Extracting Methodology Components from AI Research Papers: A Data-driven Factored Sequence Labeling Approach

Madhusudan Ghosh
madhusuda.iacs@gmail.com
Indian Associantion for the Cultivation of Science
Jadavpur, Kolkata, India

Debasis Ganguly
debasis.ganguly@glasgow.ac.uk
University of Glasgow
Glasgow, United Kingdom

Partha Basuchowdhuri
partha.basuchowdhuri@iacs.res.in
Indian Associantion for the Cultivation of Science
Jadavpur, Kolkata, India

Sudip Kumar Naskar
sudip.naskar@gmail.com
Jadavpur University
Jadavpur, Kolkata, India

ABSTRACT

Extraction of methodology component names from scientific articles is a challenging task due to the diversified contexts around the occurrences of these entities, and the different levels of granularity and containment relationships exhibited by these entities. We hypothesize that standard sequence labeling approaches may not adequately model the dependence of methodology name mentions with their contexts, due to the problems of their large, fast evolving, and domain-specific vocabulary. As a solution, we propose a factored approach, where the mention-context dependencies are represented in a more fine-grained manner, thus allowing the model parameters to better adjust to the different characteristic patterns inherent within the data. In particular, we experiment with two variants of this factored approach - one that uses the per-entity category information derived from an ontology, and the other that makes use of the topology of the sentence embedding space to infer a category for each entity constituting that sentence. We demonstrate that both these factored variants of SciBERT outperform their non-factored counterpart, a state-of-the-art model for scientific concept extraction.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**; • **Computing methodologies → Machine learning**.

KEYWORDS

Information Extraction; Factored Model; Clustering; Scientific Literature

ACM Reference Format:
Madhusudan Ghosh, Debasis Ganguly, Partha Basuchowdhuri, and Sudip Kumar Naskar. 2023. Extracting Methodology Components from AI Research Papers: A Data-driven Factored Sequence Labeling Approach. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, UK*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615258

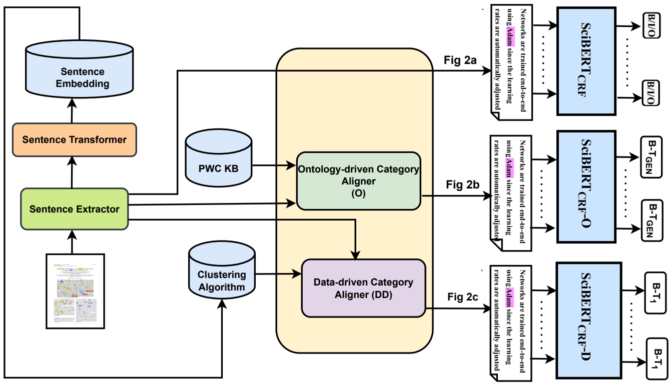


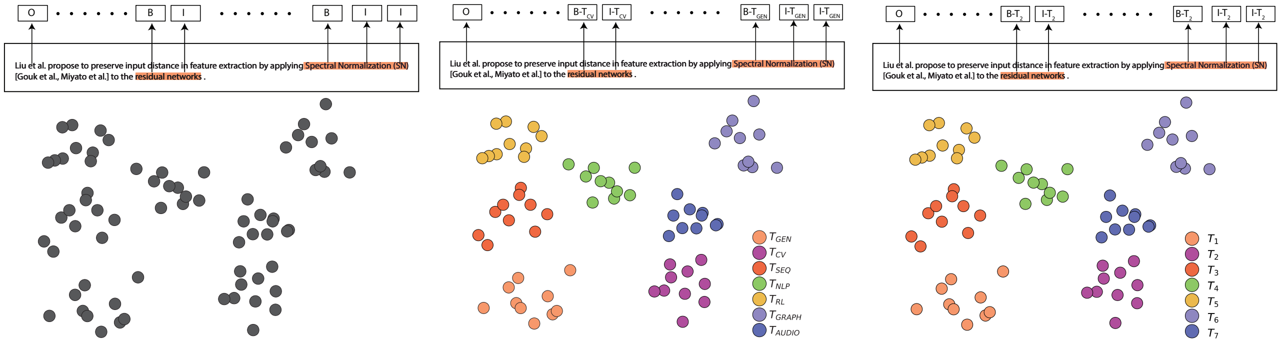
Figure 1: An overview of our proposed category-based factored sequence modeling approach for scientific concept extraction. The categories are either obtained from an ontology, e.g., the Paperswith-Code knowledge base, or via clustering the sentences.

1 INTRODUCTION

Scientific literature usually grows at a rapid rate embracing new theories, methodologies and their empirical validations or refuting [15]. Existing research [13, 26] has applied automated information extraction (IE) approaches for identifying relatively *well-defined entities* from scientific articles, e.g., task and dataset names, which often are proper nouns. However, the situation is potentially more challenging when the objective is to extract more loosely defined concepts, such as methodology names, e.g., BERT (a transformer architecture used in NLP [10]), MLM (a specific type of pre-training a language model that involves masking random words) or Adam (an optimisation method). Moreover, such rapid advances in scientific methodologies create difficulty for researchers to maintain a comprehensive and updated knowledge of the recent literature, which is critical for academic tasks, such as developing novel research ideas, selecting the correct baselines [2], peer-reviewing others' research and also scientific paper recommendation system [6]. In contrast to task and dataset names, they involve different levels of granularity and containment relationships, e.g., while MLM

CIKM '23, October 21–25, 2023, Birmingham, UK

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615258


O
B
I
B
I
I
B-T_CV
I-T_CV
B-T_GEN
I-T_GEN
I-T_GEN
B-T_2
I-T_2
B-T_2
I-T_2
I-T_2
Liu et al. propose to preserve input distance in feature extraction by applying Spectral Normalization (SN) [Gouk et al., Miyato et al.] to the residual networks .
T_GEN
T_CV
T_SEQ
T_NLP
T_RL
T_GRAPH
T_AUDIO
T_1
T_2
T_3
T_4
T_5
T_6
T_7

(a) Non-factored model **(b) Per-entity categories** **(c) Per-data instance categories**

Figure 2: A schematic illustration of how the categories affect the granularity of the BIO tags. Embeddings of individual sentences are shown as 2d points to visually illustrate the concept of partitioning them into fine-grained types.

and `Adam` refer to more fine-grained entities, `BERT`, on the other hand, is itself composed of other architectural components, such as '`attention heads`', '`positional embeddings`' etc. Manually annotating such concept mentions from scientific articles is a near impossible task [27]. Consequently, supervised models have been developed with silver standard data, where the entity names are first looked up from an ontology, following which the mentions of these entities within their contexts are then used to train sequence labeling approaches for the task of extraction [1, 3, 14, 17]. For instance, the tags assigned by authors on the uploaded versions of their papers in the PaperswithCode [23] repository provide an example of such an ontology.

Such weakly supervised approaches [4, 18, 19, 21, 24, 28] work adequately well for extraction of well-defined concepts such as task and dataset names because the contexts around these entities are likely to be more consistent, whereas the methodology component names potentially occur within more diversified contexts. This can be attributed to the following likely reasons. First, the methodology names exhibit a larger vocabulary because the number of novel methods is 2,099 as per the recent update of PwC KB, usually, higher than that of tasks or datasets, as tasks and datasets are shared across methods for a fair comparison and not the other way round. Second, the semantics associated with the methodology component names often evolve with time, e.g., while `embedding` used to correspond to static (non-contextual word embeddings) around 2012-2016, the term now, often exclusively, refers to contextual embeddings. Third, in contrast to the task and dataset names, the contexts around methodology name mentions are likely to depend on the topic of an article, e.g., the methodology name `convolutional network` is associated with different characteristics across the domains of computer vision, NLP or graphs.

Our contributions. We hypothesize that standard sequence labeling approaches may not adequately model the dependence of methodology name mentions with their contexts, due to the problems of their large, fast evolving, and domain-specific vocabulary. To this end, we propose a factored approach, where the mention-context dependencies are represented in a more fine-grained manner, thus allowing the model parameters to better adjust to the different characteristic patterns of the data, e.g., adapting to the diverse contexts of the entity '`convolutional network`' for computer vision, NLP or graphs (schematic workflow in Figure 1).

2 PROPOSED METHODOLOGY

Standard non-factored sequence labeling. We first briefly review the non-factored standard approach of sequence modeling. Let $\mathbf{x}$ be a sentence (generally speaking, a token sequence) of maximum length M in a training set $\mathcal{T}$, i.e., $\mathbf{x} = \{x_0, \ldots, x_{M-1}\}$, where each x_j is a token of that sentence. Let the set of mentions of entities occurring in $\mathbf{x}$ be $e(\mathbf{x}) = \{e_0, \ldots, e_{m-1}\} \subset \mathcal{D}$, where $\mathcal{D}$ denotes a set of such names existing as a part of an ontology resource, e.g., the set of tags in paperswithcode (PwC)[1].

Each mention, $e(\mathbf{x})$, is a subsequence of $\mathbf{x}$, which we denote by an indicator sequence of positions $\{\mathbb{I}(x_0), \ldots, \mathbb{I}(x_{M-1})\}$, where $\mathbb{I}(x_i) = 1$ if x_i is a part of some entity $e \in e(\mathbf{x})$. A continuous span $\{j, \ldots, j{+}n{-}1\}$ (where $j \in \mathbb{Z}_M$) such that $\mathbb{I}(x_i) = 1, \forall j \leq i \leq j{+}n{-}1$ indicates an entity constituted of n tokens. To differentiate the start of a span from its continuity and also its end, it is a common practice to denote the label of the first element of such an index set with B (denoting Beginning of a span), the subsequent elements as I (denoting that these are Inside a span) and the first index after the span ends, i.e., the one in the $(j+n)^{\text{th}}$ index as O (indicating that this is Outside a span), or a B if another sequence starts from this index [5]. Thus, each token sequence $\mathbf{x} \in \mathcal{T}$ of length M is mapped to a label sequence of the same length, i.e., $\mathbf{x} = \{x_0, \ldots, x_{M-1}\} \mapsto \mathbf{y} = \{y_0, \ldots y_{M-1}\}$ where each $y_i \in \{B, I, O\}$. An example of this non-factored instance is presented in Figure 2a.

2.1 Factored Model for Sequence Labeling

Data-driven partitioning. In our factored approach, we partition the training set $\mathcal{T}$ into k subsets such that $\cup_{i=0}^{k-1} \mathcal{T}_i = \mathcal{T}$ and $\mathcal{T}_i \cap \mathcal{T}_j = \emptyset \; \forall i \neq j$. Then a word sequence can be represented as, $\mathbf{x} \in \mathcal{T}_i \subset \mathcal{T}$, for some $i \in \mathbb{Z}_k$ and $0 \leq i \leq (k-1)$. Instead of using the binary labels, i.e., indicators of the form $\mathbb{I}(x_j)$, to denote if the j^{th} token is a part of a mention name, we now use k-ary labels to indicate the subset in which $\mathbf{x}$ occurs. For instance, if $\mathbf{x} \in \mathcal{T}_i$ and an entity e of length n tokens is a part of $\mathbf{x}$, say $e = \{j, \ldots, j{+}n{-}1\}$ then

[1] https://paperswithcode.com/methods

	Std. dev	**Mean**
The distribution of papers over time	3709.4	2152.1
Distribution of number of method tags per paper	4.6	6.7
Distribution of distinct tags present in the dataset	510.7	98.7

Table 1: Detailed statistics of the dataset used in our experiments.

$\{y_j, y_{j+1}, \ldots, y_{j+n-1}, y_{j+n-1}\} = \{B_i, I_i, \ldots, I_i, O_i\}$. An example of this data-driven instances is depicted in Figure 2c.

With this partitioned approach, the parameterised approximation $\theta : \mathbf{x} \mapsto \mathbf{y}$ is able to focus on not just the context of a token but also its category type; this is because instead of a generic beginning of an entity span (B), or its continuation (I), the model potentially adapts to category-specific beginning, continuation or end of spans (e.g., B_1 is different from B_2). The most generic way to obtain this partitioning that does not rely on any external ontology source is to cluster the set $\mathcal{T}$. More precisely, in this method, we cluster the set of training instances by a standard clustering approach, such as k-means, by making use of the encoded representations of the sentences (in our experiments, both tf-idf based sparse representation and SciBERT [3] encoding based dense representation).

Ontology-driven partitioning. This approach relies on the existence of a hierarchical ontology where the leaf-level entities are categorised into more coarse-grained types, e.g., the 'Category' metadata of PwC constitutes 7 different categories, namely General (GEN), Computer Vision (CV), Sequence-to-Sequence Modeling (SEQ), Reinforcement Learning (RL), NLP, Audio & Speech (AUDIO), and Graph-based Modeling (GRAPH). With this mode of partitioning, we first look up the category of each entity $e \in e(\mathbf{x})$ ($e(\mathbf{x})$ being the set of all entity mentions within a sentence), and then accordingly set the target labels corresponding to the span of e. More precisely, if $c(e)$ denotes one of k categories of an entity as defined in an ontology, the target BIO labels corresponding to the positions where e occurs are changed to $\{B_c, I_c, O_c\}$. For this way of partitioning, it may happen that the label sequence $\mathbf{y}$ for a sentence $\mathbf{x}$ can be comprised of more than one unique type of specific versions of the generic BIO types, e.g. both B_i and B_j can be a part of $\mathbf{y}$ with $i \neq j$. An annotated example in this particular setting is presented in Figure 2b. Note that this is not possible in the data-driven mode of partitioning because all the entities present in a sentence $\mathbf{x}$ belongs to only a single category. We denote these two factored approaches in our experiments with the subscripts 'D' (Data-driven) and 'O' (Ontology-driven).

3 EXPERIMENT SETUP

Dataset. For our experiments, we downloaded the PwC JSON dump [9], which comprises a total of 291,503 papers. The metadata of each paper in this dataset includes the title, abstract, a list of tags indicating methodology components (e.g., 'Adam', 'LSTM', etc.), arxiv link to the paper. Additionally, each paper is linked with a broad-level category name, which we used for the ontology-driven factored sequence modeling (c.f. Section 2.1).

For a consistent experimental setup, we removed articles that do not contain this metadata information. This yielded in a total of 34,560 papers in our dataset (summarised in Table 1). We used the SciPDF parser [12] for extracting the content from these articles. The parser outputs a structured view of the document text segmented into sections, figures and tables. In contrast to the previous work of Hou et al. [13], we exclude the captions of figures and tables since it is unlikely that a methodology name will only appear in a figure or a table caption without occurring within the text of a document.

	Model	90:10 split on ≤2017			Train: ≤ 2017, Test: > 2017		
		Prec.	Recall	F-score	Prec.	Recall	F-score
Baselines	BiLSTM$_{\text{CRF}}$	0.8257	0.5666	0.6720	**0.6460**	0.1933	0.2976
	SciBERT	0.9632	0.9863	0.9746	0.3722	0.2380	0.2903
	SciBERT$_{\text{CRF}}$	**0.9832**	0.9881	0.9857	0.4383	**0.2452**	**0.3145**
	SciREX	0.9831	**0.9887**	**0.9859**	0.4604	0.2177	0.2956
Ours	BiLSTM$_{\text{CRF}}$-O	0.8765	0.6677	0.7579	0.6628	0.2086	0.3173
	SciBERT-O	0.9616	0.9845	0.9729	0.4429	0.1938	0.2696
	SciBERT$_{\text{CRF}}$-O	0.9821	**0.9887**	0.9854	0.5193	0.2236	0.3111
	SciREX-O	**0.9853**	0.9878	**0.9865**	0.4529	0.2184	0.2947
	BiLSTM$_{\text{CRF}}$-D	0.8776	0.6169	0.7245	0.6736	0.2257	0.3381
	SciBERT-D	0.9803	0.9883	0.9843	0.5134	0.2473	0.3338
	SciBERT$_{\text{CRF}}$-D	0.9803	0.9861	0.9832	**0.5836**	**0.3065**	**0.4019**
	SciREX-D	0.9803	0.9876	0.9839	0.5473	0.2413	0.3349

Table 2: A comparison between the methods investigated on two different evaluation splits - (i) conventional 90:10 split, and (ii) chronological split for a few-shot setup. All the factored models used $k = 7$ categories. Best results for both baselines and our methods are both bold-faced; additionally, the overall best results along each column are also underlined.

Since our task is to extract the methodology components, we represented each article by concatenating the text from the following sections: 'Abstract', 'Introduction', 'Methodology', 'Experiments' and 'Results', and follow the nomenclature of convention of Hou et al. [13] to name our document representation as DocAIMER.

To simulate a more realistic situation of identifying new scientific concepts, we induced a chronological partition of the dataset for training and evaluation. Papers dated up to 2017 (inclusive) were included for training, whereas the ones published from 2018 to date were used for evaluation. This split of the data makes the experimental setup realistic in the sense that there has been significant changes in several domains of AI after the introduction of the transformer architecture.

Research Questions. As part of our investigation related to the hypothesis of factored approaches to IE via sequence labeling, we explored the following research questions.

- **RQ-1**: How effectively do our proposed two-factored approaches work on in-domain and out-domain data, the latter indicating a setup where new entities need to be identified without the trained models being aware of them?
- **RQ-2**: What is the most effective way to factor a sequence labeling model for IE, i.e., the data-driven or the ontology-driven?

Baselines. To demonstrate the effectiveness of our proposed factored-based approaches, we compare them with the following baseline methods [3, 14, 20, 25].

- **BiLSTM$_{\text{CRF}}$ [20]**: This serves as a standard baseline for many IE tasks, including NER [7]. The method employs a standard BiLSTM-based approach followed by a CRF decoder layer where embedded words are fed as input. More precisely, for word embedding, we used GloVe [22] with a dimensionality of 100.

- **SciBERT [3]**: We fine-tune SciBERT [3] as it has been reported as the state-of-the-art pre-trained language model for modeling the semantics of the scientific domain.
- **SciBERT$_{\text{CRF}}$ [25]**: In this method, similar to the approach of enhancing BiLSTM by incorporating CRF, we also include a CRF layer into SciBERT to potentially yield a stronger baseline.
- **SciREX [14]**: This approach considers SciBERT as the pretrained base language model followed by a BiLSTM encoder layer and a CRF decoder layer. Finally, the whole framework is trained in an end-to-end fashion on the downstream scientific IE task.

The difference between our proposed approach and the baseline methods is that the existing approaches do not utilize the fine-grained category label information, either at the level of individual entities (derived from an ontology) or at the broad-level topics (derived by clustering the data instances). Any improvements observed in our results can, hence, be attributed to the factoring.

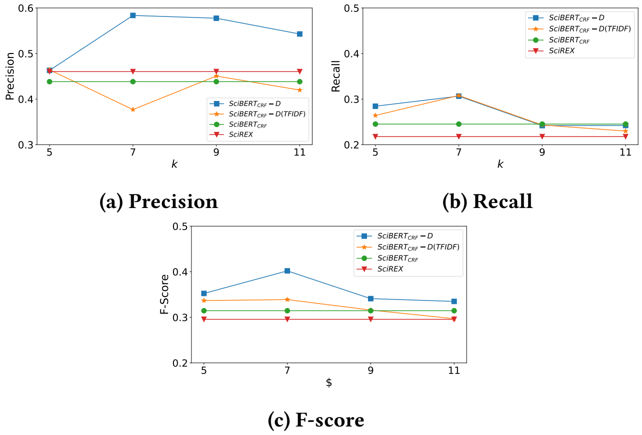
Variants of the proposed methodology. To investigate the research questions, we employ the following variants to explore the effect of data-driven vs. ontology-specific factorisation (RQ-2).

- **SciBERT-O**: This approach uses the 7 static categories from the PwC ontology by incorporating them into the labels for each entity mention. These modified fine-grained labels are then used to fine-tune SciBERT language model on our downstream methodology extraction task.
- **SciBERT$_{\text{CRF}}$-O**: Similar to SciBERT-O, SciBERT$_{\text{CRF}}$-O uses an additional CRF decoder layer on top of SciBERT-O model on our downstream task.
- **SciREX-O**: Here, SciREX is fine-tuned on the ontology-driven category specific information on the scientific IE task.
- **SciBERT-D**: It is similar to the SciBERT-O approach, but here the fine-grained labels are obtained by clustering the SciBERT encoded dense vectors of the sentences.
- **SciBERT$_{\text{CRF}}$-D** [2]: It is similar to SciBERT$_{\text{CRF}}$-O, but here we again use clustering to derive the class labels.
- **SciREX-D**: It is similar to SciREX-O, but here we use clustering to derive the class labels.

We utilize the FLAIR framework [11] to train our all the proposed models. In terms of the common neural network settings, we used AdamW [16] as the optimizer with a learning rate of 5e-5 and a stopping criterion as mentioned in [8]; the training batch size used was 32. We always report our results as the average of the results of 5 different runs of the same experiment for 5 seeds respectively.

4 RESULTS AND ANALYSIS

In relation to RQ-1, Table 2 presents a comparison between the models investigated. We can observe from Table 2 that the factored models exhibit comparable performance with respect to the baselines for in-domain evaluation. For out-domain (chronologically split data), the data-driven approach turns out to be the best one, which provides the answer to RQ-2 that the data-driven way of factoring is more effective. The non-homogeneity introduced due to the possibility of different entity categories within the same sentence potentially degrades the prediction quality.



(a) Precision

(b) Recall

(c) F-score

Figure 3: Sensitivity of our approach on the number of clusters, k.

Additionally, we also investigate how does clustering on the sparse bag-of-words encoding of the data instances compare with that of clustering the dense representation of the sentences (as reported in all the data-driven factoring based results of Table 2). In particular, we used tf-idf based vector representations of the sentences, which we denote with the suffix 'TFIDF'. We also investigate the sensitivity of the clustering-based method on the choice of the number of clusters. We conduct these sensitivity experiments only on the best approach of the out-domain data (which is the more realistic setup) as reported in Table 2.

We can observe from Figure 3 that our SciBERT$_{\text{CRF}}$-D model consistently outperforms its tf-idf counterpart and the best performing baselines (which do not depend on the number of clusters, hence shown as lines parallel to the x-axis). While the recall of the tf-idf based encoding is comparable to that of SciBERT's, the precision of the tf-idf based approach is substantially lower than its dense counterpart (precision values are also lower than the baselines).

We also observe that the the best results are obtained when $k = 7$. Both precision and recall turn out to be relatively not too sensitive to the choice of k.

CONCLUDING REMARKS.. In this work, we investigated the feasibility of applying factorized models towards extracting emerging novel methodology names from AI research papers. To this end, we propose an ontology-driven and a data-driven factored model for a more fine-grained modeling of the relation between scientific concept names and the contexts that occur around their mentions. Our experiments show that the data-driven factoring via clustering of the dense vectors (SciBERT encoding of the sentences) outperforms the existing baselines for this extraction task. In future, we envision a framework to carry out the data-driven operation in an end-to-end fashion. Additionally, we will expand our work beyond the AI domain to see how the method generalizes to a diverse range of characteristically different domains of study, e.g., economics, physics, etc.

[2] Our source code and dataset are available at https://github.com/Madhu000/ie-dd.git

REFERENCES

[1] Awais Athar and Simone Teufel. 2012. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*. 597–601.

[2] Manjot Bedi, Tanisha Pandey, Sumit Bhatia, and Tanmoy Chakraborty. 2022. Why Did You Not Compare with That? Identifying Papers for Use as Baselines. In *ECIR (1) (Lecture Notes in Computer Science, Vol. 13185)*. Springer, 51–64.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[4] Yixin Cao, Zikun Hu, Tat-seng Chua, Zhiyuan Liu, and Heng Ji. 2019. Low-Resource Name Tagging Learned with Weakly Labeled Data. 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)*. Association for Computational Linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1025

[5] Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*. 152–164.

[6] Arpita Chaudhuri, Monalisa Sarma, and Debasis Samanta. 2022. SHARE: Designing multiple criteria-based personalized research paper recommendation system. *Information Sciences* 617 (2022), 41–64.

[7] Hyejin Cho and Hyunju Lee. 2019. Biomedical named entity recognition using deep neural networks with contextual information. *BMC bioinformatics* 20 (2019), 1–11.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[9] dataset. 2022. Paperswithcodedata. https:https://production-media.paperswithcode.com/about/papers-with-abstracts.json.gz.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] git. 2018. GitFlair. https://github.com/flairNLP/flair.

[12] git. 2020. GitParser. https://github.com/titipata/scipdf_parser.git.

[13] 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*. Association for Computational Linguistics, Florence, Italy, 5203–5213. https://doi.org/10.18653/v1/P19-1513

[14] 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*. Online.

[15] Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2023. Zero-shot Entailment of Leaderboards for Empirical AI Research. *arXiv preprint arXiv:2303.16835* (2023).

[16] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7

[17] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3219–3232. https://doi.org/10.18653/v1/D18-1360

[18] Xue Mengge, Bowen Yu, Zhenyu Zhang, Tingwen Liu, Yue Zhang, and Bin Wang. 2020. Coarse-to-Fine Pre-training for Named Entity Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6345–6354. https://doi.org/10.18653/v1/2020.emnlp-main.514

[19] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, Suntec, Singapore, 1003–1011. https://aclanthology.org/P09-1113

[20] Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional LSTM-CRF for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

[21] Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly Supervised Named Entity Recognition using Positive-Unlabeled Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2409–2419. https://doi.org/10.18653/v1/P19-1231

[22] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[23] PwC. 2023. Paperswithcode. https://paperswithcode.com. Accessed.

[24] Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 148–163.

[25] Stefan Schweter and Alan Akbik. 2020. FLERT: Document-Level Features for Named Entity Recognition. arXiv:2011.06993 [cs.CL]

[26] 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)*. 719–731.

[27] Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2020. Adaptive self-training for few-shot neural sequence labeling. *arXiv preprint arXiv:2010.03680* (2020).

[28] Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised NER with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*. 2159–2169.