SciNLP: A Domain-Specific Benchmark for Full-Text Scientific Entity and Relation Extraction in NLP

Decheng Duan[1] **Yingyi Zhang**[2] **Jitong Peng**[1] **Chengzhi Zhang**[1] †
[1]Nanjing University of Science and Technology, China
[2]Soochow University, China
{akaddc,pengjit,zhangcz}@njust.edu.cn, yyzhang9@suda.edu.cn

Abstract

Structured information extraction from scientific literature is crucial for capturing core concepts and emerging trends in specialized fields. While existing datasets aid model development, most focus on specific publication sections due to domain complexity and the high cost of annotating scientific texts. To address this limitation, we introduce SciNLP—a specialized benchmark for full-text entity and relation extraction in the Natural Language Processing (NLP) domain. The dataset comprises 60 manually annotated full-text NLP publications, covering 7,072 entities and 1,826 relations. Compared to existing research, SciNLP is the first dataset providing full-text annotations of entities and their relationships in the NLP domain. To validate the effectiveness of SciNLP, we conducted comparative experiments with similar datasets and evaluated the performance of state-of-the-art supervised models on this dataset. Results reveal varying extraction capabilities of existing models across academic texts of different lengths. Cross-comparisons with existing datasets show that SciNLP achieves significant performance improvements on certain baseline models. Using models trained on SciNLP, we implemented automatic construction of a fine-grained knowledge graph for the NLP domain. Our KG has an average node degree of 3.2 per entity, indicating rich semantic topological information that enhances downstream applications. The dataset is publicly available at: https://github.com/AKADDC/SciNLP.

1 Introduction

Scientific Named Entity Recognition (SciNER) and Scientific Relation Extraction (SciRE) have emerged as two pivotal technologies for scientific literature mining. SciNER focuses on identifying and categorizing domain-specific scientific entities in publications(Jeong and Kim, 2022; Zhang et al., 2021; Zhong and Chen, 2021a; Beltagy et al., 2019), while SciRE detects semantic relationships between entities and classifies them into predefined relation types(Yan et al., 2021; Chen et al., 2020; Ji et al., 2020; Wadden et al., 2019). These tasks play essential roles in downstream applications such as scientific leaderboard construction, academic Q&A systems, knowledge graph building, text summarization, and method recommendation(Wang et al., 2024; Cao et al., 2021; Zhang et al., 2024a; Kedia et al., 2021). Compared to the construction of annotation corpora for traditional Named Entity Recognition (NER) and Relation Extraction (RE) tasks, SciNER and SciRE face significantly greater challenges in corpus development, primarily due to the complexity of scientific texts. First, scientific literature contains a vast array of highly specialized domain-specific terminology, with notable cross-disciplinary variations in terminology systems. Characteristics such as entity overlap and relational complexity require annotators to possess domain expertise. Consequently, high-quality annotated data within a domain typically demands extensive time investment from subject-matter experts. Constructing such corpora for niche disciplines often incurs prohibitively high costs. Second, scientific texts frequently exhibit entity nesting and semantic overlap. Third, scientific relation-

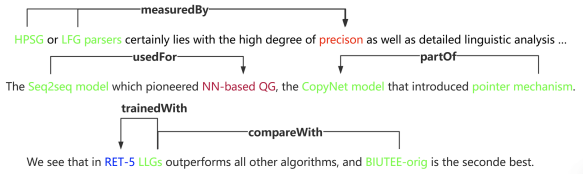

Figure 1: An annotated example from our dataset, there are four entity types:model,metric,dataset,task, The connecting lines above the sentence represent relationships between entities. In this example, the relationships include measuredBy, usedFor, partOf, trainedWith, and compareWith. Lines with arrows indicate directed relationships, while lines without arrows represent bidirectional relationships.

†Corresponding authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 14485–14498
November 4-9, 2025

ships demonstrate intricate structural characteristics. They encompass both fundamental method-task associations and complex inter-method relationships . This complexity inherently necessitates heavy reliance on domain knowledge during annotation processes.

We introduce the SciNLP benchmark dataset, the first gold-standard annotated corpus specifically designed for scientific entity and relation extraction tasks in the NLP domain. Figure 1 illustrates an annotated example from our dataset, demonstrating the predefined complex entity and relation types tailored for the NLP domain. Our dataset distinguishes itself from existing resources through two core differentiating features.

First, unlike existing scientific corpora that predominantly focus on limited annotation scope to abstracts (Ohta et al., 2002; Li et al., 2016), we systematically annotate entities and their relationships across entire scientific papers to ensure contextual completeness during entity extraction—a significantly more challenging task compared to annotating only abstracts or specific sections. This full-text annotation approach enables capturing long-range dependencies and implicit relationships that are crucial for understanding NLP research methodologies.

Second, while current domain-specific datasets mainly target general AI/machine learning concepts (Pan et al., 2024; Luan et al., 2018), our dataset is specifically designed and annotated for the NLP domain, addressing the critical gap of high-quality annotated resources in NLP, particularly those based on full-text publications. This domain specialization allows precise modeling of NLP-specific research patterns that generic scientific datasets cannot capture. Furthermore, where traditional fine-grained datasets in machine learning or AI typically adopt the simplified TDM framework (Task, Dataset, Method), our dataset introduces more granular entity types (Task, Dataset, Model, and Metric) and richer relation types to capture intricate interactions between method entities. Compared to existing relation schemas that focus on basic comparison or usage relationships, our novel relation taxonomy specifically models the evolutionary dynamics and compositional nature of NLP methods. These relationships, often requiring annotation from main paper sections that previous segment-level annotations neglect, enable deeper semantic characterization of NLP research dynamics, particularly the progression of model architectures and evaluation methodologies over time.

The primary contributions of this work are three folds:

We construct SciNLP, the first manually curated, full-text scientific publication dataset for the NLP domain, comprising 60 ACL conference long papers (2001–2024) with comprehensive semantic annotations. The dataset includes 7,072 fine-grained entities and 1,826 complex relations, addressing the limitations of existing scientific information extraction resources confined to abstracts or narrow subdomains. SciNLP pioneers systematic full-text entity annotation for NLP literature.

A hierarchical ontology framework is developed to reflect NLP's knowledge structure, defining 4 core entity types (Tasks, Methods, Datasets, Metrics) and 11 relation types. This framework captures intricate interactions among NLP's core entities, enabling precise annotation of the NLP domain's scientific knowledge.

We validate SciNLP's effectiveness through experiments on SOTA pipeline and joint extraction models. Models trained on SciNLP demonstrate successful end-to-end joint extraction from unannotated literature, proving the feasibility of constructing fine-grained NLP knowledge graphs.

2 Related Work

Within the domain of scientific information extraction research, extant scholarly endeavors have established numerous curated datasets for SciNER and SciRE, yet these resources exhibit marked heterogeneity in their annotation schemas and applicability contexts.

2.1 Scientific Named Entity Recognition

The evolution of scientific literature entity extraction datasets demonstrates a paradigm shift from partial annotation to full-text annotation and from coarse-grained classification to refined taxonomies(Li et al., 2020). Early initiatives like SEMEVAL-2017(Augenstein et al., 2017) concentrated on materials science, annotating three entity types (task/method/material) within selected passages of 500 publications, albeit constrained by limited annotation density and semantic coverage. Recent advancements, exemplified by GSAP-NER(Otto et al., 2023), target computer science and machine learning domains, applying 10 fine-

grained labels across 100 full-text articles to statistically annotate 54,598 entities, substantially enhancing entity density. The DMDD initiative (Pan et al., 2023) extends coverage to computer science, NLP, and biomedical fields, combining manual annotation of 450 scientific publications with distant supervision techniques to automatically label an additional 31,219 papers. Furthermore, SciDM(Pan et al., 2024) establishes three ML-specific entity categories (dataset/method/task), employing a hybrid "precision manual annotation + distant supervision" strategy to annotate 48,149 machine learning full-text articles. This effort yields a corpus of 1.8 million weakly annotated entities, systematically linked to the Paper with Code repository.

2.2 Scientific Entity Relation Extraction

Furthermore, several datasets concurrently support both entity and relational extraction. Early benchmarks like SEMEVAL-2018(Gábor et al., 2018) annotated abstracts in NLP domains, establishing six binary relations including USAGE, though omitting explicit entity type specifications. SCIERC(Luan et al., 2018) introduced seven scientific relations across 500 AI-focused paper abstracts. Contemporary developments witness SciREX(Jain et al., 2020) pioneering n-ary relation annotation in ML/NLP domains, albeit without formal n-ary relation definitions, ultimately documenting 360 entities, 169 binary relations, and 5 quaternary relations per paper on average. SciER(Zhang et al., 2024b) represents a paradigm advancement, defining nine relational categories based on SciDMT's framework and annotating 106 full-text AI publications to yield 24,000 entities and 12,000 relations. This progression mirrors the evolutionary trajectory observed in entity extraction datasets, demonstrating parallel developmental patterns between relational and entity annotation initiatives.

Reviewing the extant annotated datasets reveals significant imbalances in disciplinary coverage and annotation depth across current scientific publication resources for SciIE. Cross-domain corpora such as predominantly concentrate on AI and ML domains(Luan et al., 2018; Pan et al., 2024; Zhang et al., 2024b), while specialized disciplinary datasets exhibit marked fragmentation in textual coverage, typically constrained to specific paper sections or isolated sentences(Jain et al., 2020; Gábor et al., 2018), thereby compromising the integrity of holistic scientific discourse. Addressing these limitations, this study proposes the creation of SciNLP – the inaugural full-text annotation benchmark for NLP research. Anchored in comprehensive semantic units from 60 ACL proceedings (2001-2024), the corpus achieves granular annotation of 7,072 entities and 1,826 relations. Furthermore, it establishes a sophisticated taxonomy of entity and relational categories specifically tailored to the structural and conceptual composition of NLP scholarly articles.

3 SciNLP Corpus

This section we delineate the methodological framework underlying the SciNLP corpus construction, including Data Curation and Preprocessing Pipeline in 3.1, Annotation Protocol and Quality Assurance Mechanisms in 3.2, and Comparative Analysis with Benchmark Corpora in 3.3.

3.1 Data Collection and Pre-processing

Our data collection began by integrating the ACL OCL Corpus[1], which aggregates full-text papers and metadata from the ACL Anthology platform[2] up to September 2022, totaling 73,285 papers. To address the corpus's update lag, we additionally acquired 9,387 newly published papers (October 2022 to December 2024) from the ACL Anthology, extracting their PDF content and aligning metadata using the document parsing tool $GROBID$[3], resulting in a total collection of 82,672 papers.Given the academic authority and technical cutting-edge nature of ACL Long Papers (Association for Computational Linguistics Long Papers)—which best represent NLP's technological evolution—we applied stratified sampling principles to randomly select 60 papers from the full-text corpus of ACL Long Papers (2001–2024) as our annotated sample set (see Appendix A for the specific sampling strategy).

3.2 Comparison with Previous Datasets

Annotation Scheme: We constructs an information extraction annotation framework tailored to the structured characteristics of academic literature in the NLP domain. Specifically, we define four representative entity types for NLP publications—Task, Model, Dataset, and Metric—by synthesizing domain-specific requirements and prior knowledge

[1]https://github.com/shauryr/ACL-anthology-corpus
[2]https://aclanthology.org/
[3]https://github.com/kermitt2/grobid

Items	SemEval18	TDMSci	SciNLP
Annotation range	Abstract	Sentence	Full text
Entity types	-	3	4
Relationship types	6	-	11
Entities	-	2937	7072
Relationship	1595	-	1826
Documents	500	-	60

Table 1: Comparison of SciNLP and 2 datasets in NLP domain, the "-" indicates that there is no corresponding data in the dataset.

entity classifications for NLP(QasemiZadeh and Schumann, 2016). These entity types are prevalent in core sections of NLP papers, such as methodology and experimental design. Annotation guidelines explicitly instruct annotators to label only factual entities with concrete meanings, ensuring consistency and relevance.You can see Appendix A for the specific annotation scheme.

For relation annotation, building on prior work in ML, AI, and NLP domains(Luan et al., 2018; Zhang et al., 2024b) and aligning with the structural conventions of NLP literature, we define 11 relation types to represent complex interactions among the four entity categories: MeasuredBy, UsedFor, TrainedWith, EnhancedBy, SubclassOf, SubtaskOf, PartOf, EvaluatedBy, EvaluatedOn, SimilarWith, and CompareWith. Excluding the symmetric relations SimilarWith and CompareWith, all other relation types are directional. Table 2 provides some annotation examples of these relations. Compared to existing studies, our framework offers finer-grained distinctions, particularly for interactions between Model entities, where we employ three asymmetric relations—SubclassOf, EnhancedBy, and PartOf—to precisely characterize hierarchical, enhancement, and compositional dependencies. Detailed specifications of entity-relation labels are provided in Appendix A.

Annotation Strategy: We employs *Doccano*[4] as the annotation tool, with a two-stage annotation process. First, the 60 papers are evenly divided into five chronological groups (12 papers per group). Two NLP experts conducted a pilot annotation on one group: they drafted initial entity annotation guidelines, independently annotated the papers, then resolved discrepancies through discussions to refine the guidelines. These optimized guidelines were used to train five NLP graduate students for subsequent annotations. The remaining four groups were assigned to the students, with each paper annotated by at least two annotators. Inter-annotator agreement analysis revealed Cohen's kappa scores (Andres et al.) of 0.90, 0.91, 0.83, and 0.86 for entity annotation across the four groups, and 0.75, 0.81, 0.71, and 0.64 for relation annotation, demonstrating reliable annotation consistency.

[4]https://github.com/doccano/doccano

3.3 Data Analysis

Upon completing the annotation process, our dataset comprises 7,072 fine-grained entities and 1,826 relations, with an average of 121 entities and 31 relations per document. As shown in Table 1, the scale of our dataset substantially exceeds prior NLP-specific datasets designed for entity and relation extraction. The annotated data was then randomly shuffled and split into training, validation, and testing sets in an 8:1:1 ratio to support subsequent experimental analyses.

4 Document-Level Entity and Relation Extraction

This section primarily elaborates on our definition of the full-text-based Entity and Relation Extraction (ERE) task (Section 4.1) and details the evaluation metrics adopted for assessment (Section 4.2).

4.1 Task Definition

The end-to-end Entity and Relation Extraction (ERE) task aims to automatically extract structured triplets T = (e_i, r_k, e_j) from scientific literature D, where each triplet represents two scientific entities (e_i, e_j) and their semantic relation r_k. Current research typically decouples this task into two core subtasks: SciNER (Scientific Named Entity Recognition) and SciRE (Scientific Relation Extraction). Existing methods primarily follow two technical paradigms: (1) Pipeline approaches sequentially execute SciNER and SciRE—first identifying entity boundaries/types, then classifying relation types between entities; (2) Joint learning frameworks synchronously optimize both subtasks through parameter sharing or unified annotation strategies, effectively mitigating error propagation issues inherent in traditional pipeline methods.

As our dataset is annotated at the full-text level, the basic data unit is document-based. For modeling, we input a scientific document D, which comprises multiple sentences P = p_1, p_2, ..., p_n, each composed of word sequences S = w_1, w_2, ..., w_m. Based on this structure, we formally define

Relation type	Examples	Example sentences
measuredBy	HPSG→precision	HPSG or LFG parsers certainly lies with the high degree of precision as well as detailed linguistic analysis . . .
evaluatedBy	SQG→CoQA	Prior works performed SQG on CoQA
usedFor	Seq2Seq architecture→NN-based QG	Du pioneered NN-based QG by adopting the Seq2Seq architecture. . .
enhancedBy	Seq2Seq modek→variational inference	Enhancing the Seq2Seq modek into complicated strutures using variational inference. . .
evaluatedOn	QG→SQuAD ; QG→NewsQA	QG is ofen performed on existing QA datasets , e.g. , SQuAD , NewsQA , etc.
partOf	self-attention→multi-head attention	In the multi-head attention layer , there are not only lh vanilla self-attention . . .
compareWith	CorefNet—ReDR	Besides , CorefNet gets better performance among all baselines , especially ReDR
subtaskOf	SQG→conversation generation	Second , since prior works regarded SQG as a conversation generation, . . .
similarWith	Boltzmann-machine—BM	We propose to adopt the Boltzmann-machine (BM) distribution as a variational. . .
trainedWith	BERT→Persona-Chat ; GPT-2→Persona-Chat	Group 1 applies direct fine-tuning of BERT or GPT-2 on Persona-Chat. . .
subclassOf	RPART→CART	RPART , which is a CART reimplementation for the S-Plus and R statistical computing enviroments. . .

Table 2: Examples of the relations types, the four entity types: Method, Metric, Dataset, Task.

the SciNER and SciRE tasks as follows:

SciNER: For each predefined entity type E, the SciNER model identifies all possible entities in every sentence of an input document. It detects the span $Span_i w_l$, w_r and predicts the entity type $c_i \in E$ for each entity, where w_l denotes the leftmost index and w_r the rightmost index of the entity span within the word sequence W = w_1, w_2, ..., w_n.

SciRE: Given a predefined set of relation types R, the SciRE model predicts whether a relation $r_i \in R$ exists between every pair of entities (e_i, e_j) within the same sentence P, where entities are drawn from the set of entity mentions Entity.

4.2 Evaluation Settings and Results

To systematically evaluate the performance differences between joint Entity and Relation Extraction (ERE) and pipeline approaches, this study establishes granular evaluation frameworks for each subtask:

Named Entity Recognition (NER) Evaluation: Adopts strict span matching criteria, requiring models to correctly identify both entity boundaries and semantic types.

End-to-End Relation Extraction (End-to-End RE) Evaluation:For the ERE extraction results, we add two additional indicators Rel and Rel+ to evaluate the accuracy.

Boundary-Aware Metric (Rel): Following prior work(Zhong and Chen, 2021b; Ye et al., 2022; Yan et al., 2023), a relation is considered correct only if: (a) Subject entity boundaries are correctly identified; (b) Object entity boundaries are correctly identified; (c) Relation type is accurately predicted.

Strict Metric (Rel+): Extends Rel by additionally requiring correct prediction of subject/object entity types.

Pure Relation Classification (RE) Evaluation:Evaluates models' ability to discern semantic relationships between entities under gold-standard entity annotations. Specifically, for any entity pair: If a predefined relation exists, the model must predict its type correctly; If no relation exists, the model must correctly identify the absence of a relation.

5 Experiments

In this section, we utilize SciNLP to train SOTA supervised methods for SciNER and SciRE tasks, and compare them with existing datasets to validate

Methods	Sentence-level Test				Doc-level Test			
	NER	Rel	Rel+	RE	NER	Rel	Rel+	RE
PURE(Zhong and Chen, 2021b)	59.42	57.51	56.15	57.03	60.25	59.93	58.35	59.93
PL-Marker(Ye et al., 2022)	64.52	**59.51**	58.07	**60.94**	67.15	61.46	59.24	**62.16**
HGERE(Yan et al., 2023)	**92.18**	59.26	**58.89**	-	**94.15**	**61.74**	**60.76**	-

Table 3: Comparison F1 scores between full text and sentence on test set. “Rel” and “Rel+” denote the results of end-to-end relation extraction under boundaries evaluation and strict evaluation, respectively. “RE” indicates performing relation extraction with given gold standard entities, applicable only to pipeline extraction methods. "-" indicates that this model is not capable of handling the task.

Methods	SciNLP				SCIERC				SciER			
	NER	Rel	Rel+	RE	NER	Rel	Rel+	RE	NER	Rel	Rel+	RE
PURE	60.25	59.93	58.35	59.93	66.68	48.65	35.68	37.54	81.60	53.27	52.67	73.99
PL-Marker	67.15	61.46	59.24	62.16	69.94	53.25	41.62	43.05	83.31	60.06	59.24	77.11
HGERE	94.15	61.74	60.76	-	74.91	55.72	43.6	-	86.85	62.32	61.10	-

Table 4: F1 scores of different baselines on our proposed dataset.

the efficacy of our dataset. Section 5.1 introduces the supervised Baselines we have used. Section 5.2 presents the results of our dataset across these methods, followed by detailed discussions and analyses.

5.1 Supervised Baselines

To validate the effectiveness of our dataset, we selected state-of-the-art (SOTA) methods for SciNER and SciRE tasks, including the following three approaches:

PURE: (Zhong and Chen, 2021b) This model employs a dual-encoder architecture to decouple entity recognition and relation extraction tasks. The entity model performs span-based type prediction, while the relation model enhances classification through dynamic entity pair inputs and the insertion of entity-type markers. It introduces crosssentence context expansion by extending input windows, leveraging pre-trained models' ability to capture long-range dependencies for improved global semantic reasoning.

PL-Marker: (Ye et al., 2022) Proposing a floating marker mechanism, this approach dynamically binds markers to text spans during the entity recognition phase through shared positional encoding and boundary-aware attention alignment, with neighborhood clustering further optimizing boundary learning. In the relation extraction phase, it restructures spans and markers via subject-centric reorganization, jointly encoding subject-object semantics to overcome limitations of independent span-pair modeling. This unified framework enables seamless integration of entity and relation representations.

HGERE: (Yan et al., 2023) Replacing strict NER modules with a lightweight pruning module, this method generates high-recall candidate entities, transferring classification tasks to a joint module to mitigate error propagation. It constructs hypergraph structures (comprising entity nodes, relation nodes, and multi-dimensional hyperedges) to aggregate high-order semantic information through hierarchical message passing. This design significantly outperforms traditional GNNs limited to pairwise interactions, particularly in capturing complex scientific logic across long documents.

All methods were implemented using the SciBERT-scivocab-uncased pre-trained model as the text encoder. For fair comparision, experiments adopted identical random seed sets to ensure robustness, along with consistent fine-tuning settings: learning rate = 2e-5, weight decay = 0.01, batch size = 16, and context length = 100.

5.2 Result Analysis

For task evaluation, we assessed the performance of end-to-end models on each subtask and conducted ablation experiments to verify whether full-text annotation improves extraction performance compared to segment-specific approaches. Specifically, we processed SciNLP into two variants: **Full-Text Annotation**: Retains the complete document content, providing models with rich contextual information. **Segment-Restricted Annotation**: Extracts only sentences containing entities and relations, forming a sentence-level dataset with the context window length set to 0 to isolate local semantic dependencies. As demonstrated in Table 3, models trained on the full-text variant achieved superior performance in capturing cross-paragraph scientific logic and fine-grained relationships, validating the

necessity of full-text annotation for robust scientific information extraction.

Systematic analysis of the ablation study results in Table 3 reveals the following key findings: First, in document-level context modeling, all models exhibit significantly superior performance on the full-text annotated dataset compared to sentence-level settings, with SciRE demonstrating the most notable improvement. For instance, the PL-Marker model achieves a 1.17 percentage point gain on the Rel+ task under document-level training versus sentence-level training, while RE metrics show even greater enhancements, underscoring the critical role of cross-sentence contextual information in supporting complex relation reasoning. Notably, during relation-enhanced tasks (Rel+), document-level training ensures stable performance gains across models, with HGERE exhibiting the largest improvement due to the advanced capability of its dynamic graph attention mechanism in integrating global semantic information. These experimental results empirically validate that optimizing scientific information extraction models relies not only on algorithmic innovation but also on synergistic evolution with contextually complete annotation frameworks, where data richness and architectural design mutually reinforce performance.

Table 4 shows a comparison between SciNLP and scientific datasets like SCIERC and SciER, and it reveals critical differences and their impact on model performance: First, SciNLP's significantly smaller entity annotation count compared to SciER results in a notable performance gap for traditional sequence labeling models like PURE in NER tasks (SciNLP vs. SciER). However, SciNLP achieves comparable NER performance to SCIERC, validating the strong dependency of sequence models on training data volume and their performance ceilings under annotation scale constraints. Furthermore, SciNLP and SciER exhibit distinct advantages in complex relation modeling for the Rel+ task. For instance, HGERE underperforms on SCIERC's Rel+ task compared to SciNLP and SciER, a disparity attributable to SciNLP/SciER's cross-document triplet annotation framework, which deeply encodes scientific semantic networks through inter-disciplinary knowledge linkages. These findings highlight how annotation design, particularly full-text and domain-specific strategies, fundamentally shapes model capabilities in capturing scientific knowledge structures.

Cross-domain transfer experiments further reveal SciNLP's domain-specific specialization. When transferring the HGERE model from SciNLP to SCIERC, its NER performance degrades significantly—a phenomenon attributable to systematic differences in disciplinary focus: SciNLP employs NLP-centric deep annotation, while SCIERC adopts a cross-disciplinary paradigm. Experiments confirm that domain-focused annotation strategies effectively enhance intra-domain precision—under identical model architectures, SciNLP's disciplinary specialization achieves 92.4 entity recognition precision in computational linguistics, far surpassing cross-disciplinary scenarios. This "depth vs. breadth" trade-off demonstrates that specialized annotation frameworks reduce inter-domain distribution shifts, significantly improving models' ability to capture domain-specific semantic patterns, albeit at the cost of cross-domain generalization capabilities.

6 Building Fine-grained KG in NLP Domain

We construct a scientific knowledge graph from the entire ACL corpus. First, we built a comprehensive corpus containing full texts and metadata of all ACL conference papers (2001–2024) from the ACL Anthology repository, totaling 82,672 papers. Leveraging the HGERE model, we performed end-to-end triplet extraction across all full-text papers to build a fine-grained NLP knowledge graph, where nodes represent extracted scientific entities and edges denote semantic relationships between entities. This framework enables systematic encoding of domain-specific knowledge structures and cross-document scientific logic.

The construction of our knowledge graph involves two key steps: First, the model extracts all triplet data T from each paper, preserving every triplet in the form (e_i, r_k, e_j), where e_i and e_j are related entities, and r_k denotes their semantic relationship. Second, we perform entity normalization to ensure uniqueness within the knowledge graph. This process begins by creating an abbreviation-full form mapping table between entities using the predefined similar relationship. Subsequently, we unify entities using the entity normalization method proposed by(Zhang et al., 2024a), which calculates similarity scores based on lexical features and applies clustering algorithms to standardize entity representations. This dual-step approach guaran-

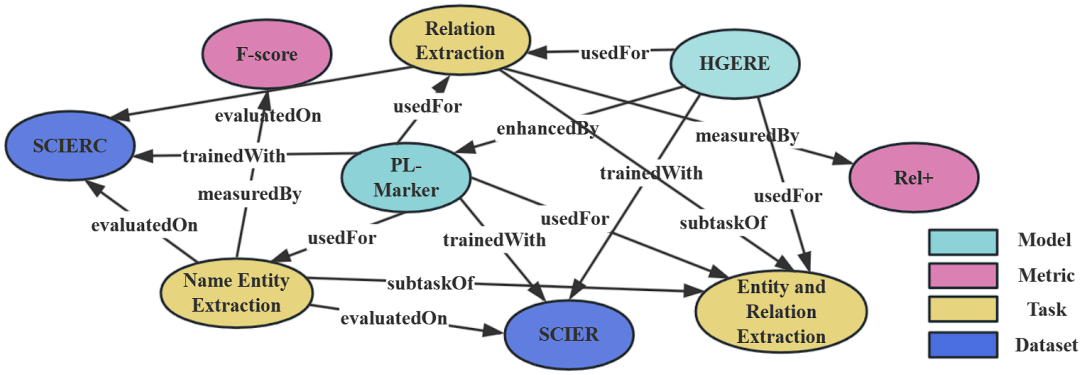

Figure 2: A subset of the NLP fine-grained KG

tees both the richness and semantic consistency of the knowledge graph. In the end, we constructed a fine-grained NLP knowledge graph comprising 232k nodes and 743k edges, with an average node degree of 3.2. Figure 2 illustrates a subset of the constructed graph, demonstrating its ability to effectively preserve domain-specific knowledge in NLP. This topology reflects the intricate semantic networks inherent to computational linguistics research.

Table 5 presents the statistics of nodes and relations in the KG we constructed. It can be observed that in the distribution of entity type quantities, entities of the "model" type account for a relatively high proportion. This indicates that NLP domain papers demonstrate a significant model-centric research paradigm, reflecting the core characteristics of rapid technological iteration and model diversification in this field. Additionally, the predominant occurrence of "usedFor" and "measuredBy" relationships highlights the characteristic of method and model-driven technological development in the NLP domain.

Nodes		**Realtions**	
Type	**Count**	**Type**	**Count**
Task	38,129	measuredBy	125,494
Dataset	35,756	evaluatedBy	26,080
Metric	34,809	usedFor	184,334
Model	124,106	enhancedBy	34660
		evaluatedOn	40,465
		partOf	47,669
		compareWith	82,599
		subtaskOf	43,372
		similarWith	40,689
		trainedWith	97,723
		subclassOf	20,179
Total	232,800	**Total**	743,264

Table 5: Nodes and relations statistics

7 Conclusion

We present SciNLP, a dataset specifically designed for fine-grained entity and relation extraction in the Natural Language Processing (NLP) domain. The dataset defines four entity types (Task, Method, Dataset, Metric) and 11 relation types (MeasuredBy, UsedFor, TrainedWith, EnhancedBy, SubclassOf, SubtaskOf, PartOf, EvaluatedBy, EvaluatedOn, SimilarWith, CompareWith), overcoming the limitations of coarse-grained relation definitions in existing resources. Constructed from 60 ACL conference papers spanning 2001–2024, SciNLP contains 7,072 entities and 1,826 relations. Unlike cross-disciplinary or abstract-level datasets, SciNLP preserves full-text scientific context through document-level annotation, providing critical support for models to capture cross-paragraph semantic dependencies. Experimental results demonstrate that models trained on SciNLP significantly outperform sentence-level settings in document-context modeling. For instance, the HGERE model achieves a 17.16 point improvement on the strict relation classification (Rel+) task compared to the cross-disciplinary SciERC dataset, validating the enhanced capability of domain-specific annotation for complex relation reasoning. Leveraging SciNLP, we automatically extracted knowledge from 82,672 NLP publications, constructing a fine-grained knowledge graph with 232k nodes and 743k edges, where entities exhibit an average node degree of 3.2. SciNLP establishes a high-confidence benchmark for NLP knowledge mining, with its full-text annotation paradigm and fine-grained ontology design enabling downstream applications such as domain-specific QA and technological trend analysis.

Limitations

While SciNLP holds pioneering value as the first full-text annotated dataset for the NLP domain, it exhibits several limitations: First, in terms of dataset scale, SciNLP's entity and relation coverage is smaller compared to cross-domain resources like SciER. Second, it does not address the widespread phenomenon of nested entities in scientific texts. Third, structural parsing errors in some full-text papers (via Grobid) introduced annotation inconsistencies. Finally, the constructed knowledge graph lacks integration with SOTA KG construction methods, resulting in suboptimal entity disambiguation and alignment rates. These gaps highlight directions for future improvements, such as incorporating nested entity modeling and robust cross-document linking frameworks.

Ethical Statement

Our **SciNLP** dataset is constructed based on resources from The ACL OCL Corpus[5], a large corpus which provides full-text and metadata to the ACL anthology collection, and it is released under the CC BY-NC 4.0[6]. All the other data are public from scientific documents. We release dataset for scientific information extraction tasks. There are no risks in our work.

In the process of the dataset construction, all human annotators are from our research group IR&TM, and all members participate voluntarily. The SciNLP dataset includes metadata extracted from the papers. No sensitive personal data (e.g., contact details or affiliations) is included. All metadata was collected in compliance with the terms of their sources and is used strictly for noncommercial academic research. We are committed to handling the data responsibly and ethically and will release our dataset under the same non-commercial license to ensure transparency and responsible data usage.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.72074113). We gratefully acknowledge all the members of the IR&TM team who participated in the annotation of data sets. We thank Yuxin Xie, Heng Zhang for providing the basic data sets and the related codes.

[5]https://github.com/shauryr/ACL-anthology-corpus

[6]https://creativecommons.org/licenses/by-nc/4.0/deed.en

References

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Xianshuai Cao, Yuliang Shi, Han Yu, Jihu Wang, Xinjun Wang, Zhongmin Yan, and Zhiyong Chen. 2021. Dekr: Description enhanced knowledge graph for machine learning method recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 203–212, New York, NY, USA. Association for Computing Machinery.

Yanguang Chen, Yuanyuan Sun, Zhihao Yang, and Hongfei Lin. 2020. Joint entity and relation extraction for legal documents with legal feature enhancement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1561–1571, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Yuna Jeong and Eunhui Kim. 2022. Scideberta: Learning deberta for science technology documents and fine-tuning information extraction tasks. *IEEE Access*, 10:60805–60813.

Bin Ji, Jie Yu, Shasha Li, Jun Ma, Qingbo Wu, Yusong Tan, and Huijun Liu. 2020. Span-based joint entity and relation extraction with attention-based span-specific and contextual semantic representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 88–99, Barcelona,

Spain (Online). International Committee on Computational Linguistics.

Akhil Kedia, Sai Chetan Chinthakindi, and Wonho Ryu. 2021. Beyond reptile: Meta-learned dot-product maximization between gradients for improved single-task regularization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 407–420, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. 2002. The genia corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 82–86, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Wolfgang Otto, Matthäus Zloch, Lu Gan, Saurav Karmakar, and Stefan Dietze. 2023. GSAP-NER: A novel task, corpus, and baseline for scholarly entity extraction focused on machine learning models and datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8166–8176, Singapore. Association for Computational Linguistics.

Huitong Pan, Qi Zhang, Cornelia Caragea, Eduard Dragut, and Longin Jan Latecki. 2024. SciDMT: A large-scale corpus for detecting scientific mentions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14407–14417, Torino, Italia. ELRA and ICCL.

Huitong Pan, Qi Zhang, Eduard Dragut, Cornelia Caragea, and Longin Jan Latecki. 2023. DMDD: A large-scale dataset for dataset mentions detection. *Transactions of the Association for Computational Linguistics*, 11:1132–1146.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1862–1868, Portorož, Slovenia. European Language Resources Association (ELRA).

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.

Zhaohui Yan, Songlin Yang, Wei Liu, and Kewei Tu. 2023. Joint entity and relation extraction with span pruning and hypergraph neural networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7512–7526, Singapore. Association for Computational Linguistics.

Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. A partition filter network for joint entity and relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 185–197, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

Heng Zhang, Chengzhi Zhang, and Yuzhuo Wang. 2024a. Revealing the technology development of natural language processing: A scientific entity-centric perspective. *Inf. Process. Manage.*, 61(1).

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. 2024b. SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13083–13100, Miami, Florida, USA. Association for Computational Linguistics.

Tao Zhang, Congying Xia, Philip S. Yu, Zhiwei Liu, and Shu Zhao. 2021. PDALN: Progressive domain adaptation over a pre-trained model for low-resource cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods*

in Natural Language Processing, pages 5441–5451, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021a. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021b. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

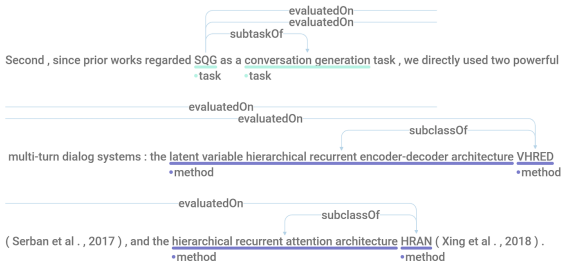

Figure 3: Doccano Annotation interface

A Annotation Guideline

This section outlines the annotation process for identifying entities and relationships in NLP literature. The workflow includes the use of annotation tools, the design of the annotation process, the definition of the entity types and relation types, and concrete examples of annotated content, which will be detailed sequentially below.

A.1 Annotation Tool

This data annotation task focuses on four types of entities in full-text NLP academic papers: algorithm entities, dataset entities, metric entities, and task entities. The original full-text corpus is sourced from the ACL Anthology Reference Corpus, where papers have been converted from XML/PDF formats to TXT format. For this study, we randomly selected 60 papers to build the training set for the automatic method of entity extraction. The manual annotation of method/task entities in these 60 papers was conducted using the online annotation platform Doccano.Figure 3 shows our annotation interface.

A.2 Entity Annotation

A.2.1 Entity Types

We annotate the following four entity types, Table 6 shows examples of correct entity annotations:

- **Task**: A task constitutes an objective requiring operational resolution through systematic processes.(e.g., Text Generation, Text Classification, Sentiment Analysis, Named Entity Recognition, and Automatic Keywords Extraction).

- **Model**: Model entities refer to specific algorithms or model architectures explicitly mentioned in scientific literature or technical documentation, designed to solve computational tasks.(e.g., GPT, LDA, SVM, LSTM, Skip-gram, Adam, BERT).

- **Dataset**: Dataset entities represent structured data resources explicitly used or referenced in scientific research for training, evaluation, or analysis. (e.g., Brown Corpus, Penn Treebank, WordNet).

- **Metric**: Metric entities denote quantitative measures used to evaluate the performance of models, algorithms, or experimental results in scientific research. (e.g., Accuracy, Precision, Recall, F1-score, BLEU, ROUGE, Kappa).

A.2.2 Annotation Note

- Determiners (e.g., "a," "the," "some," "our") placed before nouns for grammatical purposes should not be included in the entity span. For example, in the phrase "the FB15K dataset," only "FB15K dataset" should be annotated as the resource entity. However, if it is a comparison between models, it is necessary to judge whether to retain the modifiers before and after the entity according to the context.

- Annotators must mark the smallest span necessary to represent the core meaning of an algorithm/tool/resource/metric entity. Non-essential adjectival modifiers (e.g., adjectives) that are not integral to the entity's definition should be excluded. For instance, in "novel BERT-based model," only "BERT-based model" would be annotated (omitting "novel" if it is not part of the formal name).

- Do not annotate anonymous entities that include anaphoric expressions (e.g., "this algorithm," "this metric," or "the dataset"). Only annotate "factual, content-bearing" entities. Algorithm/tool/resource/metric entities typically have specific names and maintain consistent meanings across different papers. These entities are usually single terms, phrases, or abbreviations. However, even if an entity is "factual and content-bearing," it should not be annotated if its meaning is confined to a single paper's specific research topic or problem (e.g., an ad-hoc method named only within that paper's context).

- If a method/algorithm entity ends with terms such as "algorithm," "model," "framework," "strategy," "dataset," "corpus," or "toolkit," annotate the entire span, including the suffix. For example, for the algorithm "Support Vector

Machine (SVM)," if it appears as "SVM algorithm" in a paper, annotate the full term "SVM algorithm." If it appears as "SVM" alone without a suffix indicator, annotate only "SVM." This ensures consistency in capturing formal designations while adhering to the minimal span principle when no clarifying suffix is present.

- Generic terms for method categories (e.g., "classification algorithm," "semi-supervised method," "machine learning," or "deep learning") should not be annotated. Only annotate specific, named instances with well-defined meanings, such as concrete algorithm entities (e.g., SVM, KNN, Decision Trees). This ensures that annotations focus on identifiable, context-independent entities rather than broad conceptual categories.

A.3 Entity Relation Annotation

A.3.1 Entity Relation Types

And for the relations, we have defined the following 11 entity relationships, Table 7 shows examples of correct relation annotations:

- **measuredBy**: This relationship indicates that the performance of algorithm/model M is typically evaluated using metric m.
- **usedFor**:This relationship indicates that the algorithm/model M is typically applied to the NLP task T
- **trainedWith**: This relationship indicates that algorithm/model M is pre-trained on dataset D.
- **compareWith**: This relationship indicates that authors compare two entities of the same type.
- **enhancedBy**: This relationship indicates that method M_1 can address its own limitations through method M_2 to achieve better performance.
- **subclassOf**: This relationship indicates that there is a hypernym-hyponym relationship between the two entities. The hyponym has a narrower scope, while the hypernym has a broader scope. The hyponym inherits the characteristics of the hypernym, and its semantic meaning encompasses or includes that of the hypernym. Additionally, the hyponym possesses unique characteristics of its own.
- **subtaskOf**: This relationship indicates that task T_1 is a subtask or hyponym of another task T_2.
- **partOf**: This relationship indicates that entity E_1 is a component of entity E_2.
- **evaluatedBy**: This relationship indicates that the NLP task T is typically evaluated using metric M
- **evaluatedOn**: This relationship indicates that the NLP task T is typically evaluated on dataset D.
- **similarWith**: This relationship indicates that the two entities share the same or highly similar meanings, such as abbreviations and full names of certain met

A.3.2 Annotation Note

- **Do not annotate negative relationships**.For example, relationships such as "X is not used in Y" or "X is difficult to apply in Y" should not be annotated.
- **Verify that entities involved in a relationship match the specified types** (e.g., Method-Dataset for TrainedWith). Do not link incorrect entity types through these predefined relationships.
- **Annotate relationships only if there is direct textual evidence or explicit implication**. Avoid inferring relationships that are not clearly stated or strongly suggested in the text.
- **Ensure consistency in relationship annotations across documents**. If uncertain about a relationship, refrain from annotating it.
- **Do not make assumptions about relationships based on personal knowledge or external information**. Rely solely on information explicitly provided in the text.
- **Annotate all identifiable relationships inferred from context, regardless of their quantity**. If contextual clues strongly imply a relationship, annotate it even if mentioned briefly.

Examlpe	Sentence	Entity type	Annotation result
1	In this paper, we employ the bounded L-BFGS (Behm et al., 2009) algorithm for the optimization task, which works well even when the number of weights is large.	Model	L-BFGS (Behm et al., 2009) algorithm
2	The evaluation method is the case insensitive IBM BLEU-4 (Papineni et al., 2002).	Metrics	BLEU-4
3	Our IWSLT data is the IWSLT 2009 dialog task data set.	Dataset	IWSLT data; IWSLT 2009 dialog task data set
4	We used four datasets: IMDB, Elec, RCV1, and 20-newsgrous (20NG) to facilitate direct comparison with DL15.	Dataset	20-newsgrous (20NG)
5	Escudero et al.(2000)used the DSO corpus to highlight the importance of the issue of domain dependence of WSD systems, but did not propose methods such as active learning or countmerging to address the specific problem of how to perform domain adaptation for WSD.	Metrics	active learning, countmerging
6	We adapt a Kneser-Ney language model to incorporate such growing discounts, resulting in perplexity improvements over modified Kneser-Ney and Jelinek-Mercer baselines.	Model	Kneser-Ney language model; modified Kneser-Ney; JelinekMercer
7	Techniques commonly applied to automatic text indexation can be applied to the automatic transcriptions of the broadcast news radio and TV documents.	Task	text indexation

Table 6: Examples for the entity annotation

Relation type	Examples	Example sentences
subtaskOf; subtaskOf	Document-level emotion classification→classification ; sentence-level emotion classification→classification	Document-level emotion classification, sentence-level emotion classification is naturally a multi-label classification problem.
usedFor; usedFor	SEXTANT→context extraction; SEXTANT→grammatical relation extraction;	Curran and Moens compared several context extraction methods and found that the shallow pipeline and grammatical relation extraction used in SEXTANT was both extremely fast and produced high-quality results.
usedFor; usedFor; compareWith	Seq2Seq model→SQG task ; Seq2Seq model→SQG task ; SQG task—TQG task	Note that if we directly compare the performance between SQG task and TQG task under the same model (e.g., the Seq2Seq model).

Table 7: Examples of the relations annotation