Computer Science Articles Named Entity Recognition Datasets: Survey and Our Recent Development

Anonymous ACL submission

Abstract

Domain-specific named entity recognition on Computer Science (CS) scholarly articles is an information extraction task that is arguably more challenging and less studied than named entity recognition (NER) for the general domain. Given that significant progress has been made on NER, we believe that scholarly domain-specific NER will receive increasing attention in the NLP community. Nevertheless, progress on the task is currently hampered in part by its recency and the lack of standardized concept types for scientific entities/terms. This paper presents a survey of the current state of research on scholarly domain-specific NER with a focus on language resources; further, it creates a novel dataset and model for CS NER.

1 Introduction

Named entity recognition over Computer Science scholarly articles (CS NER) is an information extraction task that involves identifying and classifying scientific terminology from CS scholarly publications including articles, books, patent documents, etc. as predefined semantic concept types. To better understand the task, consider the sentence:

> Exploiting Headword Dependency and Predictive Clustering for Language Modeling

Taken from an existing language resource (Gupta and Manning, 2011), the sentence has the following scientific entity annotations. Namely, 'Headword Dependency' and 'Predictive Clustering' as *technique*; and 'Language Modeling' as *focus* and *domain*. Here the concept *technique* is expressed as method, *domain* as the research problem, and *focus* as the solution. Indeed, by the broadest definition of CS NER, the precise typing of CS entities is ambiguous since a term can have multiple conceptual roles in a single paper context or even different roles across papers.

CS NER is arguably more difficult than NER, the task of identifying and typing commonsense real-world entities such as person, location, organization, thing, or temporal information. For NER, there are well-defined linguistic constraints at the syntactic (e.g., proper noun part-of-speech for person or location names), semantic (e.g., common knowledge things such as 'bat' and 'ball' have generally agreed and unambiguous meanings), and grammatical (e.g., prepositions such as 'in' or 'at' as cues for temporal information) levels. In contrast, there are typically no clear syntactic or other surface clues for identifying CS entities. While the CL-Titles parser system (D'Souza and Auer, 2021) relied on repetitive lexicosyntactic patterns in a rule-based approach to identify CS entities, they were heuristically based.

As significant advances have been made in NER, we believe that scholarly domain-specific NER will gain increasing attention in the years to come. This is owing to the digitalization of scholarly knowledge impetus (sci; Manghi et al., 2010; Lewis et al., 2016; Auer, 2018). Semantically modeling fine-grained actionable scholarly knowledge will make their large-scale in-silico rapid surveying a new paradigm shift in scholarly digital technologies. Instead of manual human comprehension of the latest and greatest scholarly knowledge within expert silos, knowledge of discoveries can be routinely and centrally screened for information about past and novel discoveries. Further, "text mining" methods can help bridge the gap between the growing amounts of data and our continuing need for insight into their corresponding findings. Our goal in this paper is to provide a timely survey of the current state of research on scholarly domain-specific NER with a specific focus on the Computer Science (CS) domain, i.e. CS NER. Further, we release a language resource which combines our surveyistic insights on CS NER and an high-performing machine learning tool trained on our dataset.

Corpora	Domain	Coverage	Semantic Concepts	Size			Annotation
				Papers	Tokens	Entities	
FTD (2011)	CL	titles, abstracts	focus, domain, technique	426	57,182	5,382	human
ACL-RD TEC (2016)	CL	abstracts	language resource, language resource product, measures and measurements, models, other, technology and method, tool and library	300	32,758	4,391	human
ScienceIE (2017)	CS, MS, Phy	full text	material, process, task	500	83,753	10,994	human
SciERC (2018)	AI	abstracts	evaluation metric, generic, material, method, task	500	60,749	8,089	human
NLP-TDMS (2019)	CL	titles, abstracts, full text	task, dataset, metric, score	332	1,115,987	1,384	distant supervision
STEM-ECR (2020)	10 STEM	abstracts	data, material, method, process	110	26,269	6,165	human
SciREX (2020)	ML	titles, abstracts, full text	dataset, method, metric, task	438	248,7091	**156,931**	human
NCG (2021)	CL, CV	titles, abstracts	research problem	405	47,127	908	human
ORKG-TDM (2021)	AI	titles, abstracts, full text	task, dataset, metric	5,361	-	18,219	distant supervision
CL-Titles (2021)	CL	titles	language, method, research problem, resource, solution, tool	**50,237**	284,672	87,567	rule-based system
PwC (this paper)	AI	titles, abstracts	research problem, method	12,271	**1,317,256**	29,273	distant supervision
ACL (this paper)	CL	titles	language, method, research problem, resource, dataset, solution, tool	31,041	263,143	67,270	human

Table 1: Comparison of Computer Science papers centric corpora for named entity recognition (CS NER). The corpora names in bold are the corpora merged as part of the dataset of this work. Domain Acronyms. CL - Computational Linguistics; CS - Computer Science; MS - Material Science; Phy - Physics; AI - Artificial Intelligence; STEM - Science, Technology, Engineering, Medicine; ML - Machine Learning; CV - Computer Vision.

2 Definitions

The NER "named entity" recognition task, first defined in the MUC conferences (Grishman and Sundheim, 1996), basically involved identifying the names of all the people, organizations, and geographic locations in text. We perceive the CS NER task similarly, i.e. identifying all scientific entity names of relevant semantic concepts to CS scholarly articles. E.g., the entity "F1" of concept *metric*; or "SQuAD" as an entity of the *dataset* concept. In the past (2016), the word "term" has also been introduced and defined as a lexical unit carrying a specialised meaning in a particular context. This, we find, is analogous to a "named entity."

Over the years, the set of CS concepts have evolved w.r.t. the number of types, their label names, and the aspects of the paper that were annotated with the concepts. Table 1 shows a high-level overview of the existing datasets with their semantic concepts. Overall nine main concepts emerge (see Appendix B for their label mappings) which, inspired from related works (2016; 2019; 2021), are defined as follows. A *research problem* (*rp*) is the theme of a work; a *method* (*meth*) is an existing protocol to support the solution; a *solution* (*sol*) is a novel contribution of a work that solves the *rp*; a *tool* is found by asking the question "Using what?"; *resource* (*res*) refers to utilities like the Web, Encyclopedia, etc.; *dataset* is the name of a dataset; *language* (*lang*) is the natural language focus; a *metric* is the component of evaluation systems used for measuring; and *score* is the quantitative system performance number associated with a *metric*.

3 Survey of Scholarly NER Corpora

3.1 Computer Science NER (CS NER)

Table 1 shows existing CS NER corpora compared along five dimensions: (1) domain, (2) annotation coverage, (3) semantic concepts, (4) size, and (5) annotation method. Most of the corpora consist of relatively short documents. The shortest is the CL-Titles corpus (2021) with only paper titles. The longer ones have sentences from full-text articles, viz. ScienceIE (2017), NLP-TDMS (2019), SciREX (2020), and ORKG-TDM (2021). We see that the corpora have had from one (D'Souza et al., 2021) to atmost seven concepts (QasemiZadeh and Schumann, 2016). Each corpora' concepts purposefully informs an overarching knowledge extraction objective. E.g., the concepts *focus*, *technique*, and

domain in the FTD corpus (2011) helped examine the influence between research communities; ACL-RD-TEC (2016) made possible a broader trends analysis with seven concepts. Eventually, corpora began to shed light on a novel scientific community research direction toward representing the entities as knowledge graphs (Auer, 2018) with hierarchical relation annotations such as synonymy (2017) or semantic relations such '*Method Used-for* a *Task*' (2018); otherwise, concepts were combined within full-fledged semantic constructs as LEADERBOARDS with between three to four concepts (Hou et al., 2019; Jain et al., 2020; Mondal et al., 2021; Kabongo et al., 2021), viz. *rp*, *dataset*, *meth*, *metric*, and *score*; or were in extraction objectives with solely contributions-focused entities of a paper (Färber et al., 2021; D'Souza and Auer, 2021).

3.2 Biomedical NER (BioNER)

BioNER dates before CS NER. It aims to recognize concepts in bioscience and medicine. E.g., protein, gene, disease, drug, tissue, body part and location of activity such as cell or organism. The most frequently used corpora are GENETAG (full-text articles annotated with protein/gene entities) (Tanabe et al., 2005), JNLPBA (~2400 abstracts annotated with DNA, RNA, protein, cell type and cell line concepts) (Collier and Kim, 2004), GENIA (~200 Medline abstracts annotated with 36 different concepts from the Genia ontology and several levels of linguistic/semantic features) (Kim et al., 2003), NCBI disease corpus (793 abstracts annotated with diseases in the MeSH taxonomy) (Doğan et al., 2014), CRAFT (the second largest corpus with 97 full text papers annotated with over 4000 corpus) (Bada et al., 2012) linking to the NCBI Taxonomy, the Protein, Gene, Cell, Sequence ontologies etc. Finally, the MedMentions corpus (Mohan and Li, 2018) as the largest dataset with ~4000 abstracts with ~34,724 concepts from the UMLS ontology. By leveraging ontologies such as the Gene Ontology (Ashburner et al., 2000), UMLS (Bodenreider, 2004), MESH, or the NCBI Taxonomy (Schoch et al., 2020), for the semantic concepts, these corpora build on years of careful knowledge representation work and are semantically consistent with a wide variety of other efforts that exploit these community resources. This differs from CS NER which is evolving toward standardized concepts.

Structured knowledge as knowledge bases (KB) were early seen as necessary in organizing biomedical scientific findings. E.g., protein-protein (PPI) interaction databases as MINT (Chatr-Aryamontri et al., 2007) and IntAct (Kerrien et al., 2007) or the more detailed KBs as pathway (Bader et al., 2006) or Gene Ontology Annotation (Camon et al., 2004). Community challenges help curate these KBs via text mining at a large-scale. E.g., BioCreative for PPI (Krallinger et al., 2008, 2011), protein-mutation associations (Krallinger et al., 2009), and gene-disease relations (Krallinger et al., 2010); or BioNLP (Kim et al., 2011) for complex n-ary bio events. CS NER is also been addressed in equivalent series such as SemEval (2017; 2018; 2021) which is promising to foster rapid task progress.

3.3 Chemistry NER (ChemNER)

BioNER in part fosters Chemistry NER. Text mining for drug and chemical compound entities (Herrero-Zazo et al., 2013; Krallinger et al., 2015) are indispensable to mining chemical disease relations (Li et al., 2016), and drug and chemical-protein interactions (Krallinger et al., 2017, 2021). Obtaining this structured knowledge has implications in precision medicine, drug discovery as well as basic biomedical research. Corpora for ChemNER are Corbett et al.'s dataset (42 full-text papers with ~7000 chemical entities), ChemDNER (10,000 PubMed abstracts with 84,355 chemical entities) (2015), and NLM-Chem (150 full-text papers with 38,342 chemical entities normalized to 2,064 MeSH identifiers) (Islamaj et al., 2021).

4 Our Contributions-Focused Resource for CS NER

With surveyistic insights, we create a CS NER corpus with a specific IE aim, i.e. to encapsulate only the *results-focused* or, alternately, the *contributions-focused* entities of a work. This aim would further the state-of-the-art in CS NER. So far, only the LEADERBOARDS construct (2019; 2020) involving *rp*, *dataset*, *meth*, *metric*, *score* have enabled the generation of *progress overview knowledge graphs* of a field. We broaden this *results-focused entities mining* notion to other CS concepts where contributions are also approaches as in the *solution* concept. While similar concepts were annotated in other corpora, we differ with our entity selection to only the paper's results-focused entities for the concepts. SciREX (2020) and CitationIE (2021) adopt a similar "salient" entity perspective. They, however, consider a weighted citations graph for entity mentions

Types	P	R	$F1$
Method	66.8	49.13	56.62
Tool	72.01	66.05	68.9
Dataset	72.9	68.42	70.59
Research problem	68.24	79.68	73.52

Types	P	R	$F1$
Resource	75.72	78.61	77.14
Solution	78.51	82.61	80.51
Language	86.22	87.78	86.99
macro Overall	73.49	74.03	73.76

Table 2: CS Named entity recognition results on TITLES per-concept type

Types	P	R	$F1$
Research problem	81.18	75.81	78.4
Method	87.59	86.6	87.09
macro Overall	84.39	81.2	82.76

Table 3: CS NER results on ABSTRACTS per-concept type

TITLES			ABSTRACTS		
P	R	$F1$	P	R	$F1$
74.14	76.26	**75.18**	84.89	81.9	**83.37**
73.67	75.16	74.41	88.2	78.85	83.26

Table 4: Cumulative results for salient CS NER on seven concepts in TITLES and on two concepts in ABSTRACTS with GloVe embeddings (top row) and without (last row)

to determine saliency. This implies the method holds only given sufficient citations for the entity mentions. Our approach, then, is simpler where we do not deal with whether a results-focused entity is salient in the community, but merely that they are contributions-focused for each paper.

Aligned with our IE aim, we create two corpora: 1) TITLES - noting that contributions-focused entities of a paper are naturally present in titles; and 2) ABSTRACTS - for *rp* and *meth*. This latter corpus can offer a fallback for the two concepts if they are not in titles. A natural question may be why ABSTRACTS is annotated with only two of our seven entities? In this work, since our emphasis has been on reusing the existing corpora for their annotations and we found for Abstracts only *rp* and *meth* concepts satisfying our entities filter.

TITLES comprises four existing corpora. 1) ACL (last row in Table 1). This corpus was originally automatically generated as CL-Titles (2021) and includes all our concepts except *dataset*. We heuristically adapted it for *dataset* and manually verified its annotations for 31,044 of its 49,728 titles. Thus our version includes all seven of our results-focused entities. While the corpus was verified by a single annotator, we performed an IAA exercise for 50 titles involving the main annotator (a NLP Postdoc) and a secondary "outsider" annotator (a NLP PhD candidate). They had a strong IAA of 71.52% Cohen's κ. 2) PwC (second-to-last row in Table 1). It includes distant-labeled titles from `https://paperswithcode.com/` for *rp* and *meth*. Note that NLP-TDMS (2019), SciREX (2020), and ORKG-TDMS (2021) are its subsets. 3) FTD corpus (2011) for *rp*, *meth*, and *sol*. And 4) NCG (2021) for *rp* entities. The distributions are 31,041 (82%) ACL/5,885 (15%) PwC/462 (1%) FTD/398 (1%) NCG. The sizes of the FTD and NCG are the original dataset sizes. PwC was a strategically randomly selected subset which offered sufficient annotation diversity for *rp* and *meth* without biasing an automatic system to just these two types. Next, the ABSTRACTS corpus also combines four existing corpora. 1) PwC for *rp* and *meth*. 2) FTD for *rp* and *meth*. 3) NCG for *ro*. 4) SciERC (2018) for *rp*. Their distributions are 6756 (85%) PwC/462 (5%) FTD/272 (3%) NCG/431 (5%) SciERC. While only PwC was a strategically chosen subset for being representative of the two entities, the other corpora were included as is.

A Strong Baseline Model. We train TITLES CS NER and ABSTRACTS CS NER as two separate IOBES sequence tagging models. It is the state-of-the-art 3-layered model: a character sequence layer with a CNN encoder (CCNN), a word sequence layer with a BiLSTM encoder (WBiLSTM), and a CRF inference layer (CRF) (Yang and Zhang, 2018). Words in word sequences are represented as embeddings which are initialized either as pre-computed (we use GloVe (Pennington et al., 2014)) or at random. The character sequence layer automatically extracts word level features by encoding the character sequence within the word and is randomly initialized. Our results are shown in Tables 2, 3, and 4. We thus find optimal high-scores of 75.18% over seven concepts in TITLES and 83.37% over two concepts in ABSTRACTS in micro F1. In TITLES, *language* is the easiest to extract at 86.99% F1, and in ABTRACTS, it is *method* at 87.09% F1.

5 Conclusion

We reported a focused result for contributions-focused CS NER. Our work is in the broader context of existing work by conducting a multi-disciplinary corpus survey and shows how we merge existing CS NER corpora tailored to our IE aim. Our data and code is publicly released.

References

SciGraph. https://www.springernature.com/de/researchers/scigraph. Accessed: 2021-11-02.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Sören Auer. 2018. Towards an open research knowledge graph.

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.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. 2012. Concept annotation in the craft corpus. *BMC bioinformatics*, 13(1):1–20.

Gary D Bader, Michael P Cary, and Chris Sander. 2006. Pathguide: a pathway resource list. *Nucleic acids research*, 34(suppl_1):D504–D506.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267.

Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. 2004. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Research*, 32(Database issue):D262.

Andrew Chatr-Aryamontri, Arnaud Ceol, Luisa Montecchi Palazzi, Giuliano Nardelli, Maria Victoria Schneider, Luisa Castagnoli, and Gianni Cesareni. 2007. Mint: the molecular interaction database. *Nucleic acids research*, 35(suppl_1):D572–D574.

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

Peter Corbett, Colin Batchelor, and Simone Teufel. 2007. Annotation of chemical named entities. In *Biological, translational, and clinical language processing*, pages 57–64.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Jennifer D'Souza and Soeren Auer. 2021. Pattern-based acquisition of scientific entities from scholarly article titles. *arXiv preprint arXiv:2109.00199*.

Jennifer D'Souza, Sören Auer, and Ted Pedersen. 2021. SemEval-2021 task 11: NLPContributionGraph - structuring scholarly NLP contributions for a research knowledge graph. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 364–376, Online. Association for Computational Linguistics.

Jennifer D'Souza, Anett Hoppe, Arthur Brack, Mohmad Yaser Jaradeh, Sören Auer, and Ralph Ewerth. 2020. The STEM-ECR dataset: Grounding scientific entity references in STEM scholarly content to authoritative encyclopedic and lexicographic sources. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2192–2203, Marseille, France. European Language Resources Association.

Michael Färber, Alexander Albers, and Felix Schüber. 2021. Identifying used methods and datasets in scientific publications. In *Proceedings of the Workshop on Scientific Document Understanding: co-located with 35th AAAI Conference on Artificial Inteligence (AAAI 2021) ; Remote, February 9, 2021. Ed.: A. P. B. Veyseh*, volume 2831 of *CEUR Workshop Proceedings*. CEUR Workshop Proceedings (CEUR-WS).

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.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Sonal Gupta and Christopher Manning. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1–9, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of tasks, datasets, evaluation metrics, and numeric

scores for scientific leaderboards construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213, Florence, Italy. Association for Computational Linguistics.

Rezarta Islamaj, Robert Leaman, Sun Kim, Dongseop Kwon, Chih-Hsuan Wei, Donald C Comeau, Yifan Peng, David Cissel, Cathleen Coss, Carol Fisher, et al. 2021. Nlm-chem, a new resource for chemical entity recognition in pubmed full text literature. *Scientific Data*, 8(1):1–12.

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.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2021. Automated mining of leaderboards for empirical ai research. *arXiv preprint arXiv:2109.13089*.

Samuel Kerrien, Yasmin Alam-Faruque, Bruno Aranda, Iain Bancarz, Alan Bridge, Cathy Derow, Emily Dimmer, Marc Feuermann, Anja Friedrichsen, Rachael Huntley, et al. 2007. Intact—open source resource for molecular interaction data. *Nucleic acids research*, 35(suppl_1):D561–D565.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2011. Extracting bio-molecular events from literature—the bionlp'09 shared task. *Computational Intelligence*, 27(4):513–540.

Martin Krallinger, Jose MG Izarzugaza, Carlos Rodriguez-Penagos, and Alfonso Valencia. 2009. Extraction of human kinase mutations from literature, databases and genotyping studies. *BMC bioinformatics*, 10(8):1–20.

Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of biocreative ii. *Genome biology*, 9(2):1–19.

Martin Krallinger, Florian Leitner, and Alfonso Valencia. 2010. Analysis of biological processes and diseases using text mining approaches. *Bioinformatics Methods in Clinical Research*, pages 341–382.

Martin Krallinger, Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, and Alfonso Valencia. 2021. Drugprot shared task (biocreative vii track 1-2021) text mining drug-protein/gene interactions (drugprot) shared task.

Martin Krallinger, Obdulia Rabal, Saber Ahmad Akhondi, Martín Pérez Pérez, Jesus Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio Baso López, Umesh K. Nandal, Erin M. van Buel, Anjana Chandrasekhar, Marleen Rodenburg, Astrid Lægreid, Marius A. Doornenbal, Julen Oyarzábal, Anália Lourenço, and Alfonso Valencia. 2017. Overview of the biocreative vi chemical-protein interaction track.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

Martin Krallinger, Miguel Vazquez, Florian Leitner, David Salgado, Andrew Chatr-Aryamontri, Andrew Winter, Livia Perfetto, Leonardo Briganti, Luana Licata, Marta Iannuccelli, et al. 2011. The protein-protein interaction tasks of biocreative iii: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC bioinformatics*, 12(8):1–31.

Nathaniel Lewis, Jingbo Wang, Marta Poblet, and Amir Aryani. 2016. Research graph: Connecting researchers, research data, publications and grants using the graph technology. In *eResearch Australasia Conference*.

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.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreferencefor scientific knowledge graph construction. In *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*.

Paolo Manghi, Natalia Manola, Wolfram Horstmann, and Dale Peters. 2010. An infrastructure for managing ec funded research output: The openaire project. *Grey Journal (TGJ)*, 6(1).

Sunil Mohan and Donghui Li. 2018. Medmentions: A large biomedical corpus annotated with umls concepts. In *Automated Knowledge Base Construction (AKBC)*.

Ishani Mondal, Yufang Hou, and Charles Jochim. 2021. End-to-end construction of NLP knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1885–1895, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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).

Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P Sullivan, Lu Sun, Seán Turner, and Ilene Karsch-Mizrachi. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020. Baaa062.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. 2005. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6(1):1–7.

Vijay Viswanathan, Graham Neubig, and Pengfei Liu. 2021. CitationIE: Leveraging the citation graph for scientific information extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 719–731, Online. Association for Computational Linguistics.

Jie Yang and Yue Zhang. 2018. NCRF++: An open-source neural sequence labeling toolkit. In *Proceedings of ACL 2018, System Demonstrations*, pages 74–79, Melbourne, Australia. Association for Computational Linguistics.

A Inter-Annotator Agreement Scores Per-Concept

Here we report agreement scores with alternate metrics as precision, recall, F1. Additionally, detailed agreements per concept type is shown.

Type	P	R	$F1$
Tool	25	16.67	20
Method	52.17	85.71	64.86
Resource	73.33	61.11	66.67
Dataset	100	50	66.67
Research problem	62.96	77.27	69.39
Solution	86.49	71.11	78.05
Language	100	100	100
TOTAL			69

Table 5: Interannotator agreement scores on 50 titles

B Mappings between concepts for CS NER

In the fourth "Semantic Concepts" column in Table 1 in the main paper reports the original dataset concepts labels. In Table 6, we show how different labels names can be mapped as one standard name since they have the same semantic definitions. For our language resource section 4 reported in the main paper, we adopt the standard names.

	Types	Mappings in Related Work
1	*research-problem*	domain; application; task; research problem
2	*method*	technique; technology and method; method
3	*solution*	focus; solution
4	*tool*	tool and library; tool
5	*resource*	language resource; resource
6	*dataset*	language resource product; dataset
7	*language*	language
8	metric	measures and measurements; evaluation metric; metric
9	score	measures and measurements

Table 6: Mappings of nine scientific semantic types across Computer Science papers centric corpora for CS NER. The italicized types are in the dataset of this work.

C Detailed Baseline Model Ablations

Neural Architectures	micro P	micro R	micro $F1$
word CNN + CRF	70.28	71.24	70.76
	69.32	69.16	69.24
word LSTM + CRF	69.24	70.08	69.65
	68.41	66.76	67.58
word BiLSTM + CRF	71.92	73.34	72.62
	71.44	72.91	72.17
word CNN + char CNN + CRF	71.31	72.96	72.13
	72.50	71.01	71.75
word LSTM + char CNN + CRF	72.01	72.4	72.21
	71.59	69.65	70.61
word BiLSTM + char CNN + CRF	**74.14**	**76.26**	**75.18**
	73.67	75.16	74.41

Table 7: Results with different neural architectures for CS NER over seven semantic concepts with embeddings (top row) and without (bottom row) on TITLES.

Neural Architectures	micro P	micro R	micro $F1$
word CNN + CRF	90.55	72.51	80.53
	91.78	73.58	81.68
word LSTM + CRF	85.45	75.54	79.62
	90.02	71.82	79.9
word BiLSTM + CRF	88.22	76.24	81.79
	90.14	76.36	82.68
word CNN + char CNN + CRF	78.61	71.08	74.65
	88.59	66.33	75.86
word LSTM + char CNN + CRF	85.48	78	81.57
	87.71	76.49	81.71
word BiLSTM + char CNN + CRF	**84.89**	**81.9**	**83.37**
	88.2	78.85	83.26

Table 8: Results with different neural architectures for CS NER over two semantic concepts with embeddings (top row) and without (bottom row) on ABSTRACTS.