JSON formats, demonstrate high compatibility and data exchange capabilities, making them suitable for broad application across various natural language processing tools. The ScienceIE dataset is meticulously annotated, but its use of multiple formats increases processing complexity. The FTD and ACL datasets use plain text formats, with ACL employing the standard BIOES annotation scheme, offering good compatibility. The ORKG-TDM dataset is stored in TSV format, which is simple and user-friendly, but may be limited when handling complex data. To enhance interoperability, it is recommended to standardize formats. Standardized data formats and unified annotation schemes are crucial for improving the interoperability of scientific literature entity datasets, facilitating data sharing and the application of various tools, and driving the broader adoption and development of NLP technologies.

5.4 Data Application

In the data application stage, the actual impact of the dataset is mainly examined. This includes its performance and contributions to scientific research or practical applications. The citation count of the dataset partly reflects its popularity in the academic community. However, its more profound impact comes from its contributions to the research field and its driving force in practical applications. The high citation count of SciERC is not only a recognition of its data quality and utility. It also reflects its central position in academic research. The underlying reasons should be investigated thoroughly for datasets with lower citation counts like ACL. These reasons could be the dataset's specialization, narrow application scope, short release time, or inadequate promotion. These factors are not directly equivalent to inadequate dataset quality, but they do affect its impact on the academic community. The influence of a dataset is closely related to its maintenance and update strategies. Regularly updated datasets can better adapt to changes in the research field. Thus maintaining their influence and application value in the long term. With the rapid development of artificial intelligence technology, datasets are updated promptly. Reflecting the latest research trends and technological advancements are more likely to be widely cited and applied.

5.5 Comprehensive Evaluation

This study proposes a quality evaluation framework for scientific literature entity datasets that encompasses four critical stages: data collection, annotation, publication, and application. The applicability of this framework was validated through an empirical analysis of six datasets: ACL, ScienceIE, FTD, SciERC, ORKG-TDM, and SCIREX. The findings suggest that the credibility of data sources and a balanced data scale are key to ensuring the initial quality of the datasets. High-quality data annotation relies not only on rigorous annotation standards and processes but also requires the involvement of expert teams to enhance consistency and reliability. Furthermore, the completeness of metadata, the balance of label distribution, and data interoperability are crucial in the publication stage, as they significantly impact the effectiveness of model training. In addition to citation metrics, the actual impact of a dataset is closely linked to its update and maintenance frequency, though many of the analyzed datasets lack regular updates. Overall, the proposed evaluation framework effectively identifies the core dimensions of dataset quality. However, due to limitations in available information, this study did not fully explore aspects such as dataset scalability, long-term maintenance, and community engagement. These areas warrant further investigation in future research.

6. CHALLENGES AND PROSPECTS

6.1 Overcoming the Challenge of Dataset Label Inconsistency: Improving Annotation Methods and Noise Management Strategies

In scientific and technological literature analysis research, the dataset's quality is directly related to the performance of the model and the credibility of the research results. Among them, the problem of noise in the dataset, especially the inconsistency of labels, has become a significant challenge affecting the accuracy of model evaluation and the reliability of research conclusions. The inconsistency of labels may mislead the model to learn incorrect information, affecting the model's judgment and reasoning abilities and, thus, its performance. In the empirical study of the SCIERC dataset [41], it is pointed out that a high proportion of label errors will directly affect the reliability of the model evaluation. On the SCIERC dataset with corrected labels, performance improvements were achieved in five different NER models. Therefore, accurate data annotation and effective noise management are essential ways to improve the quality of datasets. Faced with these challenges, this article proposes several suggestions: First, by adopting more refined entity annotation methods and introducing remote supervision technology, the accuracy and consistency of annotation can be significantly improved. Second, implementing effective noise filtering mechanisms can help mitigate the impact of noise in the dataset. Finally, strengthening the standardization and transparency of dataset construction is also crucial, ensuring that every step of the construction process is thoroughly documented to enhance the reusability and scalability of the dataset.

6.2 The Necessity of Constructing Full-Text Scientific and Technological Literature Entity Datasets: Enhancing the Depth of Literature Analysis

With the in-depth scientific and technological literature study, the demand for constructing full-text datasets is increasing. Unlike those containing only abstracts or specific sentences, full-text datasets provide researchers with more comprehensive and in-depth scientific research information, which is crucial for deepening the understanding and analysis of scientific literature. By presenting the complete content of the literature, full-text datasets reveal the deep and complex information structure, which helps in scientific literature's multi-level and multi-dimensional deep exploration. Analysis of 15 million full-text English scientific and technological literature published between 1823 and 2016 shows that compared to text mining using only abstracts [42], full-text datasets can more effectively extract relationship information such as protein-protein, disease-gene, and protein subcellular localization and have a significant advantage in accuracy.

However, the analysis results show that the performance of full-text datasets in practical applications has not met expectations, highlighting the urgency and importance of strengthening the construction of full-text datasets. Faced with the dual challenges of copyright laws and technology, copyright issues require researchers to establish partnerships with publishers and academic institutions to ensure datasets' legal use and ethical standards. In addition, dealing with large volumes of long-text data and complex information structures imposes higher requirements on natural language processing technologies and machine learning algorithms. To address these challenges, developing semi-automated annotation tools that integrate artificial intelligence and manual review is crucial, aiming to improve annotation efficiency and reduce associated costs. At the same time, promoting interdisciplinary cooperation and utilizing experts' deep knowledge and rich experience in various fields ensure high quality and comprehensive coverage in dataset construction. The open access and sharing of full-text data provide a solid foundation for disseminating and deepening scientific knowledge, effectively promoting the development and innovation of the academic field.

6.3 Enhancing the Efficiency of Scientific Literature Entity Dataset Utilization: Transfer Learning and Data Interoperability

This paper explores the influence of datasets in scientific research and practical applications, especially the effectiveness of dataset utilization, reflected by the Google Scholar citation count (Figure 4), which is less than ideal. Given this, in scientific and technological literature research, especially in specialized tasks such as named entity recognition, the challenges faced by deep learning technologies highlight the urgent need to improve the efficiency of dataset utilization. Due to insufficient samples and uneven label distribution in scientific and technological literature entity datasets, there are scenarios of small samples or sparse annotated data, which affect the performance of models. To address this issue, the research community has begun to explore various new learning strategies, including multi-task transfer learning, small sample learning, and zero-shot learning [43]. In particular, the approach of multi-task transfer learning [44], which transfers knowledge learned from one task to other related tasks, has been proven to significantly improve model performance in scenarios with scarce data or limited resources. Furthermore, utilizing open-source, public datasets to construct new datasets or re-annotating existing datasets is hoped to further enhance the application value and universality of datasets.

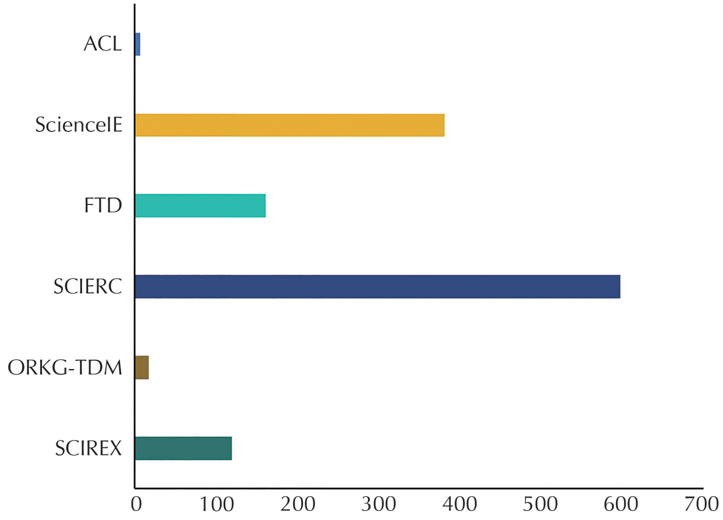


Figure 4. Google Scholar Citation Count for Datasets.

6.4 Open Source and Sharing of Scientific and Technological Literature Entity Datasets: Data Management, Incentive Mechanisms, and Maintenance

In scientific and technological literature research, especially in entity extraction and identification studies, the sharing and management of open-source datasets play a crucial role in improving research efficiency and promoting academic progress. Many research results fail to share or open-source their datasets, and researchers often build datasets only for their use, limiting the accumulation of knowledge and the iterative development of technology. To address this issue, focusing on the long-term maintenance and effective management of dataset-sharing processes is necessary. Establishing a stable and reliable data storage and sharing platform is particularly important, as it not only ensures the accessibility and permanence of datasets but also supports the continuity and in-depth exploration of scientific research. Secondly, establishing effective incentive mechanisms to encourage researchers to share datasets is crucial. Measures such as academic publishing support and financial rewards can enable researchers to actively participate in data sharing, which is of profound significance for accumulating knowledge and developing innovation. Finally, regular maintenance and timely updates are essential to ensure that datasets reflect the latest research results and developments in the field. Similar to software projects, technical updates and content maintenance of datasets are crucial to ensuring data quality and relevance. It is strongly recommended that research communities or researchers open-source and share datasets, especially in key technology areas such as entity extraction and identification, to jointly promote the formation of an open-source culture and contribute to the advancement of scientific research.

7. CONCLUSION

This study conducts an in-depth analysis of 22 open-source datasets based on scientific literature. To explore their role in advancing the application of natural language processing technology in scientific research. From the perspective of dataset lifecycle, this paper constructs an evaluation index system for scientific literature datasets. Evaluate dataset quality and annotation accuracy and examine their effects in academic and practical applications. Discover the importance of metadata completeness for datasets. Provide necessary background and structural information for academic research and practical applications. High-quality annotation processes and strict review mechanisms can significantly improve data consistency. Consequently, this enhances the effectiveness of training in machine learning models. The balance of fine-grained labels is an essential factor in assessing the effectiveness of datasets in classification tasks. It revealed the balance of category label distribution in datasets. Thus affecting the effectiveness and bias issues of subsequent model training. The study focuses on academic citation volume, reflecting its contribution to the research field and its driving role in practical applications. Perform well in specific dimensions but face challenges like data noise filtering, standardization, and continuous updating. To address these challenges, it is recommended that future research appropriately increase the analysis of more datasets, adopt more diverse assessment methods, and consider the specific needs of different field applications.

ACKNOWLEDGEMENTS

This work is supported by the Youth Backbone Talents Project of National Science Library, CAS titled "Research on the Construction of General Entity Corpus for Scientific Content Mining" (Grant No. E3550201).

AUTHOR CONTRIBUTIONS

Qinya Xu (xuqinya@mail.las.ac.cn): Theme design, data collection, and paper writing.

Qianqian Yu (yuqianqian@mail.las.ac.cn): Served as the corresponding author and contributed significantly to the conception, design, and drafting of the manuscript.

Li Qian (qianl@mail.las.ac.cn): Served as the corresponding author, theme design, revise the paper.

REFERENCES

[1] SHEN X Y, OU S Y. Research on extraction and application of knowledge units in scientific literature: reviews and prospects[J]. Information studies: theory & application, 2022, 45(12): 195–207

[2] Hou L, Zhang J, Wu O, et al. Method and dataset entity mining in scientific literature: a CNN + BiLSTM model with self-attention[J]. Knowledge-Based Systems, 2022, 235: 107621

[3] WANG Ch W, DONG Q X, SUI ZH F, et al. Quality Evaluation of Public NLP Dataset[J]. Journal of Chinese Information Processing. 2023, 37(2): 26–40

[4] QasemiZadeh B, Handschuh S. The ACL RD-TEC: a dataset for benchmarking terminology extraction and classification in computational linguistics[C]//Proceedings of the 4th International Workshop on Computational Terminology (Computerm). 2014: 52–63

[5] QasemiZadeh B, Schumann A K. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016: 1862–1868

[6] Klie J C, de Castilho R E, Gurevych I. Analyzing dataset annotation quality Management in the Wild[J]. Computational Linguistics, 2024: 1–48

[7] Krause J, Sapp B, Howard A, et al. The unreasonable effectiveness of noisy data for fine-grained recognition[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer International Publishing, 2016: 301–320

[8] Wang Z, Shang J, Liu L, et al. Crossweigh: Training named entity tagger from imperfect annotations[J]. arXiv preprint arXiv:1909.01441 (2019)

[9] Gan J, Luo J, Wang H, et al. Multimodal entity linking: a new dataset and a baseline[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 993–1001

[10] Zhu Y, Zhang P, Haq E U, et al. Can chatgpt reproduce human-generated labels? a study of social computing tasks[J]. arXiv preprint arXiv:2304.10145 (2023)

[11] Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks[J]. Proceedings of the National Academy of Sciences, 2023, 120(30): e2305016120

[12] Reiss M V. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark[J]. arXiv preprint arXiv:2304.11085 (2023)

[13] Tejaswin P, Naik D, Liu P. How well do you know your summarization datasets?[J]. arXiv preprint arXiv:2106.11388 (2021)

[14] Gururangan S, Swayamdipta S, Levy O, et al. Annotation artifacts in natural language inference data[J]. arXiv preprint arXiv:1803.02324 (2018)

[15] Picard S, Chapdelaine C, Cappi C, et al. Ensuring dataset quality for machine learning certification[C]// 2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW). IEEE, 2020: 275–282

[16] Chug S, Kaushal P, Kumaraguru P, et al. Statistical learning to operationalize a domain agnostic data quality scoring[J]. arXiv preprint arXiv:2108.08905 (2021)

[17] DONG Q X, SUI Zh F, ZHAN W D, et al. Problems and Countermeasures in Natural Language Processing Evaluation[J]. Journal of Chinese Information Processing. 2021, 35(6): 1–15

[18] DeepLearning.AI, Landing AI. Data-Centric AI Competition [EB/OL]. [2024-10-11]. https://https-deeplearning-ai.github.io/data-centric-comp

[19] Chen H, Cao G, Chen J, et al. A practical framework for evaluating the quality of knowledge graph[C]// Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding: 4th China Conference, CCKS 2019, Hangzhou, China, August 24–27, 2019, Revised Selected Papers 4. Springer Singapore, 2019: 111–122

[20] Xu L, Liu J, Pan X, et al. Dataclue: A benchmark suite for data-centric nlp[J]. arXiv preprint arXiv:2111.08647 (2021)

[21] Gupta S, Manning C D. Analyzing the dynamics of research by extracting key aspects of scientific papers[C]//Proceedings of 5th international joint conference on natural language processing. 2011: 1–9

[22] Luan Y, He L, Ostendorf M, et al. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction[J]. arXiv preprint arXiv:1808.09602 (2018)

[23] Brack A, D'Souza J, Hoppe A, et al. Domain-independent extraction of scientific concepts from research articles[C]//European Conference on Information Retrieval. Cham: Springer International Publishing, 2020: 251–266

[24] Augenstein I, Das M, Riedel S, et al. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications[J]. arXiv preprint arXiv:1704.02853 (2017)

[25] Jain S, van Zuylen M, Hajishirzi H, et al. Scirex: A challenge dataset for document-level information extraction[J]. arXiv preprint arXiv:2005.00512 (2020)

[26] Mondal I, Hou Y, Jochim C. End-to-end construction of NLP knowledge graph[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 1885–1895

[27] Hou Y, Jochim C, Gleize M, et al. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction[J]. arXiv preprint arXiv:1906.09317 (2019)

[28] D'Souza J, Auer S. Pattern-based acquisition of scientific entities from scholarly article titles[C]//International Conference on Asian Digital Libraries. Cham: Springer International Publishing, 2021: 401–410

[29] D'Souza J, Auer S. Computer science named entity recognition in the open research knowledge graph[C]//International Conference on Asian Digital Libraries. Cham: Springer International Publishing, 2022: 35–45

[30] D'Souza J, Hoppe A, Brack A, et al. The STEM-ECR dataset: grounding scientific entity references in STEM scholarly content to authoritative encyclopedic and lexicographic sources[J]. arXiv preprint arXiv:2003.01006 (2020)

[31] D'Souza J. Overview of STEM Science as Process, Method, Material, and Data Named Entities[J]. Knowledge, 2022, 2(4): 735–754

[32] Kabongo S, D'Souza J, Auer S. Automated mining of leaderboards for empirical ai research[C]//Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings 23. Springer International Publishing, 2021: 453–470

[33] Färber M, Albers A, Schüber F. Identifying Used Methods and Datasets in Scientific Publications[C]//SDU@AAAI (2021)

[34] Schindler D, Bensmann F, Dietze S, et al. Somesci-A 5 star open data gold standard knowledge graph of software mentions in scientific articles[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 4574–4583

[35] Istrate A M, Li D, Taraborelli D, et al. A large dataset of software mentions in the biomedical literature[J]. arXiv preprint arXiv:2209.00693 (2022)

[36] Heddes J, Meerdink P, Pieters M, et al. The automatic detection of dataset names in scientific articles[J]. Data, 2021, 6(8): 84

[37] Pan H, Zhang Q, Dragut E, et al. DMDD: A Large-Scale Dataset for Dataset Mentions Detection[J]. arXiv preprint arXiv:2305.11779 (2023)

[38] Hou Y, Jochim C, Gleize M, et al. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics[J]. arXiv preprint arXiv:2101.10273 (2021)

[39] Duck G, Nenadic G, Brass A, et al. bioNerDS: exploring bioinformatics' database and software use through literature mining[J]. BMC bioinformatics, 2013, 14(1): 1–13

[40] Du C, Cohoon J, Lopez P, et al. Softcite dataset: A dataset of software mentions in biomedical and economic research publications[J]. Journal of the Association for Information Science and Technology, 2021, 72(7): 870–884

[41] Zeng Q, Yu M, Yu W, et al. Validating Label Consistency in NER Data Annotation[J]. arXiv preprint arXiv:2101.08698 (2021)

[42] Westergaard D, Stærfeldt H H, Tønsberg C, et al. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts[J]. PLoS computational biology, 2018, 14(2): e1005962

[43] Kim H, Sung M, Yoon W, et al. Improving tagging consistency and entity coverage for chemical identification in full-text articles[J]. arXiv preprint arXiv:2111.10584 (2021)

[44] Brack A, Hoppe A, Buschermöhle P, et al. Cross-domain multi-task learning for sequential sentence classification in research papers[C]//Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries. 2022: 1–13

AUTHOR BIOGRAPHY

Qinya Xu is currently a doctoral student in the National Science Library, Chinese Academy of Sciences. He received his Master's degree from Yanshan University in 2018. His research interests mainly focus on knowledge organization, knowledge extraction and semantic evaluation. E-mail address: xuqinya@mail.las.ac.cn

Qianqian Yu is currently an associate research librarian of National Science Library, Chinese Academy of Sciences, Beijing, China. Her main fields of interest are data organization, data management, knowledge discovery. E-mail address: yuqianqian@mail.las.ac.cn

Li Qian is currently an senior engineer of National Science Library, Chinese Academy of Sciences, Beijing, China. His main fields in intelligent data and intelligent knowledge services, has led multiple national-level research projects, and has published over 80 academic papers.
E-mail address: qianl@mail.las.ac.cn