Event Detection Explorer: An Interactive Tool for Event Detection Exploration

Wenlong Zhang*
wzhang71@stevens.edu
Stevens Institute of Technology
USA

Bhagyashree Ingale*
bingale@stevens.edu
Stevens Institute of Technology
USA

Hamza Shabir*
hbuch@stevens.edu
Stevens Institute of Technology
USA

Tianyi Li†
li4251@purdue.edu
Purdue University
USA

Tian Shi‡
researchtianshi@gmail.com
Independent Researcher
USA

Ping Wang*
ping.wang@stevens.edu
Stevens Institute of Technology
USA

ABSTRACT

Event Detection (ED) is an important task in natural language processing. In the past few years, many datasets have been introduced for advancing ED machine learning models. However, most of these datasets are under-explored because not many tools are available for people to study events, trigger words, and event mention instances systematically and efficiently. In this paper, we present an interactive and easy-to-use tool, *ED Explorer*, for ED dataset and model exploration. ED Explorer consists of an interactive web application, an API, and an NLP toolkit, which can help both domain experts and non-experts to better understand ED tasks. We use ED Explorer to analyze a recently proposed large-scale ED dataset (referred to as MAVEN). With ED Explorer, we discovered several underlying issues of the dataset, including data sparsity, label bias, label imbalance, and debatable annotations. Such insights are essential for guiding the continuous improvement of existing ED datasets and the advances of ED models. The ED Explorer system[1] and the demonstration video[2] have both been made publicly available.

CCS CONCEPTS

• **Applied computing** → *Annotation*; • **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **Visualization toolkits**.

KEYWORDS

Event Detection, Natural Language Processing, Interactive Tool

ACM Reference Format:
Wenlong Zhang, Bhagyashree Ingale, Hamza Shabir, Tianyi Li, Tian Shi, and Ping Wang. 2023. Event Detection Explorer: An Interactive Tool for Event Detection Exploration. In *28th International Conference on Intelligent User Interfaces (IUI '23 Companion), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3581754.3584178

[1]http://edx.leafnlp.org/
[2]https://www.youtube.com/watch?v=6QPnxPwxg50

IUI '23 Companion, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0107-8/23/03.
https://doi.org/10.1145/3581754.3584178

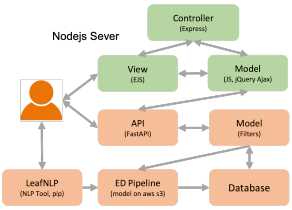

Figure 1: The architecture of the ED Explorer.

1 INTRODUCTION

One crucial element of event understanding is event detection (ED), which detects event triggers from unstructured texts and classifies them into predefined event types [3, 9]. It is one of the most important steps for extracting structured events from unstructured texts [1]. Traditional feature-based models [2, 7, 11] rely on incorporating related features into the models. Many recent deep learning models formulated the ED task as a sequence labeling problem and achieved state-of-the-art results [5, 8, 12, 21, 23]. The advances in deep ED models are attributed to the development of ED datasets for training and benchmarking the models, such as ACE 2005 [17] and TAC KBP [14].

However, these datasets suffer from several limitations [19]. (1) *Data Scarcity*. These datasets are on a small scale and cover a small number of instances. (2) *Low Event Type Coverage*, i.e., only a small number of event types are considered in these datasets. (3) *Label Imbalance*. Not all events are mentioned with equitable frequency in the datasets. In ACE 2005, 60% of event types have less than 100 annotated event mention instances. Recently, a large-scale ED dataset, MAVEN [19], has been introduced with more than 100K event mention instances for 168 event types, which alleviates the data scarcity and low event type Coverage problems. Several other ED datasets are also available [10, 15, 16]. RAMS [6] was originally annotated for document-level argument linking. ALDG [3] and FewEvent [4] are automatically labeled datasets for improving ED models with augmented datasets.

With more ED datasets being introduced and different models being developed [18, 22], the needs to address the following problems become even more pressing: (1) *Uniqueness*. What are the advantages of each ED dataset? (2) *Reliability*. Most ED datasets

are recently introduced without comprehensive validation of their limitations by domain experts. (3) *Accessibility*. Although there are many tools for domain experts to explore ED datasets, it is still difficult for most non-expert users and stakeholders to get access to and understand the ED task.

These problems motivate us to develop *ED Explorer*, a system that can be easily accessed by a broader audience to systematically explore different ED datasets. *ED Explorer* (Figs. 1 and 2) allows both domain experts and non-experts to systematically and efficiently explore different public ED datasets and the models trained on them. There are three toolkits for users: A web application, an API, and an NLP toolkit. The interactive front end of the web application is designed to facilitate users of varied ED expertise to navigate across different event types and trigger words to better understand the datasets and efficiently identify underlying problems in the annotations. There is also a home-maintained and easy-to-use NLP toolkit in Python, *LeafNLP*, for the ED task. Consequently, ED explorer allows users to test the ED models via the integrated and interactive web application, API and LeafNLP.

Triggers	Event Types (# Ins.)	N.T. (# Ins.)
crash	Catastrophe(174), Damaging(4) Motion(2), Attack(2)	153
damage	Damaging(619), Causation(1) Destroying(1), Bodily Harm(1)	275
storm	Catastrophe(925), Attack(14) Self Motion(5), Damaging(1) Motion(1), Bodily Harm(1)	771

(a)

Problems	Instance Examples
Negative Trigger	It formed on October 1 in the Caribbean Sea as the seventeenth tropical **storm::Negative Trigger**, and initially moved slowly to the north.
Trigger Wrong Events	Unknown to the hijackers, passengers aboard **made::Manufacturing** telephone calls to friends and family and relayed information on the hijacking.
Events Ambiguity	**S1:** The hurricane reached peak winds of 125 mph (205 km/h) on October 6 while **moving::Motion** through the Bahamas. **S2:** By midday on June 25, the hurricane reached peak winds of before **moving::Self Motion** inland well south of the U.S. Mexico border.

(b)

Table 1: (a) Examples of trigger words with their annotated event types and frequencies in MAVEN. N.T. represents Negative Triggers that are not annotated. (b) Common annotation problems in MAVEN. The pattern moving::Motion represents the word *moving* triggers an event *Motion*.

2 ED EXPLORER

In this section, we describe the pipelines and usage of our Event Detection (ED) Explorer.

Datasets. Three representative ED datasets are explored. (1) MAVEN [19]: an open-domain and general-purpose data for detecting multiple triggers and events in a single sentence. (2) RAMS [6]: a crowdsourced data for identifying arguments of different roles for an event from multiple sentences. (3) ALDG [3]: an automatically generated data using distant supervision [13]. There are also several other public data, such as CASIE [15] and Commodity News Corpus [10]. Since our platform is designed for public use, we do not include the well-known ACE 2005 and TAC KBP datasets.

Architecture. ED Explorer enables end-users to explore ED datasets and models by interacting with a Web Application and API (see Fig. 1). The Web Application is an HTTP server in Node.js developed following the Model-View-Controller design pattern. For the Controller, we adopt *express* as the primary framework and *express router* to handle routing and user navigation. For View (front-end), we use *EJS* as our template engine to generate HTML and use Bootstrap to style web pages. For Model, they send and receive JSON content via HTTP requests (e.g., GET and POST) to a REST API. The Web API is built with FastAPI, a high-performance Python web framework. Different models, i.e., functions, in FastAPI handle different requests, interact with databases or machine learning pipelines, and respond to the requests.

Following this design, EP Explorer provides three entry points for end-users: (1) Web Application, where users can explore ED datasets and models. (2) Web API, with which users can get processed data and output of ED models. (3) LeafNLP, which users can download and install (via *pip install leafnlp*) for data annotation.

ED Dataset Explorer (EDDE). EDDE helps users understand and explore ED datasets with three primary components (Fig. 2). (1) **Events Overview** presents the distribution of different events and event types and underlying annotation issues.(2) **Event Type Explorer** shows the 10 most frequent trigger words (with the count of their corresponding instances) and all event mention instances (i.e., sentences) for each event type. We use RED and BLUE colors to highlight trigger and negative words, respectively. With this component, end-users can efficiently explore different event types, trigger words, and event mention instances. (3) **Trigger Word Explorer** supports the systematic exploration of trigger words, event types, and event mention instances. This component played an important role in identifying incorrect annotations. For example, we find that most instances of *storm* are annotated as *Catastrophe*, but it is also occasionally labeled as *Attack* and *Self motion*. By manually checking these rare instances, we found many debatable annotations.

ED Model Explorer. We implemented a deep learning model, where input texts are encoded by a BERT encoder [20] and further passed to a randomly initialized two-layer BiLSTM before the classification layer. With the front-end interface, end-users can retrieve annotations by typing an input sentence or article. The annotated trigger words and event types are linked to the Trigger Words Explorer and Event Types Explorer. This integrated interactive system helps end-users better understand the model outputs and the ED datasets.

3 ED DATASETS ANALYSIS

We systematically explored multiple datasets with a focus on MAVEN and observed several **common limitations**.

- **Sparsity**. In the MAVEN training set, there are 50,388 unique candidate trigger words, out of which 7,074 words triggered at least one event. The total number of annotated instances is 96,897. Among the 7,074 trigger words, only 963 (14%) have 20 or more annotated instances. In total, these 963 trigger words cover 75,950 annotated instances (78%). Consequently, most trigger words have very few instances to train ED models.
- **Label Bias**. We observed that most documents are about *military conflict*, *hurricane*, *civilian attack*, and *civil conflict*, which may lead to label bias and limit the applications of trained ED models. For example, for event *Building*, the most common trigger words include *established*, *built*, *building*, *constructed*, and *build*. With the ED Model Explorer, we found that the ED model cannot detect event *Building* in any of these sentences: "We will build a house.", "We will construct a new building.", "We will expand the runway.".
- **Label Imbalance**. For the 7,074 trigger words, we further inspected the events they may trigger. We found that 4,648 words (66%) have triggered only one event in different instances. Regarding the other words that trigger more than one event, 61% of them have dominant events, which results in the label imbalance

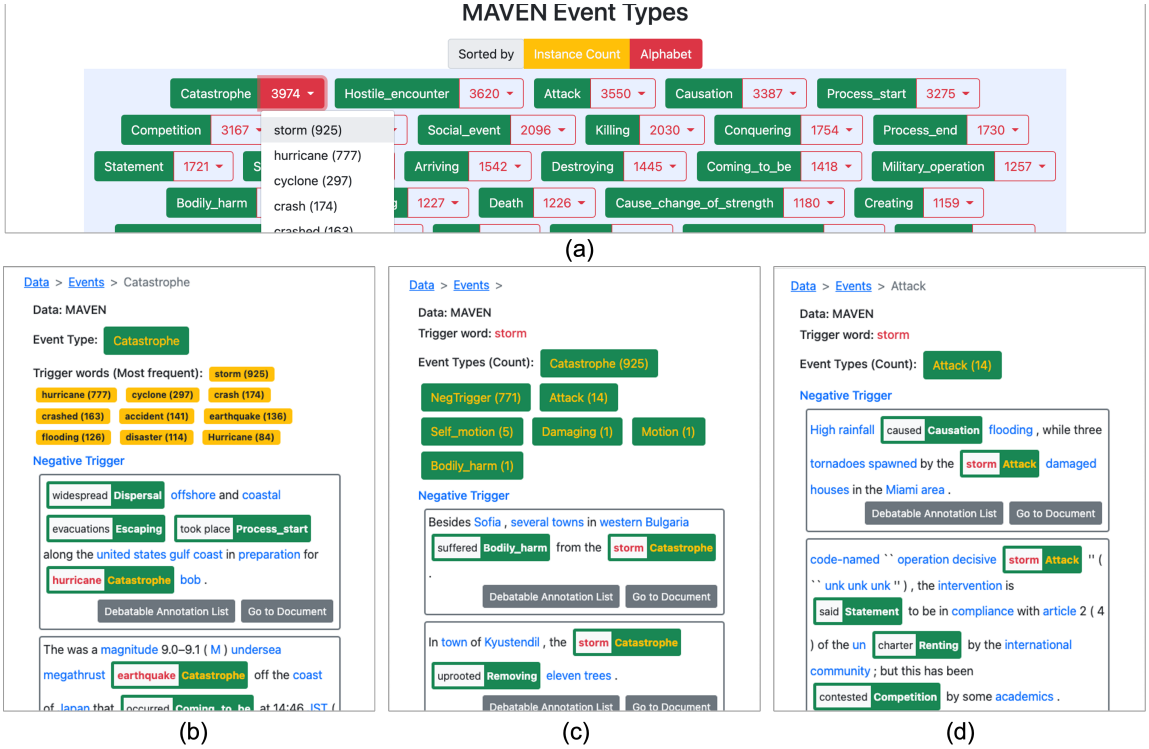

Figure 2: Front-end design of ED Explorer, which includes three primary components, including (a) Events Overview, (b) Event Type Explorer, and (c-d) Trigger Word Explorer.

problem. For example, in Table 1(a), *Catastrophe* is the dominant events for *crash* and *storm*. Accordingly, the ED model trained on MAVEN may suffer from the problem that it predicts only one event for each trigger word despite different scenarios.

Debatable Annotations in MAVEN. We also manually inspected 10,000 annotated instances in MAVEN and found 2,579 debatable instances (25%) which can be grouped into three types. (1) **Negative Trigger** represents the situation where annotating a word triggers an event as a negative trigger. (2) **Trigger Wrong Events** indicates that the word does not trigger the annotated event types. (3) **Events Ambiguity** means that it is difficult to distinguish two event types (such as *Motion* and *Self Motion*). We have shown examples of each of the annotation problems in Table 1(b).

4 CONCLUSION

In this paper, we introduced an event detection exploration tool, ED Explorer, to facilitate a better understanding of the ED task, datasets, and models. ED Explorer consists of an interactive web application, an API, and a LeafNLP toolkit, which allow end users to access datasets and models in a variety of ways. With ED Explorer, we conduct a systematic analysis of a recently developed MAVEN dataset and discover several underlying issues in ED datasets, such as label imbalance and debatable annotations. In the future, we plan to further develop the ED Explorer with more features and address the issues discovered in existing ED datasets.

REFERENCES

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

[2] Jun Araki and Teruko Mitamura. 2015. Joint event trigger identification and event coreference resolution with structured perceptron. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2074–2080.

[3] Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically Labeled Data Generation for Large Scale Event Extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 409–419. https://doi.org/10.18653/v1/P17-1038

[4] Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. 2020. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 151–159.

[5] Ning Ding, Ziran Li, Zhiyuan Liu, Haitao Zheng, and Zibo Lin. 2019. Event detection with trigger-aware lattice neural network. 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)*. 347–356.

[6] Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-Sentence Argument Linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8057–8077.

[7] Prashant Gupta and Heng Ji. 2009. Predicting unknown time arguments based on cross-event propagation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. 369–372.

[8] Viet Dac Lai, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Event detection: Gate diversity and syntactic importance scoresfor graph convolution neural networks. *arXiv preprint arXiv:2010.14123* (2020).

[9] Duong Le and Thien Huu Nguyen. 2021. Fine-grained event trigger detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2745–2752.

[10] Meisin Lee, Lay-Ki Soon, and Eu-Gene Siew. 2021. Effective Use of Graph Convolution Network and Contextual Sub-Tree for Commodity News Event Extraction. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*. 69–81.

[11] Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 73–82.

[12] Jian Liu, Yubo Chen, and Kang Liu. 2019. Exploiting the ground-truth: An adversarial imitation based knowledge distillation approach for event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6754–6761.

[13] Mike Mintz, Steven Bills, Rion Snow, and Dan 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*. 1003–1011.

[14] Teruko Mitamura, Zhengzhong Liu, and Eduard H Hovy. 2015. Overview of TAC KBP 2015 Event Nugget Track.. In *In Proceedings of the Text Analysis Conference (TAC), Gaithersburg, Maryland. National Institute of Standards and Technology (NIST)*.

[15] Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. Casie: Extracting cybersecurity event information from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8749–8757.

[16] 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*. Association for Computational Linguistics, Florence, Italy, 3623–3634. https://doi.org/10.18653/v1/P19-1353

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

[18] Peiyi Wang, Runxin Xun, Tianyu Liu, Damai Dai, Baobao Chang, and Zhifang Sui. 2021. Behind the Scenes: An Exploration of Trigger Biases Problem in Few-Shot Event Classification. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1969–1978.

[19] 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)*. 1652–1671.

[20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[21] Haoran Yan, Xiaolong Jin, Xiangbin Meng, Jiafeng Guo, and Xueqi Cheng. 2019. Event detection with multi-order graph convolution and aggregated attention. 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)*. 5766–5770.

[22] Pengfei Yu, Heng Ji, and Prem Natarajan. 2021. Lifelong Event Detection with Knowledge Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 5278–5290.

[23] Yue Zhao, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng. 2018. Document embedding enhanced event detection with hierarchical and supervised attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 414–419.