Computer Science Named Entity Recognition in the Open Research Knowledge Graph*

Jennifer D'Souza[0000−0002−6616−9509] and Sören Auer[0000−0002−0698−2864]

TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{jennifer.dsouza,auer}@tib.eu

Abstract. Domain-specific named entity recognition (NER) on Computer Science (CS) scholarly articles is an information extraction task that is arguably more challenging for the various annotation aims that can hamper the task and has been less studied than NER in the general domain. Given that significant progress has been made on NER, we anticipate that scholarly domain-specific NER will receive increasing attention in the years to come. Currently, progress on CS NER – the focus of this work – is hampered in part by its recency and the lack of a standardized annotation aim for scientific entities/terms. This work proposes a standardized task by defining a set of seven *contribution-centric scholarly entities* for CS NER viz., *research problem*, *solution*, *resource*, *language*, *tool*, *method*, and *dataset*. The main contributions are: (1) combining existing CS NER resources that maintain their annotation focus on the set or subset of *contribution-centric scholarly entities* we consider; (2) noting the need for big data to train neural NER models, this work additionally supplies thousands of *contribution-centric entity* annotations from article titles and abstracts, thus releasing a cumulative large novel resource for CS NER; and, (3) training a sequence labeling CS NER model inspired by state-of-the-art neural architectures from the general domain NER task. Throughout the work, several practical considerations are discussed aiming to guide information technology designers of digital libraries.

Keywords: Named entity recognition · Information extraction · Neural sequence labeling.

1 Introduction

Named entity recognition (NER) is an essential Natural Language Processing (NLP) function for *the lifting of entities-of-interest from unstructured text.* NER powers contemporary knowledge-graph based search engines as demonstrated in industry e.g. by Facebook [83] and Google [2]; but also the open data community with Wikidata [92]. NER has proven indispensable to machine readers of unstructured texts of common or general knowledge. Commonsense machine reading is

* Supported by TIB Leibniz Information Centre for Science and Technology, the EU H2020 ERC project ScienceGRaph (GA ID: 819536)

an area where significant progress can be tracked via state-of-the-art systems such as Babelfy [81], DBpedia Spotlight [75], NELL [77], and FRED [39], to name a few. However, the same cannot be said for all domains of discourse text. Consider scholarly literature, as an exemplar, which remains relatively understudied in terms of advanced information retrieval applications that go beyond keywords toward content-based entity-centric machine readers. In the scholarly domain, obtaining fine-grained entity-centric knowledge facilitated by well-established NER systems is not yet feasible. As significant advances have been made on NER in the general domain, 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 via crowdsourcing that is growing [1,72,66,5,7,12,93,42]. While expert-based crowdsourcing is effective to obtain high-quality data, it is not necessarily a scalable solution in the face of growing volumes of scientific literature [14], the processing of which would need the support of automated NLP techniques, one among which is NER. Obtaining the critical digitalized data mass via scalable methods warrant the paradigm shift toward the standardized adoption of digitalized scholarly knowledge. This data representation is advantageous for several reasons, mainly by its meaningful structured connections across entity-centric research progress, research redundancy [46] can be readily alleviated – a problem predominant in document-based silos of research records where core conceptual entities, buried within text volumes, need to be manually uncovered through human comprehension. Thus, instead of manual human comprehension of the latest and greatest scholarly knowledge within expert silos, digitalized scholarly knowledge can be routinely and centrally screened for information about past and novel discoveries.

Notably, next-generation scholarly digital library (DL) infrastructures are already emerging. The Open Research Knowledge Graph (ORKG) [4] digital research and innovation infrastructure, argues for obtaining a semantically rich, interlinked KG representations of the "content" of the scholarly articles, specifically, focused on *research contributions*.[1] With intelligent analytics enabled over such contributions-focused KGs, researchers can track research progress without the cognitive overhead that reading dozens of articles imposes. Via expert crowdsourcing, the information can be readily structured based on human judgements. However, the issue of needing an automated NLP system as a scalable complementary assistance technique remains; one that could even serve the purpose of making it easier for experts to structure scholarly knowledge via drag-and-drop recommendations. A typical dilemma then with leveraging automated NLP for the ORKG, specifically, w.r.t. implementing an NER module is deciding the scholarly/scientific entity extraction targets. In other words, aligned with the ORKG objective of structuring research contributions, the key challenge is: *How to select only the contribution-centric entities and what would be their types*? While this question has a broad scope across Science, in this paper, we tackle the scholarly *contribution-centric* NER problem for Computer Science (CS).

[1] The ORKG platform can be accessed online: `https://orkg.org/`.

We define *contribution-centric* CS NER as involving the identification of a word or a phrase as an entity from Computer Science scholarly articles, either from its title, or abstract, or full-text or one or more or all places, which satisfies one of the following seven types, viz. *research problem*, *solution*, *resource*, *language*, *tool*, *method*, and *dataset*. Furthermore, since, in the context of this work, CS NER is qualified as being *contribution-centric*, only those entities that are either the outcome of a particular work or used to support the outcome of that work are candidate extraction targets. Related to the model we propose is the Task, Dataset, Metric (TDM) model attemped in several works [50,79,45] originally inspired from `https://paperswithcode.com/`. However, the TDM model targets only the automated construction of Leaderboards in the domain of Artificial Intelligence. In this work, we try to generically define a broader set of *contribution-centric* entities which also naturally subsumes the TDM model.

Summarily, the contributions of this paper are the following. 1) Existing CS NER language resources are examined and the problem of the lack of standardized entities therein is clarified. 2) A standardized set of entities is elicited by proposing a standardization on semantic types for extraction that satisfy the aim of being *contribution-centric* extraction targets for CS. 3) Existing resources that fulfill the aim of *contribution-centric* extraction targets are combined and, further still, additional data is annotated resulting in a large corpus which is made publicly available. 4) Finally, based on empirical evaluations from six different state-of-the-art neural architectures an automated CS NER system is created. The best model performances obtained are between 75% and 85% on the task. Our empirical analysis could serve informaticians working in the ecosphere of the contemporary digital libraries. Our system developed for the Open Research Knowledge Graph [4] is also a community release at `https://tinyurl.com/orkg-cs-ner`; including its underlying dataset `https://github.com/jd-coderepos/contributions-ner-cs`.

2 Background

NER is a long-standing task in the NLP community backed by over 3 decades of NLP research tracing it to the Message Understanding Conference series [40] where the name was first coined. In contrast, the earliest work on CS NER is relatively recent in 2015 [41]. Since MUC [40], NER has seen a broad flurry of activity with numerous scientific events among which are CONLL [90,91] and ACE [30]. This was also complemented by steady research. NER systems have been variously implemented including handcrafted rules, bootstrapping approaches relying on seed extraction patterns [23,35]; as feature-engineered machine learning approaches based on hidden markov models, support vector machines, conditional random fields, and decision trees using features such as orthographic, prefixes suffixes, labels, neighboring words, etc [71,73,16,68,87]. Starting with [24], neural network NER systems became popular. The systems differ based on representations relying on words and sub-word units viz. word level architectures [24,17], character level architectures [54], character+word level architectures [20,70], and character+word+affix models [27,94]. This mature field is also rife with various tools such as GATE (General Architecture for Text Engineering) [26], Mallet (Machine learning for language toolkit) [74], Natural Language Toolkit (Suite of Python libraries for NLP) [11], DBpedia Spotlight [76], Stanford NER [37], and SpaCy [44]. In contrast, CS NER has only recently started gaining traction following being organized as a task in a SemEval series having then garnered interest in the broader community [6,38,69,45,28,48,33].

3 Definitions

The NER "named entity" task from the MUC conferences [40] identified the names of all the people, organizations, and geographic locations in a text. Analogously, we define the CS NER task as that of identifying scientific entities of specific semantic types from CS scholarly articles. E.g., the entity "F1" of type *metric*; or the entity "SQuAD" as an instance of the *dataset* semantic type. In the past [86], the word "term" was introduced as a lexical unit carrying a specialised meaning in a particular context. This, we understand, is synonymous to a "named entity."

Over the years, the CS entity semantic types have evolved w.r.t. the number of types, the type labels, and the extraction target aim they served. Table 1 shows a high-level overview of the existing semantic types in the context of their datasets. Following which, listed in Table 2 are the nine main semantic types that have emerged, viz. *research problem*, *method*, *solution*, *tool*, *resource*, *dataset*, *language*, *metric*, and *score*. These tables are discussed in detail in Section 4.3. These semantic types are defined as follows [86,45,32]. A *research problem* is the theme of the investigation; a *method* is existing protocols used to support the solution; a *solution* is a novel contribution of a work that solves the *research problem*; entities of type *tool* are found by asking the question "Using what?" or "By which means?;" *resource* are names of existing data and other references to utilities like the Web, Encyclopedia, etc., used to address the *research problem* or used in the *solution*; *dataset* refers to the name of a dataset; *language* is the natural language focus of a work; a *metric* is the component of evaluation systems used for measuring and measurement processes; and *score* is the system performance number associated with an evaluation *metric*.

4 Related Work — Corpora

This section provides an overview of existing corpora used for scholarly NER research organized by three main research fields.[2] While this paper only addresses CS NER, in this section the Biomedical and Chemistry domains are undertaken in a related work context to offer a broader perspective on the NER problem.

[2] Please note access to some datasets may be possible only by contacting the authors.

4.1 Biomedical NER (BioNER)

Biomedical NER aims to identify and classify technical terms in the domain of molecular biology for instances of concepts of interest in bioscience and medicine. Examples of such concepts include protein, gene, disease, drug, tissue, body part and location of activity such as cell or organism. Concerning annotated gold-standard corpora to facilitate machine learning, the most frequently used corpora are GENETAG (composed of full text articles annotated with protein/gene mentions) [89], JNLPBA (composed of 2400 abstracts annotated with DNA, RNA, protein, cell type and cell line concepts) [22], GENIA (composed of 200 Medline abstracts annotated with 36 different concepts from the Genia ontology and several levels of linguistic/semantic features) [52], NCBI disease corpus (composed of 793 abstracts annotated with diseases in the MeSH taxonomy) [31], CRAFT (the second largest corpus consisting of 97 full text papers annotated with over 4000 concepts) [8] linking to the NCBI Taxonomy, the Protein, Gene, Cell, Sequence ontologies etc. Finally, the MedMentions corpus [78] posits itself as the largest language resource with over 4000 abstracts annotated with around 34,724 concepts from the UMLS ontology. By leveraging ontologies such as the Gene Ontology [3], UMLS [13], MESH,[3] or the NCBI Taxonomy [88], for scientific concept annotations, 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.

An evolutionary perspective on BioNER shows following endeavors that aimed to establish meaningful semantic structural relations between the bio-molecules such as protein-protein interactions (PPI) [57,62], protein-mutation associations [56], and gene-disease relations [58] that were the focus of the BioCreative challenge series; or more complex n-ary bio-molecular events which were addressed in the BioNLP challenge [53]. These attempts help curate large-scale knowledge by mining large volumes of articles for PPI databases as MINT [18] and IntAct [51] or the more detailed knowledge databases such as pathway [9] or Gene Ontology Annotation (GOA) [15]. Further, the ability to construct such connections between bio-molecules also provide a global view on different biological entities and their interactions, such as disease, genes, food, drugs, side effects, pathways, and toxins, opening new routes of research. A more comprehensive survey is available here `https://orkg.org/comparison/R163265/`.

4.2 Chemistry NER (ChemNER)

BioNER in part fosters Chemistry NER (ChemNER). ChemNER has evolved from extracting drugs [43], to chemicals [61], chemical disease relations [67], and drug and chemical-protein interactions [60,59]. Text mining for drug and chemical compound entities [43,61] are indispensable to mining chemical disease relations [67], and drug and chemical-protein interactions [60,59]. Obtaining this structured knowledge has implications in precision medicine, drug discovery as

[3] `https://www.nlm.nih.gov/mesh/meshhome.html`

well as basic biomedical research. Corpora for ChemNER are [25]'s dataset (42 full-text papers with ~7000 chemical entities), ChemDNER (10,000 PubMed abstracts with 84,355 chemical entities) [61], and NLM-Chem (150 full-text papers with 38,342 chemical entities normalized to 2,064 MeSH identifiers) [47].

Increasingly, text mining initiatives are seeking out recipes or formulaic semantic patterns to automatically mine machine-actionable information from scholarly articles [55,63,82,64]. In [63], wet lab protocols are annotated, covering a large spectrum of experimental biology, including neurology, epigenetics, metabolomics, cancer and stem cell biology, with actions corresponding to lab procedures and their attributes including materials, instruments and devices. Thereby the protocols then constituted a prespecified machine-readable format as opposed to the ad-hoc documentation norm. Kulkarni et al. [63] even release a large human-annotated corpus of semantified wet lab protocols to facilitate machine learning of such shallow semantic parsing over natural language instructions. Within scholarly articles, such instructions are typically published in the Materials and Method section of papers from Biology and Chemistry. [55,82] present semantically structured material synthesis procedures capturing synthesis operations (i.e. predicates), and the materials, conditions, apparatus and other entities participating in each synthesis step. These initiatives lend further evidence to the evolution of digital library technologies toward digitalization.

4.3 Computer Science NER (CS NER)

Table 1 shows existing CS NER corpora compared along five dimensions: (1) domain — the CS subarea(s) covered in the annotated data, (2) annotation coverage — the aspect of the paper annotated, (3) scientific entity semantic types — the semantic types assigned to the entities, (4) size — the number of papers, tokens, entities in the respective corpora, and (5) annotation method — the method by which the annotations were obtained. To offer a few summary observations. Most of the corpora consist of relatively short documents. The shortest is the CL-Titles corpus [32] with only paper titles. The longer ones have sentences from full-text articles, viz. ScienceIE [6], NLP-TDMS [45], SciREX [48], and ORKG-TDM [50]. We see that the corpora have had from one (e.g., the NCG corpus [33]) to atmost seven entity types (e.g., ACL-RD-TEC [86]). Each corpora' types purposefully informs an overarching knowledge extraction objective. E.g., the *focus*, *technique*, and *domain* entity types in the FTD corpus [41] helped examine the influence between research communities; ACL-RD-TEC [86] made possible a broader trends analysis with seven types. Eventually, corpora began to shed light on a novel scientific community research direction toward representing the entities as knowledge graphs [5] with hierarchical relation annotations such as synonymy [6] or semantic relations such '*Method Used-for* a *Task*' [69]; otherwise, scientific types were combined within full-fledged semantic constructs as LEADERBOARDS with between three to four concepts [45,48,80,50], viz. *research problem*, *dataset*, *method*, *metric*, and *score*; or were in extraction objectives with solely *contribution-centric* entities of a paper [36,32]. Overall, the corpora served two main information extraction aims. 1) Some resources offered all mentions of CS relevant scientific entities: ACL-RD-TEC [86], ScienceIE [6], and STEM-ECR [34]; and 2) While others offered only *contribution-centric* entities: FTD [41], SciERC [69], NLP-TDMS [45], SciREX [48], NCG [33], ORKG-TDM [50], and CL-Titles [32].

5 Our Corpus

To build our corpus for *contribution-centric* information extraction targets of scientific entities, we aimed: 1) to reuse existing resources for their entity annotations that already fulfill our extraction target aim (described further in Section 5.1); and 2) to append additional annotations to create a larger corpus for neural machine learning system development (described in Section 5.2).

5.1 Combining Existing Resources

This step first entailed normalizing different semantic label names with the same semantic definitions as one standard name. The mappings we used are elicited in Table 2. The table lists nine main semantic types whose semantic interpretations or definitions were offered in Section 3. After obtaining the label names mappings as semantic type normalizations, we selected only those corpora, and specifically the semantic types within the corpora, that satisfied our CS NER *contribution-centric* entities aim. Overall, our corpus was organized as aggregations of similar parts of the scholarly article. Thus, article titles constitute one corpus called the **Titles** corpus and article abstracts constitute a second corpus called the **Abstracts** corpus. Next we describe how some of the existing resources could be reused and combined to form the two respective corpora.

The **Titles** corpus combines annotations from two different corpora: 1) the FTD corpus [41] (row 1 in Table 1) for all three of its entities, viz. *research problem*, *method*, and *solution* entities. In all, 462 titles could be obtained from the FTD corpus which originally also includes exactly 462 total annotated paper titles and abstracts with one or more of the three entities' annotations. And, 2) NCG [33] (row 8 in Table 1) for its *research problem* entities. In all, 398 titles were obtained from NCG with *research problem* annotations which had a total of approximately 450 papers [33]. Thus the data from these two corpora were merged as the **Titles** corpus finally containing three entities, viz. *research problem*, *method*, and *solution* deemed by their original corpora respective annotation aims as *contribution-centric*, in turn fulfilling our CS NER aim.

The second corpus, i.e. the **Abstracts** corpus combines: 1) the FTD corpus paper abstracts annotated with *research problem* and *method* entities. Since no annotations for *solution* entities could be obtained in the abstracts, this type could not be included. As such, abstracts in all 462 of the FTD annotated papers were included in our corpus. Next, 2) 272 abstracts from the NCG corpus with annotations for *research problem*. And, lastly, 3) the SciERC corpus [69] (row 4 in Table 1) annotated abstracts for its *contribution-centric research problem* entity. 431 of its 500 total annotated papers could be combined. Note that SciERC had

Table 1. Comparison of Computer Science papers 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. **Size** column name acronyms: *P* - papers; *T* - tokens; *E* - entities. A detailed version of this table is available here https://orkg.org/comparison/R150058.

Corpora	**Domain**	**Coverage**	**Entity Semantic Types**	**Size**		
				P	*T*	*E*
FTD [41]	CL	titles, abstracts	focus, domain, technique	426	57,182	5,382
ACL-RD TEC [86]	CL	abstracts	language resource, language resource product, measures and measurements, models, other, technology and method, tool and library	300	32,758	4,391
ScienceIE [6]	CS, MS, Phy	full text	material, process, task	500	83,753	10,994
SciERC [69]	AI	abstracts	evaluation metric, generic, material, method, task	500	60,749	8,089
NLP-TDMS [45]	CL	titles, abstracts, full text	task, dataset, metric, score	332	1,115,987	1,384
STEM-ECR [34]	10 STEM	abstracts	data, material, method, process	110	26,269	6,165
SciREX [48]	ML	titles, abstracts, full text	dataset, method, metric, task	438	248,7091	**156,931**
NCG [33]	CL, CV	titles, abstracts	research problem	405	47,127	908
ORKG-TDM [50]	AI	titles, abstracts, full text	task, dataset, metric	5,361	-	18,219
CL-Titles [32]	CL	titles	language, method, research problem, resource, solution, tool	**50,237**	284,672	87,567
PwC (this paper)	AI	titles, abstracts	research problem, method	12,271	**1,317,256**	29,273
ACL (this paper)	CL	titles	language, method, research problem, resource, dataset, solution, tool	31,044	263,143	67,270

annotations for additional semantic types as well, e.g., *generic*, *material*, and *method*. However, these annotations could not be included since they did not satisfy our *contribution-centric* entities inclusion criteria.

Table 2. Mappings of nine scientific semantic types across Computer Science papers for CS NER. The first seven italicized types are in the dataset of this work.

	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 3. Inter-annotator percentage agreement scores on 50 titles in the ACL corpus over six scientific semantic entity types

entity type	P	R	$F1$
solution	86.49	71.11	78.05
tool	25	16.67	20
dataset	100	50	66.67
language	100	100	100
method	52.17	85.71	64.86
research problem	62.96	77.27	69.39
TOTAL			69

5.2 Our Annotated Data

ACL This corpus of Computational Linguistics paper titles was originally released as part of the CL-Titles parser software resources [32] and was automatically annotated using the rule-based CL-Titles parser. It included all the titles in the ACL Anthology at a specific download dump timestamp (`https://aclanthology.org/anthology.bib.gz`). This corpus was re-reviewed in this work for annotation quality and additional scientific semantic types that should be included. As such we noted that the semantic type *dataset* relevant particularly in the domain of Computational Linguistics was not originally included in the annotated types. We heuristically modified the annotations to include the *dataset* semantic type additionally and we further manually verified as many of the annotations as were possible in a fixed time-frame of 2 weeks. In this time, 31,041 of its 49,728 titles could be verified and amended for incorrect annotations. Thus the new verified and adapted version (which we call simply ACL) includes seven *contribution-centric* entities, viz. *language*, *method*, *research problem*, *resource*, *dataset*, *solution*, and *tool*. See last row in Table 1 for details.

While the corpus verification exercise was done by a single primary annotator (a NLP Postdoc), an IAA exercise for 50 randomly selected titles involving the primary annotator and a secondary "outsider" annotator (a NLP PhD candidate) was also conducted to gauge the replicability of the primary annotator's judgements. For this, a relatively straightforward process was followed. The primary annotator created definitions for the considered types with example titles. These were then shown to the "outsider" annotator and were available to him as reference material during the annotation task itself. He then proceeded with annotating a randomly selected set of 50 titles which were already annotated by the primary annotator but with the annotations unavailable to him. Thus his annotations of the 50 titles were performed in a blind protocol. Following the completion of the annotation task, inter-annotator agreement (IAA) scores were computed using the Cohen's κ [21] as well as standard F1 metrics. The results were promising. In terms of Cohen's κ, they had a strong IAA of 71.52%. Their IAA scores in terms of the F1 metric are shown in Table 3 with detailed agreement scores breakdown per semantic type. From the Table, we see that the annotators had perfect agreement over identifying entities of type *language*, with least agreement when identifying entities of type *tool*. This latter low agreement score could be ascribed to the second annotator preference of *method* vs. *tool*.

The 31,041 verified titles were appended to our **Titles** corpus.

PwC A second data source was leveraged from which additional annotations were appended to the **Titles** corpus and **Abstracts** corpus, respectively. The data was originally sourced from PapersWithCode (`https://paperswithcode.com/`) hence it is referred as **PwC**. Note that three of the datasets listed in Table 1, viz. NLP-TDMS [45], SciREX [48], and ORKG-TDMS [50], were indeed subsets of the PwC source. While their papers' subsets selected may have respectively varied, they all unanimously relied on annotation by distance labeling. PwC, itself, is a publicly available leaderboard of tasks in AI predominantly representing the AI NLP and Computer Vision research fields among others such as Knowledge Graph Embeddings, Robotics, etc. They release a public download dump of crowdsourced leaderboards in scholarly articles on research problems in AI annotated w.r.t. *task*, *dataset*, *metric*, *score*, and *method* entities. We downloaded the dump from the online source https://paperswithcode.com/about (timestamp 19-10-2021) and obtained data annotations via distance labeling of their crowdsourced annotations for only the *research problem* and *method* entities (see Table 2 for label mappings of *task* as *research problem*). Note that among their five available entities, three entities, viz. *dataset*, *metric*, *score*, were not considered since they were often not direct mentions in the text but were inferable candidates and hence did not satisfy the sequence labeling objective of this work.

Both our **Titles** and **Abstracts** corpora were appended with available PwC annotations for *research problem* and *method* entities. This was done by following two selection criteria: 1) since PwC provided over 50,000 papers, we wanted to select only a subsample of the data to avoid skewing our overall dataset annotations to just the two PwC entities (i.e., *research problem* and *method*), since owing to the *ACL* data (Table 5.2), **Titles** has six semantic types in all. And 2) we wanted to select a subsample size representative enough of PwC to capture the different nature of their crowdsourced annotations. Table 4 provides an insight of how the selection criteria were implemented specifically in terms of how much data was additionally included. Starting with the **Titles** corpus, the PwC titles were grouped in three categories: those with both *research problem* and *method* mentions; those with either one. From each group, roughly 2000 titles were added to dataset. Similarly, for the **Abstracts** corpus, the paper abstracts were grouped in three categories: those with both *research problem* and *method* mentions; those with either one. From each of the three groups, roughly 2000 abstracts on average were added to the overall data.

Table 4. Selection criteria and statistics for Titles and Abstracts, respectively, from https://paperswithcode.com/ (PwC) that were appended in our overall corpus

Titles criteria	**statistics**	**Abstracts criteria**	**statistics**
1, 2, or 3 tasks and 1, 2, or 3 methods	1,855	1, 2, 3, 4, or 5 tasks and methods	1,756
1, 2, 3, or 4 tasks and no method for titles	2,100	1, 2, or 3 tasks and no method	2,500
no task and 1 or 2 methods for titles	2,100	no task and 1, 2, or 3 methods	2,500

Our final resulting corpus distributions in terms of the constituent corpora after "combining existing resources" and adding "our annotated data" was as follows. **Titles** corpus constituent subcorpora distributions: 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. The considered annotations were for seven *contribution-centric* entities, viz. *solution*, *tool*, *dataset*, *language*, *method*, *resource*, and *research problem.* And, the **Abstracts** corpus constituent subcorpora distributions were: 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 by their original sizes and *contribution-centric* entities annotations availability. Our corpus is released at `https://github.com/jd-coderepos/contributions-ner-cs`.

6 Our CS NER Sequence Labeler

6.1 Experimental Setup

Model. Overall, we experimented with six different sequence labeling neural architectures. The basic building blocks to these six architectural variants, inspired from state-of-the-art neural sequence labelers in the general domain [20,70,65,96,85], were are follows: 1) a CNN, a LSTM, or a BiLSTM first layer over word representations of the data, 2) with and without a second char CNN layer, and 3) an output layer as a CRF decoder since CRFs outperformed the softmax function in sequence labeling tasks. Thus the following six architectures were experimented with: i) word CNN + CRF, ii) word LSTM + CRF, iii) word BiLSTM + CRF, iv) word CNN + char CNN + CRF, v) word LSTM + char CNN + CRF, and

vi) word BiLSTM + char CNN + CRF. As mentioned before, each of these architectural configurations, i.e. leveraging only word representations in the first layer or the character-based CNN as the second layer are deconstructions of state-of-the-art sequence labelers for NER in the general domain [20,70,65,96,85]. Further, the word representations for the first layer were computed one of two ways: either directly from the data, or as precomputed vectorized embedding representations.[4]

For implementing the sequence labelers, we leveraged the open-source toolkit called NCRF++ [95] (`https://github.com/jiesutd/NCRFpp`) that is based on PyTorch. Our experimental configuration files for model hyperparameter details including learning rate, dropout rate, number of layers, hidden size etc., are on Github `https://github.com/jd-coderepos/contributions-ner-cs`.

Evaluaton Metrics We leverage the micro Precision (P), Recall (R), and F1-score ($F1$) measures for overall task performance. By micro-measures, the total true positive (tp), false positive (fp), true negative (tn), and false negative (fn) counts are computed over all types across all test data instances, i.e. across all titles or across all abstracts, respectively. Following which the precision, recall, and F1 measures are obtained. This helps evaluate task performance at a fine-grained level per semantic type which would take into account low performances on minority semantic types as well. Further, we also leverage the standard P, R, and $F1$ scores per entity type.

Table 5. Results with different neural architectures for CS NER over seven semantic concepts with embeddings computed on the data source (top row) and with pretained embeddings (bottom row) on the Titles corpus (columns 2 to 4) and Abstracts corpus (columns 5 to 7).

Neural Architectures	**micro**	**micro**	**micro**	**micro**	**micro**	**micro**
	P	R	$F1$	P	R	$F1$
word CNN + CRF	70.28	71.24	70.76	90.55	72.51	80.53
	69.32	69.16	69.24	91.78	73.58	81.68
word LSTM + CRF	69.24	70.08	69.65	85.45	75.54	79.62
	68.41	66.76	67.58	90.02	71.82	79.9
word BiLSTM + CRF	71.92	73.34	72.62	88.22	76.24	81.79
	71.44	72.91	72.17	90.14	76.36	82.68
word CNN + char CNN	71.31	72.96	72.13	78.61	71.08	74.65
+ CRF	72.50	71.01	71.75	88.59	66.33	75.86
word LSTM + char CNN	72.01	72.4	72.21	85.48	78	81.57
+ CRF	71.59	69.65	70.61	87.71	76.49	81.71
word BiLSTM + char	**74.14**	**76.26**	**75.18**	**84.89**	**81.9**	**83.37**
CNN + CRF	73.67	75.16	74.41	88.2	78.85	83.26

[4] We used GloVe embeddings [84].

Table 6. CS NER percentage scores per scientific entity type

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
resource	75.72	78.61	77.14
solution	78.51	82.61	80.51
language	86.22	87.78	86.99

6.2 Results and Analysis

Table 5 shows the results from our six neural sequence labeling architectural configurations over our two respective corpora, viz. **Titles** and **Abstracts**, respectively, for the task of *contribution-centric* CS NER. Our best performing configuration on both datasets is *word-based BiLSTM + character CNN + CRF*. From the first three columns for **Titles** results, the highest performance is 75.18% in micro F1 over its seven entities obtained with word embeddings computed directly on the data source. And from the last three columns for **Abstracts** corpus results, we see the highest performance is 83.37% in micro F1 over its two semantic types again using word embeddings computed directly on the data source. Their performances are analyzed in detail next.

CNN versus LSTM in the first layer? From the results, we observe that word-based BiLSTMs outperform word-based CNNs which in turn outperform word-based LSTMs. Thus word-based BiLSTMs are clearly the best neural model for the first layer for *contributions-centric* CS NER. This observation is aligned with the state-of-the-art NER model configuration in the general domain as well.

Is a char CNN layer preferable in the second layer of the sequence learning neural architecture? We find that it is. Comparing the results in the last three rows with the first three rows of Table 5, shows the models discriminative ability significantly increases. This is more evident for the **Titles** corpus which had seven semantic types compared to the **Abstracts** corpus with only two. In the former case, a more robust model would be needed. The added character CNN layer satisfies this need.

Is it beneficial to leverage pre-trained embeddings? We see that it is more beneficial to compute embeddings directly on the dataset rather than using the pretrained embeddings out-of-the-box. However, present advanced embedding models based on transformers such as BERT [29] and its variants [10] also allow finetuning the pretrained embeddings on respective experimental datasets. We relegate this experiment to future work. We hypothesize that such embeddings could be leveraged with effective results in a sequence labeling setting as well. However, considering the case presented in this work, i.e. leveraging word embeddings directly computed on a large enough underlying data source versus using pretrained word embeddings, we empirically verify that the former method is better suited to the task.

Which scientific entity type is easiest versus hardest to classify? These results are shown in Table 6. Viewing the scores of the seven entity types with the **Titles** and the **Abstracts** datasets as one combined resource, we find the five semantic types, viz. *dataset*, *research problem*, *resource*, *solution*, and *language* obtained scores above 70%. Thus for these types our model proves practically suitable. Of these five types, *language* was the highest performing extraction target. This is also consistent with the IAA scores (see Table 3) between the human annotators who annotated the *language* entity type with perfectly matching annotation consistency. The sequence labeler performed worst on *method* entity type. This score discrepancy can be explained by the confusion between the *tool* entity type and *method* entity type which were shown not easily distinguishable during the IAA experiments as well.

7 Conclusion and Future Directions

This work has reported a research direction on unifying prior work on scholarly domain-specific NER, specifically for CS NER. It discussed the reuse of existing resources and the complementary addition of new annotations as a contributing publicly available language resource in the community. Furthermore, drawing on observations of state-of-the-art NER systems in the general domain where the NER task itself has garnered much research interest, six neural sequence labeling architectural variants were empirically tested for CS NER. Consequently new machine learning empirical insights could be supplied as a result of this work regarding the strengths of suitable architectural components. We show that the overall CS NER task of extracting *contribution-centric* entities involving seven semantic types has performances above 75% demonstrating itself as a reliable predictor of entities in practical, real-world system usage settings. The code base is publicly released `https://tinyurl.com/orkg-cs-ner`, as well as service calls via a REST API `https://tinyurl.com/csner-rest-api` and as a Python package `https://tinyurl.com/cs-ner-pypi`.

Given rapid scholarly publication trends [49] — one that is only further bolstered with the sharing of PDF preprints ahead (or even instead) of peer-reviewed publications [19] — the research community is faced with a crucial dilemma. *How to stay on-track with the past and the current rapid-evolving research progress?* The manual comprehension of this information is nearly impossible for humans. Furthermore, an enormous amount of digital information is expressed as natural-language (NL) text that is not easily processable by computers. Knowledge Graphs (KG) offer a widely used format for representing information in computer-processable form which are increasingly being supported in next-generation scholarly digital library platforms, e.g., ORKG [4]. Such systems currently rely on crowdsourcing methods to obtain good quality data. NLP is therefore needed for mining knowledge graphs from texts. A central part of the problem is to extract the named entities. To this end, this work has taken on the branch of CS NER with a *contribution-centric* entities information extraction aim. As part of future work, the existing set of CS NER entities will be investigated to increase the types coverage. Further, other domains of the Science will also be explored for the scholarly-domain-specific NER task.

References

1. SciGraph. https://www.springernature.com/de/researchers/scigraph, accessed: 2021-11-02
2. A reintroduction to our knowledge graph and knowledge panels. https://blog.google/products/search/about-knowledge-graph-and-knoswledge-panels/ (2020), accessed: 2020-07-16
3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000)
4. Auer, S., Oelen, A., Haris, M., Stocker, M., D'Souza, J., Farfar, K.E., Vogt, L., Prinz, M., Wiens, V., Jaradeh, M.Y.: Improving access to scientific literature with knowledge graphs. Bibliothek Forschung und Praxis **44**(3), 516–529 (2020)
5. Auer, S.: Towards an open research knowledge graph (Jan 2018). https://doi.org/10.5281/zenodo.1157185, https://doi.org/10.5281/zenodo.1157185
6. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 546–555. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/S17-2091, https://aclanthology.org/S17-2091
7. Baas, J., Schotten, M., Plume, A., Côté, G., Karimi, R.: Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies **1**(1), 377–386 (2020)
8. Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W.A., Cohen, K.B., Verspoor, K., Blake, J.A., et al.: Concept annotation in the craft corpus. BMC bioinformatics **13**(1), 1–20 (2012)
9. Bader, G.D., Cary, M.P., Sander, C.: Pathguide: a pathway resource list. Nucleic acids research **34**(suppl_1), D504–D506 (2006)
10. Beltagy, I., Lo, K., Cohan, A.: 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). pp. 3615–3620 (2019)
11. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media, Inc., 1st edn. (2009)
12. Birkle, C., Pendlebury, D.A., Schnell, J., Adams, J.: Web of science as a data source for research on scientific and scholarly activity. Quantitative Science Studies **1**(1), 363–376 (2020)
13. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Research **32**(Database issue), D267 (2004)
14. Bornmann, L., Mutz, R.: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology **66**(11), 2215–2222 (2015)

15. Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler, R.: The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. Nucleic Acids Research **32**(Database issue), D262 (2004)
16. Carreras, X., Màrquez, L., Padró, L.: A simple named entity extractor using adaboost. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. pp. 152–155 (2003)
17. Chalapathy, R., Zare Borzeshi, E., Piccardi, M.: An investigation of recurrent neural architectures for drug name recognition. In: Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis. pp. 1–5. Association for Computational Linguistics, Auxtin, TX (Nov 2016). https://doi.org/10.18653/v1/W16-6101, https://aclanthology.org/W16-6101
18. Chatr-Aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L., Cesareni, G.: Mint: the molecular interaction database. Nucleic acids research **35**(suppl_1), D572–D574 (2007)
19. Chiarelli, A., Johnson, R., Richens, E., Pinfield, S.: Accelerating scholarly communication: The transformative role of preprints (2019)
20. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional lstm-cnns. Transactions of the Association for Computational Linguistics **4**, 357–370 (2016)
21. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
22. Collier, N., Kim, J.D.: 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). pp. 73–78. COLING, Geneva, Switzerland (Aug 28th and 29th 2004), https://aclanthology.org/W04-1213
23. Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (1999), https://aclanthology.org/W99-0613
24. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**(null), 2493–2537 (Nov 2011)
25. Corbett, P., Batchelor, C., Teufel, S.: Annotation of chemical named entities. In: Biological, translational, and clinical language processing. pp. 57–64 (2007)
26. Cunningham, H., Humphreys, K., Gaizauskas, R., Wilks, Y.: Gate - a general architecture for text engineering. pp. 29–30 (01 1997). https://doi.org/10.3115/974281.974299
27. Dernoncourt, F., Lee, J.Y., Szolovits, P.: NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 97–102. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-2017, https://aclanthology.org/D17-2017
28. Dessì, D., Osborne, F., Recupero, D.R., Buscaldi, D., Motta, E., Sack, H.: Ai-kg: an automatically generated knowledge graph of artificial intelligence. In: Pan, J., Tamma, V., d?Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web ? ISWC 2020. vol. 12507, pp. 127–143. Springer (2020), http://oro.open.ac.uk/71736/
29. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)

30. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The automatic content extraction (ACE) program – tasks, data, and evaluation. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). European Language Resources Association (ELRA), Lisbon, Portugal (May 2004), http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf
31. Doğan, R.I., Leaman, R., Lu, Z.: Ncbi disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics **47**, 1–10 (2014)
32. D'Souza, J., Auer, S.: Pattern-based acquisition of scientific entities from scholarly article titles. arXiv preprint arXiv:2109.00199 (2021)
33. D'Souza, J., Auer, S., Pedersen, T.: 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). pp. 364–376. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.semeval-1.44, https://aclanthology.org/2021.semeval-1.44
34. D'Souza, J., Hoppe, A., Brack, A., Jaradeh, M.Y., Auer, S., Ewerth, R.: 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. pp. 2192–2203. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.268
35. Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence **165**(1), 91–134 (2005). https://doi.org/https://doi.org/10.1016/j.artint.2005.03.001, https://www.sciencedirect.com/science/article/pii/S0004370205000366
36. Färber, M., Albers, A., Schüber, F.: 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. CEUR Workshop Proceedings, vol. 2831. CEUR Workshop Proceedings (CEUR-WS) (2021)
37. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. p. 363–370. ACL '05, Association for Computational Linguistics, USA (2005). https://doi.org/10.3115/1219840.1219885, https://doi.org/10.3115/1219840.1219885
38. Gábor, K., Buscaldi, D., Schumann, A.K., QasemiZadeh, B., Zargayouna, H., Charnois, T.: SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 679–688. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/S18-1111, https://aclanthology.org/S18-1111

39. Gangemi, A., Presutti, V., Recupero, D.R., Nuzzolese, A.G., Draicchio, F., MongiovÃ¬, M.: Semantic Web Machine Reading with FRED. Semantic Web **8**(6), 873–893 (2017)
40. Grishman, R., Sundheim, B.: Message Understanding Conference- 6: A brief history. In: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics (1996), `https://aclanthology.org/C96-1079`
41. Gupta, S., Manning, C.: Analyzing the dynamics of research by extracting key aspects of scientific papers. In: Proceedings of 5th International Joint Conference on Natural Language Processing. pp. 1–9. Asian Federation of Natural Language Processing, Chiang Mai, Thailand (Nov 2011), `https://aclanthology.org/I11-1001`
42. Hendricks, G., Tkaczyk, D., Lin, J., Feeney, P.: Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies **1**(1), 414–427 (2020)
43. Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., Declerck, T.: The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of biomedical informatics **46**(5), 914–920 (2013)
44. Honnibal, M.: SpaCy. `https://github.com/explosion/spaCy` (2015), accessed: 2021-11-09
45. Hou, Y., Jochim, C., Gleize, M., Bonin, F., Ganguly, D.: 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. pp. 5203–5213. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1513, `https://aclanthology.org/P19-1513`
46. Ioannidis, J.P.: The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. The Milbank Quarterly **94**(3), 485–514 (2016)
47. Islamaj, R., Leaman, R., Kim, S., Kwon, D., Wei, C.H., Comeau, D.C., Peng, Y., Cissel, D., Coss, C., Fisher, C., et al.: Nlm-chem, a new resource for chemical entity recognition in pubmed full text literature. Scientific Data **8**(1), 1–12 (2021)
48. Jain, S., van Zuylen, M., Hajishirzi, H., Beltagy, I.: SciREX: A challenge dataset for document-level information extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7506–7516. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.670, `https://aclanthology.org/2020.acl-main.670`
49. Jinha, A.E.: Article 50 million: an estimate of the number of scholarly articles in existence. Learned Publishing **23**(3), 258–263 (2010)
50. Kabongo, S., D'Souza, J., Auer, S.: Automated mining of leaderboards for empirical ai research. In: International Conference on Asian Digital Libraries. pp. 453–470. Springer (2021)
51. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., et al.: Intact—open source resource for molecular interaction data. Nucleic acids research **35**(suppl_1), D561–D565 (2007)
52. Kim, J.D., Ohta, T., Tateisi, Y., Tsujii, J.: Genia corpus—a semantically annotated corpus for bio-textmining. Bioinformatics **19**(suppl_1), i180–i182 (2003)
53. Kim, J.D., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Extracting bio-molecular events from literature—the bionlp'09 shared task. Computational Intelligence **27**(4), 513–540 (2011)

54. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. p. 2741–2749. AAAI'16, AAAI Press (2016)
55. Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., Tshitoyan, V., Ceder, G.: Text-mined dataset of inorganic materials synthesis recipes. Scientific data **6**(1), 1–11 (2019)
56. Krallinger, M., Izarzugaza, J.M., Rodriguez-Penagos, C., Valencia, A.: Extraction of human kinase mutations from literature, databases and genotyping studies. BMC bioinformatics **10**(8), 1–20 (2009)
57. Krallinger, M., Leitner, F., Rodriguez-Penagos, C., Valencia, A.: Overview of the protein-protein interaction annotation extraction task of biocreative ii. Genome biology **9**(2), 1–19 (2008)
58. Krallinger, M., Leitner, F., Valencia, A.: Analysis of biological processes and diseases using text mining approaches. Bioinformatics Methods in Clinical Research pp. 341–382 (2010)
59. Krallinger, M., Miranda, A., Mehryary, F., Luoma, J., Pyysalo, S., Valencia, A.: Drugprot shared task (biocreative vii track 1-2021) text mining drug-protein/gene interactions (drugprot) shared task (2021)
60. Krallinger, M., Rabal, O., Akhondi, S.A., Pérez, M.P., Santamaría, J., Rodríguez, G.P., Tsatsaronis, G., Intxaurrondo, A., López, J.A.B., Nandal, U.K., van Buel, E.M., Chandrasekhar, A., Rodenburg, M., Lægreid, A., Doornenbal, M.A., Oyarzábal, J., Lourenço, A., Valencia, A.: Overview of the biocreative vi chemical-protein interaction track (2017)
61. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D.M., et al.: The chemdner corpus of chemicals and drugs and its annotation principles. Journal of cheminformatics **7**(1), 1–17 (2015)
62. Krallinger, M., Vazquez, M., Leitner, F., Salgado, D., Chatr-Aryamontri, A., Winter, A., Perfetto, L., Briganti, L., Licata, L., Iannuccelli, M., et al.: 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 (2011)
63. Kulkarni, C., Xu, W., Ritter, A., Machiraju, R.: An annotated corpus for machine reading of instructions in wet lab protocols. In: NAACL: HLT, Volume 2 (Short Papers). pp. 97–106. New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-2016
64. Kuniyoshi, F., Makino, K., Ozawa, J., Miwa, M.: Annotating and extracting synthesis process of all-solid-state batteries from scientific literature. In: LREC. pp. 1941–1950 (2020)
65. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 260–270 (2016)
66. Lewis, N., Wang, J., Poblet, M., Aryani, A.: Research graph: Connecting researchers, research data, publications and grants using the graph technology. In: eResearch Australasia Conference (2016), `https://eresearchau.files.wordpress.com/2016/03/eresau2016_paper_95.pdf`
67. Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.H., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C., Lu, Z.: Biocreative v cdr task corpus: a resource for chemical disease relation extraction. Database **2016** (2016)
68. Li, Y., Bontcheva, K., Cunningham, H.: Svm based learning system for information extraction. In: International Workshop on Deterministic and Statistical Methods in Machine Learning. pp. 319–339. Springer (2004)

69. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-task identification of entities, relations, and coreferencefor scientific knowledge graph construction. In: Proc. Conf. Empirical Methods Natural Language Process. (EMNLP) (2018)
70. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1064–1074 (2016)
71. Malouf, R.: Markov models for language-independent named entity recognition. In: Proceedings of the 6th Conference on Natural Language Learning - Volume 20. p. 1–4. COLING-02, Association for Computational Linguistics, USA (2002). https://doi.org/10.3115/1118853.1118872, `https://doi.org/10.3115/1118853.1118872`
72. Manghi, P., Manola, N., Horstmann, W., Peters, D.: An infrastructure for managing ec funded research output: The openaire project. Grey Journal (TGJ) **6**(1) (2010)
73. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4. p. 188–191. CONLL '03, Association for Computational Linguistics, USA (2003). https://doi.org/10.3115/1119176.1119206, `https://doi.org/10.3115/1119176.1119206`
74. McCallum, A.K.: Mallet: A machine learning for language toolkit (2002), http://mallet.cs.umass.edu
75. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th international conference on semantic systems. pp. 1–8. ACM (2011)
76. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: Shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems. p. 1–8. I-Semantics '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2063518.2063519, `https://doi.org/10.1145/2063518.2063519`
77. Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Yang, B., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., et al.: Never-ending learning. Communications of the ACM **61**(5), 103–115 (2018)
78. Mohan, S., Li, D.: Medmentions: A large biomedical corpus annotated with umls concepts. In: Automated Knowledge Base Construction (AKBC) (2018)
79. Mondal, I., Hou, Y., Jochim, C.: End-to-end construction of NLP knowledge graph. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 1885–1895. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.165, `https://aclanthology.org/2021.findings-acl.165`
80. Mondal, I., Hou, Y., Jochim, C.: End-to-end construction of NLP knowledge graph. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 1885–1895. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.165, `https://aclanthology.org/2021.findings-acl.165`
81. Moro, A., Cecconi, F., Navigli, R.: Multilingual Word Sense Disambiguation and Entity Linking for Everybody. In: ISWC. pp. 25–28. Riva del Garda, Italy (2014)
82. Mysore, S., Jensen, Z., Kim, E., Huang, K., Chang, H.S., Strubell, E., Flanigan, J., McCallum, A., Olivetti, E.: The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In: Proceedings of the 13th Linguistic Annotation Workshop. pp. 56–64 (2019)

83. Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., Taylor, J.: Industry-scale knowledge graphs: lessons and challenges. Queue **17**(2), 48–75 (2019)
84. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1162, `https://aclanthology.org/D14-1162`
85. Peters, M., Ammar, W., Bhagavatula, C., Power, R.: Semi-supervised sequence tagging with bidirectional language models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1756–1765 (2017)
86. QasemiZadeh, B., Schumann, A.K.: 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). pp. 1862–1868. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), `https://aclanthology.org/L16-1294`
87. Rocktäschel, T., Huber, T., Weidlich, M., Leser, U.: WBI-NER: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 356–363. Association for Computational Linguistics, Atlanta, Georgia, USA (Jun 2013), `https://aclanthology.org/S13-2058`
88. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J.P., Sun, L., Turner, S., Karsch-Mizrachi, I.: NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database **2020** (08 2020). https://doi.org/10.1093/database/baaa062, `https://doi.org/10.1093/database/baaa062`, baaa062
89. Tanabe, L., Xie, N., Thom, L.H., Matten, W., Wilbur, W.J.: Genetag: a tagged corpus for gene/protein named entity recognition. BMC bioinformatics **6**(1), 1–7 (2005)
90. Tjong Kim Sang, E.F.: Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In: COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002) (2002), `https://aclanthology.org/W02-2024`
91. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. pp. 142–147 (2003), `https://aclanthology.org/W03-0419`
92. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
93. Wang, K., Shen, Z., Huang, C., Wu, C.H., Dong, Y., Kanakia, A.: Microsoft academic graph: When experts are not enough. Quantitative Science Studies **1**(1), 396–413 (2020)
94. Yadav, V., Sharp, R., Bethard, S.: Deep affix features improve neural named entity recognizers. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. pp. 167–172. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/S18-2021, `https://aclanthology.org/S18-2021`

95. Yang, J., Zhang, Y.: Ncrf++: An open-source neural sequence labeling toolkit. In: Proceedings of ACL 2018, System Demonstrations. pp. 74–79 (2018)
96. Yang, Z., Salakhutdinov, R., Cohen, W.W.: Transfer learning for sequence tagging with hierarchical recurrent networks. arXiv preprint arXiv:1703.06345 (2017)