On Event Detection in Scientific Papers: A Multi-Domain Dataset

Anonymous ACL submission

Abstract

Given the growing number of scientific papers, automatic information extraction in scientific documents is important for efficient knowledge update and discovery. A key component in scientific papers involves rhetorical activities/events to convey new knowledge and convince readers of the correctness. This work explores a new information extraction problem for scientific documents, aiming to identify event trigger words of rhetorical events/activities, i.e., event detection (ED). To promote future research in this area, we present SciEvent, the first and new dataset for event detection in scientific documents. SciEvent annotates scientific papers of four different domains (i.e., computer science, biology, physics, and mathematics) using 8 popular event types. Our experiments on SciEvent demonstrate the challenges of scientific ED for existing models and call for further research effort in this area. We will publicly release SciEvent to facilitate future research.

1 Introduction

Scientific papers retain most of the knowledge discovery for academia. However, the current exponential growth of scientific literature makes it difficult for even the most experienced reader to keep track of all the ongoing research. Hence, automatic information extraction from scientific documents is helpful for researchers to efficiently comprehend new knowledge and boost scientific discovery.

In natural language processing (NLP), automatic information extraction (IE) has been an ambitious goal for which a wide range of topics have been considered, including summarization (Collins et al., 2017), keyphrase extraction (Gupta and Manning, 2011; Augenstein et al., 2017), entity extraction (Hobbs, 2002; Luan et al., 2017; Jain et al., 2020), relation extraction (Kruiper et al., 2020), clause tagging (Li et al., 2021), and knowledge population (Groth et al., 2018). However, despite extensive progress, prior work on scientific IE has not explored the task of event detection (ED) whose goal is to identify event trigger words of predefined event types in text (Chen et al., 2015). In particular, a scientific paper is a discourse between the authors and readers rather than a simple writing (Li et al., 2021): the authors aim to provide new findings, and more importantly, convince expert readers of the correctness of the presented knowledge. As such, to achieve the persuasion purpose, rhetorical activities/events are prevalent in scientific papers, characterizing different intents such as proposing ideas, reporting results, making claims, or citing issues. Extracting such rhetorical events is thus important to reveal structures of scientific discourses and aid scientific document understanding. Consequently, our goal in this paper is to perform the first study on event detection for rhetorical events in scientific documents. For instance, in the following sentence, an event detection system for scientific documents should be able to predict the words *"apply"*, *"provide"*, and *"comparison"* as event trigger words of the event types *Apply, Claim*, and *Compare*, respectively.

*In this study we **apply** what we hope is a representative set of anomaly detection algorithms and in so doing we are able to **provide** a robust **comparison** of these algorithms.*

In the literature, there have been some studies on recognizing rhetorical activities in scientific documents (De Waard and Maat, 2012; Dernoncourt and Lee, 2017; Huang et al., 2020; Li et al., 2021); however, these works mainly focus on the analysis of discourse at the sentence/clause level, seeking to classify sentences/clauses according to rhetorical components. Our work is different from these prior works in that we consider rhetorical events at the trigger word level that is necessary for scientific document analysis. In particular, as sentences/clauses in scientific documents tend to be long and contain compound information (potentially with more than one event), sentence/clause level modeling might introduce noisy/irrelevant information for rhetorical event representation, thus hindering event analysis and structuring for document understanding. In contrast, extracting event trigger words allows systems to focus on each specific event to facilitate event-centric analysis, e.g., how rhetorical events are expressed in scientific documents or how rhetorical events can be linked to reveal knowledge discovery processes.

A hindrance for event detection research in scientific papers involves the lack of annotated datasets that are necessary for model development and evaluation. This is more pronounced for recent advanced deep learning models for ED (Chen et al., 2015; Cui et al., 2020) where annotated benchmark datasets are critical to measure progress and drive research agenda. To resolve this bottleneck for ED research, our work presents a new dataset for Scientific Event Detection, called **SciEvent**. The dataset is manually annotated by skilled crowd annotators for 8 rhetorical event types in scientific papers. Different from prior datasets for rhetorical components in scientific discourses that mainly focus on biomedical articles (Dernoncourt and Lee, 2017; Huang et al., 2020), SciEvent annotates papers in four different domains, i.e., biology, physics, computer science, and mathematics, to provide a more diverse dataset. Our experiments reveal the challenging nature of SciEvent for ED where the performance of state-of-the-art ED models on SciEvent is far behind those on existing general-domain datasets, thus calling for more research effort in this area. We will publicly release SciEvent to promote scientific ED research.

2 Data Preparation

The documents in SciEvent are collected from `arxiv.org`, focusing on four domains: computer science (CS), quantitative biology, physics, and mathematics. To prepare for the rhetorical event annotation, we design an event taxonomy with 8 event types to capture the most relevant and impactful events in our collected data. As such, we divide the event types into three groups depending on whether they are concerned about current or prior proposed methods/studies.

The first group involves five event types to characterize different aspects of novel methods or studies proposed in the current paper, i.e., **PROPOSE**, **WISDOM**, **APPLY**, **EVALUATE**, and **CLAIM**. In particular, the **PROPOSE** event type involves expressions to describe the novelty or procedure of the proposed methods/studies. An example for this event type is: "*In this Letter, we* ***investigate*** *the feasibility of generating multi-MeV gamma-rays.*"[1]. A **WISDOM** event, on the other hand, presents established wisdom/knowledge that serves as the direct foundation or motivation to develop the proposed methods/studies, e.g., "*It has been demonstrated that humans can* ***perform*** *even one-shot classification.*". In contrast, an **APPLY** event indicates a direct application of existing concrete tools, methods, or systems in the current paper, e.g., "*We* ***leverage*** *BERT to encode sentences.*". In addition, the **EVALUATE** event type captures evaluations of the proposed methods/studies that are usually associated with experiment procedures and performance metrics (e.g., "*One motherset in particular, yeast,* ***has*** *a failure rate approaching 100%*"). Finally, a **CLAIM** event captures an expression of claim for characteristics or achievements of the proposed methods/studies, e.g., "*We* ***achieve*** *state-of-the-art results on the CU-Birds dataset*".

The second event type group has two event types to exclusively represent existing or prior methods/studies which are mentioned in the current paper, including **ISSUE** and **DISCUSS**. As such, an **ISSUE** event expresses a potential issue or gap of existing methods/studies for which the current paper aims to directly avoid/address, e.g., "*Prior methods only* ***consider*** *English data.*". In contrast, **DISCUSS** events cover general discussions or judgements of existing methods that are related but not directly comparable with the proposed methods/studies in the current paper, e.g., sentences in the *Related Work* sections such as "*These methods can* ***perform*** *well in their generalized forms.*". Finally, the third event group contains one event type, i.e., **COMPARE**, to capture comparison expressions between the proposed methods/studies in the current paper and and those in prior works, e.g., "*The proposed method significantly* ***outperforms*** *the baselines.*".

3 Annotation

We recruit 8 annotators from the crowdsourcing platform `upwork.com` to annotate scientific papers in the four domains (i.e., two annotators for each domain). As Upwork provides resumes of annotators, we explicitly select annotators that have demonstrated experience in reading and writing

[1] Event trigger words are underlined in this work

scientific papers in their corresponding domains (e.g., M.S. and Ph.D. students). Detailed annotation guidance with many examples and explanation are provided to train the annotators. Following the practice in prior ED research (Nguyen and Grishman, 2015), we also instruct annotators to select a single most important word to serve as the trigger word for an event mention. Overall, we annotated 115 papers (29 for computer science, 45 for biology, 28 for physics, and 13 for mathematics), amounting to over 22K sentences and 633K words. Our annotators for each domain co-annotate the documents in their domain and achieve Cohen's Kappa scores of 0.76, 0.65, 0.83, and 0.61 on CS, biology, physics, and mathematics, respectively. These inter-agreement scores thus indicate substantial agreements between our annotators. Eventually, the annotators are engaged in discussions to resolve any conflict to produce a final consolidated version of our SciEvent dataset. Table 1 presents some statistics for SciEvent while Figure 1 shows the event type distributions in different domains.

	CS	Biology	Physics	Math
#Event	10,352	18,528	7,730	3,372
#Doc	29	45	28	13
#Sent	5,742	9,825	4,439	2,019
#Token	142K	289K	141K	61K

Table 1: Statistics for SciEvent in different domains.

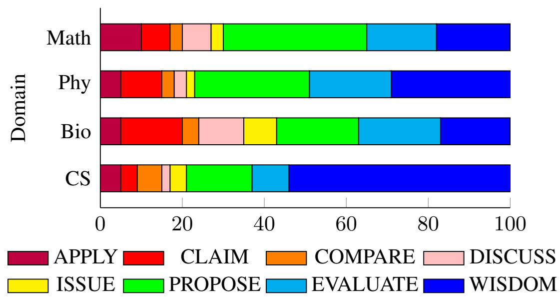

Figure 1: Distributions of event types in SciEvent.

Dataset Challenges: Our dataset reveals several challenges for event detection in scientific papers. One particular challenge involves appropriate coreference resolution of noun phrases/entity mentions to accurately determine event types for trigger words. For example, in the sentence "*This capability greatly **facilitates** experiments related to weak nonlinearity-based quantum computing.*", the event trigger word "*facilitates*" might belong to the CLAIM or DISCUSS event type depending

Word	Count	Total	Rate
be	2,752	10,208	27.0%
use	1,516	1,858	81.6%
show	1,053	1,177	89.5%
have	601	1,534	39.2%
give	286	471	60.7%

Table 2: Event rates of the most frequent triggers.

on whether "*This capacity*" is referring to a characteristics of the currently proposed method (i.e., CLAIM) or prior one (i.e., DISCUSS). As such, solving event detection for scientific papers often requires modeling context beyond sentence boundary to identify accurate references. In addition, domain knowledge and expertise are critical to predict event types for triggers in scientific documents. An example for this challenge can be seen in the sentence "*This approach is only able to **handle** a single domain in the input.*". The event trigger "*handle*" should be assigned to the ISSUE event type if the current paper is directly addressing the single-domain issue in prior work. However, "*handle*" should instead be a DISCUSS event if this sentence is merely contributing to the discussion of relevant works for the current paper. As such, understanding the research topics and directions in the current paper is essential to event type recognition.

To demonstrate the ambiguity in SciEvent, Table 2 presents five words with the highest frequency of being labeled as event triggers (i.e., Count). The table also shows the occurrence frequency of the words (i.e., Total) and the percentages that they are annotated as event triggers (i.e., Rate) in SciEvent. As can be seen, even for the most frequent event trigger words, there is still a probability that they do not trigger/evoke any event of interest. Consequently, capturing the context of the words is crucial to detect event triggers in scientific documents. Finally, we find that sentences in scientific papers often contain multiple event triggers. Among sentences containing at least an event trigger in SciEvent, 51% contains one event trigger, 31% contains 2 event triggers, 12% contains 3 event triggers, and 6% contains more than 3 event triggers. This suggests potential correlations between events at sentence level, which might be helpful to improve ED models for scientific documents.

4 Experiments

To reveal the complexity of the ED task in SciEvent, we evaluate the performance of the state-of-the-art deep learning models for ED. In particular, we consider two major groups of ED models, i.e., sequence-based and graph-based. In the sequence-based approach, we evaluate three typical models: (i) **CNN**: a convolutional neural network for ED (Nguyen and Grishman, 2015), (ii) **DMBERT**: a model that employs dynamic pooling over BERT-based representations (Wang et al., 2019), and (iii) **BERTED**: a model that uses a feed-forward layer over BERT embeddings for representation learning (Yang et al., 2019). For graph-based models, we also consider three recent ED models. In particular, **BERTGCN** is a graph convolutional network (GCN) based on dependency trees (Nguyen and Grishman, 2018), **GatedGCN** augments GCN with trigger-aware gating mechanism for each layer (Lai et al., 2020), and **EEGCN** introduces both dependency labels and structures into GCN for ED (Cui et al., 2020). For all the methods, we leverage SciBERT (Beltagy et al., 2019), a customized BERT model for scientific documents to encode input texts. The Stanza toolkit (Qi et al., 2020) is used for dependency parsing. We also follow the word-classification formulation for ED models as prior work (Nguyen and Grishman, 2015; Chen et al., 2015). As such, to facilitate the experiments, we divide the annotated documents for each domain in SciEvent into three separate portions for training/test/development data. In particular, the numbers of documents in training/test/development data for CS, biology, physics, and math are 24/3/2, 40/3/2, 23/3/2, and 8/3/2 respectively. We fine-tune the hyper-parameters for the models on the development set of CS. Appendix A provides a reproducibility checklist that includes the selected values for the hyper-parameters.

Model	CS	Biology	Physics	Math
CNN	49.8	39.0	52.6	48.1
DMBERT	60.2	40.4	57.2	47.9
BERTED	60.0	39.5	57.1	48.2
BERTGCN	60.0	39.1	54.9	48.9
GatedGCN	59.8	39.1	54.3	47.5
EEGCN	60.3	41.2	56.4	49.9

Table 3: Performance (F1 scores) on the test sets of different domains in SciEvent.

Table 3 presents performance of the models on the test sets of different domains in SciEvent. One observation is that the graph-based models do not exhibit significant improvement over the sequence-based models. We attribute this to the potential noises in the general-domain dependency parser (i.e., Stanza) that cannot achieve its optimal performance on scientific documents, thus hindering the effectiveness of the dependency trees for ED in graph-based models. Importantly, we find that the performance of existing ED models over different domains in SciEvent is far behind those for the general domains (i.e., at least 77% on the popular ACE 2005 dataset (Walker et al., 2006; Lai et al., 2020)). This result suggests new challenges for ED in scientific documents and call for more research effort in this domain. Finally, to further demonstrate the ED challenges in SciEvent, we present an additional experiment for cross-domain evaluation of ED (i.e., trained and tested on different domains) in Appendix B.

5 Related Work

Prior studies on ED involve feature engineering for statistical models (Ahn, 2006; Ji and Grishman, 2008; Hong et al., 2011; Li et al., 2013; Mitamura et al., 2015) and recent deep learning models (Chen et al., 2015; Liu et al., 2017, 2018; Wang et al., 2019). However, such prior work has mainly considered the general domain (i.e., on ACE 2005) that might not be useful for specific domains. Recently, there have been some effort on creating new datasets for ED in specific domains, including biomedical texts (Kim et al., 2009), literary texts (Sims et al., 2019), cybersecurity texts (Satyapanich et al., 2020), and Wikipedia texts (Wang et al., 2020). However, existing ED datasets have not explored rhetorical events in scientific documents as we do. Finally, as discussed in the introduction, scientific document understanding has been studied for different tasks in NLP, including summarization (Teufel and Moens, 2002; AbuRa'ed et al., 2020), keyword extraction (Augenstein et al., 2017), entity/relation extraction (Luan et al., 2017; Jain et al., 2020; Kruiper et al., 2020), and discourse tagging (Dernoncourt and Lee, 2017; Li et al., 2021). However, trigger-based event detection has not been explored for scientific documents.

6 Conclusion

We present SciEvent, the first dataset for ED on scientific documents. SciEvent is manually annotated for 8 popular rhetorical event types in four different domains. Extensive evaluation of state-of-the-art ED models highlights the challenges of SciEvent for ED. In the future, we plan to extend SciEvent to annotate arguments for scientific events.

References

Ahmed AbuRa'ed, Horacio Saggion, and Luis Chiruzzo. 2020. A multi-level annotated corpus of scientific papers for scientific document summarization and cross-document relation discovery. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6672–6679, Marseille, France. European Language Resources Association.

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8.

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

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

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 195–205, Vancouver, Canada. Association for Computational Linguistics.

Shiyao Cui, Bowen Yu, Tingwen Liu, Zhenyu Zhang, Xuebin Wang, and Jinqiao Shi. 2020. Edge-enhanced graph convolution networks for event detection with syntactic relation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2329–2339, Online. Association for Computational Linguistics.

Anita De Waard and Henk Pander Maat. 2012. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 47–55.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Paul Groth, Mike Lauruhn, Antony Scerri, and Ron Daniel Jr. 2018. Open information extraction on scientific text: An evaluation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3414–3423, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Jerry R Hobbs. 2002. Information extraction from biomedical text. *Journal of biomedical informatics*, 35(4):260–264.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1127–1136, Portland, Oregon, USA. Association for Computational Linguistics.

Ting-Hao Kenneth Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C. Lee Giles. 2020. CODA-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

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

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pages 254–262, Columbus, Ohio. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*.

Ruben Kruiper, Julian F.V. Vincent, Jessica Chen-Burger, Marc P.Y. Desmulliez, and Ioannis Konstas. 2020. A scientific information extraction dataset for nature inspired engineering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2078–2085, Marseille, France. European Language Resources Association.

Viet Dac Lai, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Event detection: Gate diversity and syntactic importance scores for graph convolution neural networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5405–5411.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *ACL*.

Xiangci Li, Gully Burns, and Nanyun Peng. 2021. Scientific discourse tagging for evidence extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2550–2562, Online. Association for Computational Linguistics.

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1789–1798, Vancouver, Canada. Association for Computational Linguistics.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Brussels, Belgium. Association for Computational Linguistics.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2641–2651, Copenhagen, Denmark. Association for Computational Linguistics.

Teruko Mitamura, Zhengzhong Liu, and Eduard Hovy. 2015. Overview of tac kbp 2015 event nugget track. In *TAC*.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371.

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Thirty-second AAAI conference on artificial intelligence*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. Casie: Extracting cybersecurity event information from text. In *Proceedings of the AAAI conference on artificial intelligence*.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Simone Teufel and Marc Moens. 2002. Articles summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. In *Technical report, Linguistic Data Consortium*.

Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun, and Peng Li. 2019. Adversarial training for weakly supervised event detection. 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)*, pages 998–1008, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294, Florence, Italy. Association for Computational Linguistics.

A Reproducibility checklist

- **Dataset**: The statistics of the created dataset SciEvent and the annotation process are presented in Section 3. The dataset is included in the submission. We will publicly release the dataset upon the acceptance of the paper. A URL to the publishing site will be included in the paper.

- **Source code with the specification of all dependencies, including external libraries**: We will publicly release the code to run the models upon the acceptance of the paper.

- **Description of computing infrastructure used**: All the experiments were run on a machine with 2 Intel Xeon E5-2620 v4 CPUs, 128GB of RAM, and 4 NVIDIA RTX 2080 Ti GPUs with 11GB RAM. We only train the models with one GPU. The amount of GPU memory for each run ranges from 5 to 7 GB, depending on the models being used.

- **Average runtime for each approach**: We train the models for 100 epochs; each takes approximately 2 minutes. The best epoch is chosen based on the performance on the development sets.

- **Number of parameters in the model**: Every model uses the pre-trained BERT model with non-trainable 110M parameters. The CNN, DMBERT, BERTED, BERTGCN, GatedGCN, EEGCN models have additional 20M, 250K, 80K, 10M, 10M, and 8M trainable parameters, respectively.

- **Explanation of evaluation metrics used, with links to code**: Follow prior work in ED (Nguyen and Grishman, 2015; Chen et al., 2015), we use the precision, recall, and F1 scores for performance metrics.

- **Hyperparameter bounds and configurations for best-performing models**: We use the *allenaiscibert_scivocab_cased* version of BERT in all the considered models (Beltagy et al., 2019). The Stanza toolkit (Qi et al., 2020)[2] (version 1.3.0) is employed for dependency parsing. To obtain the representation vector for a trigger word candidate in a sentence, the hidden vectors of 12 layers of BERT are concatenated. We fine-tune the hyper-parameters for the models in this work over the development data of the computer science domain. For consistency, the same hyper-parameters are applied for the models in other domains of SciEvent. As such, to train the models, we use the Adam optimizer with the learning rate of $2e-5$ (searched in the range of $\{2e-5, 3e-5, 4e-5, 5e-5\}$) and batch size of 128 (searched in the range of $\{32, 64, 128, 256\}$). For the CNN model, we use 4 kernel sizes of $2, 3, 4, 5$, each with 150 filters (searched in the range $\{100, 150, 200, 250, 300\}$. The BERTGCN, GatedGCN, EEGCN models employ two GCN layers (searched in the range $\{2, 3, 4, 5\}$), each with 256 hidden units (searched in the range $\{128, 256, 512\}$). The edge embedding size of the EEGCN model is set to 50. Finally, we use two layers for all the feed-forward neural networks in the models with 256 hidden units in the layers (searched in the range $\{128, 256, 512\}$).

[2] https://stanfordnlp.github.io/stanza/

Model	In-domain	Out-of-domain		
	CS	Biology	Physics	Math
CNN	49.8	19.0	26.3	18.1
DMBERT	60.2	28.8	36.9	29.1
BERTED	60.0	27.3	35.9	29.0
BERTGCN	60.0	27.7	40.7	27.6
GatedGCN	59.8	27.8	39.2	27.9
EEGCN	60.3	28.0	37.9	28.1

Table 4: Performance (F1 scores) in the cross-domain setting. CS is the source domain while Biology, Physics, and Math serve as the target domains.

B Cross-Domain Evaluation

As SciEvent involves four different domains, we further explore the cross-domain evaluation setting where the models are trained on a source domain and evaluated on different domains. In particular, we choose CS as the source domain and treat the others as the target domains. As such, we train the models on the training data portion of CS and evaluate the models on the test data in different domains. Table 4 reports in-domain (i.e., trained and tested on CS) and out-of-domain (i.e., trained on CS and tested on other domains) performance of the models. It is clear from the table that the performance of all the ED models degrades significantly when they are evaluated on new domains (i.e., Biology, Physics, and Math). This clearly demonstrates the divergence between different scientific domains and introduces room for further development of ED research in this setting.

C Annotation Guideline

Table 5 and 6 present a detailed description of event types and examples for each event type in our SciEvent dataset.

Type	Description	Examples
PROPOSE	This event type captures expressions to describe the novelty or procedure of the proposed methods or studies in the papers.	A new method is explicitly introduced: – *We* ***propose*** *prototypical networks for the problem of few-shot classification* – *In this Letter, we* ***investigate*** *the feasibility of generating multi-MeV gamma-rays of several hundreds attoseconds duration ...*
		The description of how the new method works: – *Our model* ***learns*** *a metric space in which classification can be* ***performed*** *by* ***computing*** *distances to prototype representations of each class.*
WISDOM	This event type involves expressions to present established wisdom/knowledge that serves as the direct foundation or motivation to develop the proposed methods/studies in the current paper.	A knowledge that the current paper relies on to motivate its methods/studies: – *While the problem is quite difficult, it has been demonstrated that humans* ***have*** *the ability to perform even one-shot classification.*
		Some benefit of an approach that the current paper follows: – *The concept... has* ***gained*** *popularity because of its flexibility when* ***dealing*** *with complex models and large data sets, in contrast with maximum likelihood estimation.*
APPLY	This event type involves expressions to indicates a direct application of existing concrete tools, methods, or systems in the current pape	An explicit mention: – *The* ***use*** *of episodes makes the training problem more faithful to the test environment.* – *The package solver* ***uses*** *Golang's native data structures and interfaces.*
		An implicit mention: – *The game* ***comes*** *with an built-in package named solver that powers some features of it.*
EVALUATE	This event type involves expressions for the evaluations of the proposed methods/studies that are associated with experiment procedures and performance metrics.	– *One motherset in particular, yeast,* ***has*** *a failure rate approaching 100%, a strong indication that it is a very poor choice for benchmark construction.* – *As each benchmark construction factor has a well defined control group, we* ***compute*** *the mean difference in performance of best algorithm between...*
CLAIM	This event type captures an expression of claim for characteristics or achievements of the proposed methods/studies. This is different from EVALUATE where the experiment and performance details of evaluations are mentioned.	Some achievements of the proposed methods is mentioned: – *We further extend prototypical networks to zero-shot learning and* ***achieve*** *state-of-the-art results on the CU-Birds dataset.* – *We provide an analysis showing that some simple design decisions can* ***yield*** *substantial improvements over recent approaches involving complicated architectural choices and meta-learning.*
		Some characteristics of the proposed methods are mentioned: – *The generated OD's are* ***commensurate*** *with those required for all previous demonstrations of optical modulation using the Rb - PBGF system.* – *This capability greatly* ***facilitates*** *performing experiments related to weak nonlinearity - based quantum computing.*

Table 5: Event types with their descriptions and examples in the SciEvent dataset. Event trigger words are underlined. Continued in Table 6.

Type	Description	Examples
ISSUE	This event type capture expressions to indicate a potential issue or gap of existing methods/studies for which the current paper aims to directly avoid/address.	*– A naive approach, such as re-training the model on the new data, would severely* ***overfit****.* *– However, the RDR regime is only* ***achievable*** *with extremely intense lasers* $\xi \gg 1$*.* *– MCMC methods used for parameter estimation within a model* ***use*** *only ratios of posterior densities, and are therefore unable to* ***measure*** *its normalisation in general.*
DISCUSS	This event type captures expressions for general discussions or judgements of existing methods that are related but not directly comparable with the proposed methods/studies in the current paper. This is in contrast to ISSUE that necessitates a direct connection with the proposed methods/studies.	*– Two recent approaches have* ***made*** *significant progress in few-shot learning.* *– These algorithms* ***focus*** *on finding a feasible solution and do not* ***contain*** *support for preferences (optimization criteria).* *– On the other hand, they can* ***perform*** *well in their generalised forms, but if and only if an adequate reference distribution is used.*
COMPARE	This event type captures expressions to indicate comparisons between the proposed methods/studies in the current paper and and those in prior works.	*–* ***Compared*** *to recent approaches for few-shot learning, they reflect a simpler inductive bias that is beneficial in this limited-data regime ...* *– Our results show that despite the lack of task-specific tuning our model performs surprisingly well,* ***yielding*** *better results than all previously reported models ...*

Table 6: Event types with their descriptions and examples in the SciEvent dataset. Event trigger words are underlined.