MAVEN: A Massive General Domain Event Detection Dataset

Xiaozhi Wang[1], **Ziqi Wang**[1], **Xu Han**[1], **Wangyi Jiang**[1], **Rong Han**[1],
Zhiyuan Liu[1,2*], **Juanzi Li**[1,2], **Peng Li**[3], **Yankai Lin**[3], **Jie Zhou**[3]
[1]Department of Computer Science and Technology, BNRist;
[2]KIRC, Institute for Artificial Intelligence,
Tsinghua University, Beijing, 100084, China
[3]Pattern Recognition Center, WeChat AI, Tencent Inc, China
{wangxz20, ziqi-wan16, hanxu17}@mails.tsinghua.edu.cn

Abstract

Event detection (ED), which means identifying event trigger words and classifying event types, is the first and most fundamental step for extracting event knowledge from plain text. Most existing datasets exhibit the following issues that limit further development of ED: (1) **Data scarcity**. Existing small-scale datasets are not sufficient for training and stably benchmarking increasingly sophisticated modern neural methods. (2) **Low coverage**. Limited event types of existing datasets cannot well cover general-domain events, which restricts the applications of ED models. To alleviate these problems, we present a MAssive eVENt detection dataset (MAVEN), which contains 4,480 Wikipedia documents, 118,732 event mention instances, and 168 event types. MAVEN alleviates the data scarcity problem and covers much more general event types. We reproduce the recent state-of-the-art ED models and conduct a thorough evaluation on MAVEN. The experimental results show that existing ED methods cannot achieve promising results on MAVEN as on the small datasets, which suggests that ED in the real world remains a challenging task and requires further research efforts. We also discuss further directions for general domain ED with empirical analyses. The source code and dataset can be obtained from https://github.com/THU-KEG/MAVEN-dataset.

1 Introduction

Event detection (ED) is an important task of information extraction, which aims to identify event triggers (the words or phrases evoking events in text) and classify event types. For instance, in the sentence "*Bill Gates* `founded` *Microsoft in 1975*", an ED model should recognize that the word "`founded`" is the trigger of a `Found` event. ED

* Corresponding author: Z.Liu (liuzy@tsinghua.edu.cn)

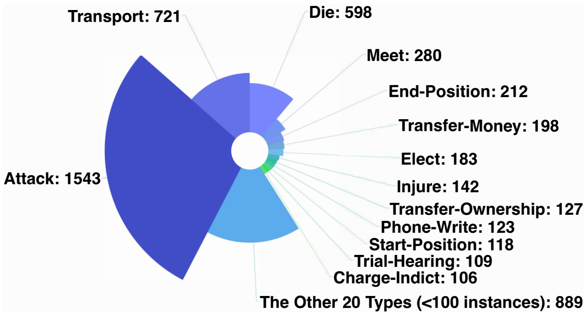

Figure 1: Data distribution of the most widely-used ACE 2005 English dataset. It contains 33 event types, 599 documents and 5,349 instances in total.

is the first stage to extract event knowledge from text (Ahn, 2006) and also fundamental to various NLP applications (Yang et al., 2003; Basile et al., 2014; Cheng and Erk, 2018; Yang et al., 2019).

Due to the rising requirement of event understanding, many efforts have been devoted to ED in recent years. The advanced models have been continuously proposed, including the feature-based models (Ji and Grishman, 2008; Gupta and Ji, 2009; Li et al., 2013; Araki and Mitamura, 2015) and advanced neural models (Chen et al., 2015; Nguyen and Grishman, 2015; Nguyen et al., 2016; Feng et al., 2016; Ghaeini et al., 2016; Liu et al., 2017; Zhao et al., 2018; Chen et al., 2018; Ding et al., 2019; Yan et al., 2019). Nevertheless, the benchmark datasets for ED are upgraded slowly. As event annotation is complex and expensive, the existing datasets are mostly small-scale. As shown in Figure 1, the most widely-used ACE 2005 English dataset (Walker et al., 2006) only contains 599 documents and 5,349 annotated instances. Due to the inherent data imbalance problem, 20 of its 33 event types only have fewer than 100 annotated instances. As recent neural methods are typically data-hungry, these small-scale datasets are not sufficient for training and stably benchmarking mod-

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1652–1671,
November 16–20, 2020.

ern sophisticated models. Moreover, the covered event types in existing datasets are limited. The ACE 2005 English dataset only contains 8 event types and 33 specific subtypes. The Rich ERE ontology (Song et al., 2015) used by TAC KBP challenges (Ellis et al., 2015, 2016) covers 9 event types and 38 subtypes. The coverage of these datasets is low for general domain events, which results in the models trained on these datasets cannot be easily transferred and applied on general applications.

Recent research (Huang et al., 2016; Chen et al., 2017) has shown that the existing datasets suffering from the data scarcity and low coverage problems are now inadequate for benchmarking emerging methods, i.e., the evaluation results are difficult to reflect the effectiveness of novel methods. To tackle these issues, some works adopt the distantly supervised methods (Mintz et al., 2009) to automatically annotate data with existing event facts in knowledge bases (Chen et al., 2017; Zeng et al., 2018; Araki and Mitamura, 2018) or use bootstrapping methods to generate new data (Ferguson et al., 2018; Wang et al., 2019b). However, the generated data are inevitably noisy and homogeneous due to the limited number and low diversity of event facts and seed data instances.

In this paper, we present MAVEN, a human-annotated massive general domain event detection dataset constructed from English Wikipedia and FrameNet (Baker et al., 1998), which can alleviate the data scarcity and low coverage problems:

(1) Our MAVEN dataset contains $111,611$ different events, $118,732$ event mentions, which is twenty times larger than the most widely-used ACE 2005 dataset, and $4,480$ annotated documents in total. To the best of our knowledge, this is the largest human-annotated event detection dataset until now.

(2) MAVEN contains 168 event types, which covers a much broader range of general domain events. These event types are manually derived from the frames defined in the linguistic resource FrameNet (Baker et al., 1998), which has been shown to have good coverage of general event semantics (Aguilar et al., 2014; Huang et al., 2018). Furthermore, we construct a tree-structure hierarchical event type schema, which not only maintains the good coverage of FrameNet but also avoids the difficulty of crowd-sourced annotation caused by the original sophisticated schema, and may help future ED models with the hierarchy information.

We reproduce some recent state-of-the-art ED models and conduct a thorough evaluation of these models on MAVEN. From the experimental results, we observe significant performance drops of these models as compared with on existing ED benchmarks. It indicates that detecting general-domain events is still challenging and the existing datasets are difficult to support further explorations. We also explore some promising directions with empirical analyses, including modeling the multiple events shown in one sentence, using the hierarchical event schema to handle long-tail types and distinguish close types, and improving low-resource ED tasks with transfer learning. We hope that all contents of MAVEN could encourage the community to make further breakthroughs.

2 Event Detection Definition

In our dataset, we mostly follow the settings and terminologies defined in the ACE 2005 program (Doddington et al., 2004). We specify the vital terminologies as follows:

An **event** is a specific occurrence involving participants (Consortium, 2005). In MAVEN, we mainly focus on extracting the basic events that can be specified in one or a few sentences. Each event will be labeled with a certain **event type**. An **event mention** is a sentence within which the event is described. As the same event may be mentioned multiple times in a document, there are typically more event mentions than events. An **event trigger** is the key word or phrase in an event mention that most clearly expresses the event occurrence.

The ED task is to identify event triggers and classify event types for given sentences. Accordingly, ED is conventionally divided into two subtasks: Trigger Identification and Trigger Classification (Ahn, 2006). **Trigger identification** is to identify the annotated triggers from all possible candidates. **Trigger classification** is to classify the corresponding event types for the identified triggers. Both the subtasks are evaluated with micro precision, recall, and F-1 scores. Recent neural methods typically formulate ED as a token-level multi-class classification task (Chen et al., 2015; Nguyen et al., 2016) or a sequence labeling task (Chen et al., 2018; Zeng et al., 2018), and only report the trigger classification results (add an additional type `N/A` to be classified at the same time, indicating that the candidate is not a trigger). In MAVEN, we inherit all the above-mentioned settings in both dataset construction and model evaluation.

3 Data Collection of MAVEN

3.1 Event Schema Construction

The event schema used by the existing ED datasets like ACE (Doddington et al., 2004), Light ERE (Aguilar et al., 2014) and Rich ERE (Song et al., 2015) only includes limited event types (e.g. `Movement`, `Contact`, etc). Hence, we need to construct a new event schema with a good coverage of general-domain events for our dataset.

Inspired by Aguilar et al. (2014), we mostly use the frames in FrameNet (Baker et al., 1998) as our event types for a good coverage. FrameNet follows the frame semantic theory (Fillmore, 1976, 2006) and defines over $1,200$ semantic frames along with corresponding frame elements, frame relations, and lexical units. From the ED perspective, some frames and lexical units can be used as event types and triggers respectively.

Considering FrameNet is primarily a linguistic resource constructed by linguistic experts, it prioritizes lexicographic and linguistic completeness over ease of annotation (Aguilar et al., 2014). To facilitate the crowd-sourced annotation with large numbers of annotators, we simplify the original frame schema into our event schema. We collect 598 event-related frames from FrameNet by recursively selecting the frames having "Inheritance", "Subframe" or "Using" relations with the `Event` frame like Li et al. (2019). Then we manually filter out abstractive frames (e.g. `Process_resume`), merge similar frames (e.g. `Choosing` and `Adopt_selection`), and assemble too fine-grained frames into more generalized frames (e.g. `Visitor_arrival` and `Drop_in_on` into `Arriving`). We finally get 168 event types to annotate, covering 74.4% (selected or inherit from the selected frames) of the 598 event-related frames, and the mapping between event types and frames are shown in Appendix D.

Based on the FrameNet inheritance relation and the HowNet event schema (Dong and Dong, 2003), we organize the event types into a tree-structure **hierarchical event type schema**. During annotation, we ask the annotators to label the triggers with the most fine-grained type (e.g. `Theft` and `Robbery`). The coarse-grained types (e.g. `Committing_crime`) are only used for those rare events without appropriate fine-grained types so that to recall more events with fewer labels. Appendix C shows the overall hierarchical schema.

Topic	#Documents	Percentage
Military conflict	$1,458$	32.5%
Hurricane	480	10.7%
Civilian attack	287	6.4%
Concert tour	255	5.7%
Music festival	170	3.8%
Total	$2,650$	59.2%

Table 1: Count and % of MAVEN documents in Top-5 EventWiki (Ge et al., 2018) topics.

3.2 Document Selection

To support the annotation, we need a large number of informative documents as our basic corpus. We adopt English Wikipedia as our data source considering it is informative and widely-used (Rajpurkar et al., 2016; Yang et al., 2018). Meanwhile, Wikipedia articles contain rich entities, which will benefit event argument annotation in the future.

To effectively select the articles containing enough events, we follow a simple intuition that the articles describing grand "topic events" may contain much more basic events than the articles about specific entity definitions. We adopt EventWiki (Ge et al., 2018) to help select the event-related articles. It is a knowledge base for major events and each major event is described with a Wikipedia article. We thus utilize the articles indexed by EventWiki as the base and manually select some articles to annotate their basic events covered by our event schema. To ensure the quality of articles, we follow the previous settings (Yao et al., 2019) to use the introductory sections for annotation. Moreover, we filter out the articles with fewer than 5 sentences or fewer than 10 event-related frames labeled by a semantic labeling tool (Swayamdipta et al., 2017).

Finally, we select $4,480$ documents in total, covering 90 of the 95 major event topics defined in EventWiki. Table 1 shows the top 5 EventWiki topics of our selected documents.

3.3 Candidate Selection and Automatic Labeling

We have massive data to be annotated with 168 event types. To facilitate efficiency and improve consistency of our annotators, who are not all linguistic experts, we adopt some heuristic methods to narrow down trigger candidates and the corresponding event type candidates, and automatically label some triggers to provide information.

	Dataset	#Documents	#Tokens	#Sentences	#Event Types	#Events	#Event Mentions
	ACE 2005	599	303k	15,789	33	4,090	5,349
Rich ERE	LDC2015E29	91	43k	1,903	38	1,439	2,196
	LDC2015E68	197	164k	8,711	37	2,650	3,567
	LDC2015E78	171	114k	4,979	31	2,285	2,933
	TAC KBP 2014	351	282k	14,852	34	10,719	10,719
	TAC KBP 2015	360	238k	11,535	38	7,460	12,976
	TAC KBP 2016	169	109k	5,295	18	3,191	4,155
	TAC KBP 2017	167	99k	4,839	18	2,963	4,375
	Total	1,272	854k	41,708	38	29,293	38,853
	MAVEN	**4,480**	**1,276k**	**49,873**	**168**	**111,611**	**118,732**

Table 2: Statistics of MAVEN compared with existing widely-used ED datasets. The #Event Type shows the number of the most fine-grained types (i.e. the "subtype" of ACE and ERE). For the multilingual datasets, we report the statistics of the English subset (typically the largest subset) for direct comparisons to MAVEN. We merge all the Rich ERE datasets and remove the duplicate documents to get the "Total" statistics.

Candidate selection We first do POS tagging with the NLTK toolkit (Bird, 2006), and select the content words (nouns, verbs, adjectives, and adverbs) as the trigger candidates to be annotated. As event triggers can also be phrases, the phrases in documents that can be matched with the phrases provided in FrameNet are also selected as trigger candidates. For each trigger candidate, we provide 15 event types as label candidates. The 15 type candidates are automatically recommended with the cosine similarities between trigger word embeddings and the average of the word embeddings of event types' corresponding lexical units in FrameNet. The word embeddings we used here are the pre-trained Glove (Pennington et al., 2014) word vectors. To verify the effectiveness of these candidate selection methods, we randomly choose 50 documents and invite an expert to directly label all the words with the 168 event types. The results show that 100% of the expert-provided labeled triggers appeared among the automatically listed trigger candidates provided to annotators. Furthermore, the results also show that 96.8% of the expert-provided event types appeared among the 15 event type candidates automatically recommended to the annotators.

Automatic labeling We label some trigger candidates with a state-of-the-art frame semantic parser (Swayamdipta et al., 2017) and use the corresponding event types of the predicted frames as the default event types. The annotators can replace them with more appropriate event types or just keep them to save time and effort. Evaluated on the final dataset, the frame semantic parser can achieve 52.4% precision and 49.7% recall, which indicates that the automatic labeling process can help to save about a half of the overall annotation effort.

3.4 Human Annotation

The final step requires the annotators to label the trigger candidates with appropriate event types and merge the event mentions (annotate which mentions are expressing the same event).

Annotation process As the event annotation is complicated, to ensure the accuracy and consistency of our annotation, we follow the ACE 2005 annotation process (Consortium, 2005) to organize a two-stage iterative annotation. In the first stage, 121 crowd-source annotators are invited to annotate the documents given the default results and candidate sets described in the last section. Each document is annotated twice by two independent annotators in this stage. In the second stage, 17 experienced annotators and experts will give the final results on top of the annotation results of the two first-stage annotators. Each document will be annotated only once in the second stage.

Data quality To evaluate the dataset quality, we randomly sample 1,000 documents and invite different second-stage annotators to independently annotate these documents for one more time. We measure the inter-annotator agreements of the event type annotation between two annotators with Cohen's Kappa (Cohen, 1960). The results for the first stage trigger and type annotation are 38.2% and 42.7%, respectively. And the results for the second stage trigger and type annotation are 64.1% and 73.7%. One of the authors also manually examined

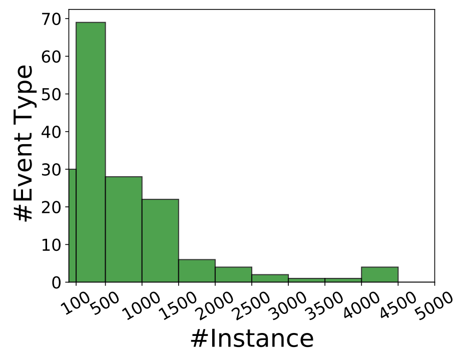

Figure 2: Distribution of MAVEN event types by their instance numbers.

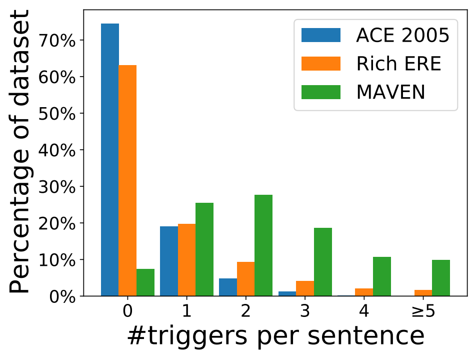

Figure 3: Distribution of sentences containing different numbers of (golden) triggers of three datasets.

Top-level Event Type	Subtype Examples	Percentage
Action	Telling, Attack, Building	46.9%
Change	Change_event_time, Change_of_leadership	27.5%
Scenario	Emergency, Catastrophe, Incident	13.4%
Sentiment	Supporting, Convincing, Quarreling	6.4%
Possession	Commerce_buy, Giving, Renting	5.7%

Table 3: Five top-level event types and their percentages of MAVEN. Appendix C shows more details.

50 random documents. The estimated accuracies of event type annotation and event mention merging are 90.1% and 86.0% respectively. These results show that although the general domain event annotation is difficult (the first-stage inter-agreement is low), MAVEN's quality is satisfactory.

4 Data Analysis of MAVEN

4.1 Data Size

We show the main statistics of MAVEN and compare them with some existing widely-used ED datasets in Table 2, including the most widely-used ACE 2005 dataset (Walker et al., 2006) and a series of Rich ERE annotation datasets provided by TAC KBP competition, which are DEFT Rich ERE English Training Annotation V2 (LDC2015E29), DEFT Rich ERE English Training Annotation R2 V2 (LDC2015E68), DEFT Rich ERE Chinese and English Parallel Annotation V2 (LDC2015E78), TAC KBP Event Nugget Data 2014-2016 (LDC2017E02) (Ellis et al., 2014, 2015, 2016) and TAC KBP 2017 (LDC2017E55) (Getman et al., 2017). The Rich ERE datasets can be combined as used in Lin et al. (2019) and Lu et al. (2019), but the combined dataset is still much smaller than MAVEN. MAVEN is larger than all existing ED datasets, especially in the number of events. Hopefully, the large-scale dataset can accelerate the research on general domain ED.

4.2 Data Distribution

Figure 2 shows the histogram of MAVEN event types by their instance numbers. We can observe that the inherent data imbalance problem also exists in MAVEN. However, as MAVEN is large-scale, 41% and 82% event types have more than 500 and 100 instances respectively. Compared with existing datasets like ACE 2005 (only 39% event types have more than 100 instances), MAVEN significantly alleviates the data scarcity problem, which will benefit developing strong ED models and various event-related downstream applications.

We want MAVEN to serve as a real-world ED dataset, and the distribution of real-world data is inherently long-tail. To evaluate the ED ability on the long-tail scenario is also our goal. Hence, we do not apply data augmentation or balancing during dataset construction and maintain the real-world distribution in MAVEN. To support future exploration of handling the long-tail problem, we design a hierarchical event type schema, which may help transfer knowledge from the coarse-grained event types to the long-tail fine-grained types. We show the five top-level (most coarse-grained) types and their proportions in Table 3 and the detailed hierarchical schema in Appendix C.

4.3 Multiple Events in One Sentence

A key phenomenon of ED datasets is that a sentence can express multiple events at the same time, and ED models will better classify the event types with the help of correlations between multiple events. Although the multiple event phenomenon has been investigated by existing works (Li et al., 2013;

Subset	#Document	#Event	#Mention	#Negative.
Train	2,913	73,496	77,993	323,992
Dev	710	17,726	18,904	79,699
Test	857	20,389	21,835	93,570

Table 4: The statistics of splitting MAVEN. "#Negative." is the number of negative instances.

Chen et al., 2018; Liu et al., 2018) on ACE 2005 dataset, we observe that this phenomenon is much more common and complex on MAVEN.

In Figure 3, we compare MAVEN's percentages of sentences containing different numbers of triggers with ACE 2005 and the combined Rich ERE dataset (corresponding to the "Total" row in Table 2). We can observe that because MAVEN's coverage on general domain events is much higher, the multiple events in one sentence phenomenon is much more common in MAVEN than existing datasets. Moreover, as more event types are defined in MAVEN, the association relations between event types will be much more complex than on ACE 2005. We hope MAVEN can facilitate ED research on modeling multiple event correlations.

5 Experiments

Our experiments and analyses will show the challenges of MAVEN and promising ED directions.

5.1 Benchmark Setting

We firstly introduce the MAVEN benchmark setting here. MAVEN is randomly split into training, development, and test sets and the statistics of the three sets are shown in Table 4. After splitting, there are 32% and 71% of event types that have more than 500 and 100 training instances respectively, which ensures the models can be well-trained.

Conventionally, the existing ED datasets only provide the standard annotation of positive instances (the annotated event triggers) and researchers will sample the negative instances (non-trigger words or phrases) by themselves, which may lead to potential unfair comparisons between different methods. In MAVEN, we provide official negative instances to ensure fair comparisons. As described in Section 3.3, the negative instances are the content words labeled by the NLTK POS tagger or the phrases which can be matched with the FrameNet lexical units. In other words, we only filter out those empty words, which will not influence the application of models developed on MAVEN.

5.2 Experimental Setting

Models Recently, various neural models have been developed for ED and achieved superior performances compared with traditional feature-based models. Hence, we reproduce six representative state-of-the-art neural models and report their performances on both MAVEN and widely-used ACE 2005 to assess the challenges of MAVEN, including: (1) **DMCNN** (Chen et al., 2015) is a convolutional neural network (CNN) model, which leverages a CNN to automatically learn sequence representations and a dynamic multi-pooling mechanism to aggregate learned features into trigger-specific representations for classification. (2) **BiLSTM** (Hochreiter and Schmidhuber, 1997) is a vanilla recurrent neural network baseline, which adopts the widely-used bi-directional long short-term memory network to learn textual representations, and then uses the hidden states at the positions of trigger candidates for classifying event types. (3) **MOGANED** (Yan et al., 2019) is an advanced graph neural network (GNN) model. It proposes a multi-order graph attention network to effectively model the multi-order syntactic relations in dependency trees and improve ED. (4) **DMBERT** (Wang et al., 2019b) is a vanilla BERT-based model. It takes advantage of the effective pre-trained language representation model BERT (Devlin et al., 2019) and also adopts the dynamic multi-pooling mechanism to aggregate features for ED. We use the $\text{BERT}_{\text{BASE}}$ architecture in our experiments. (5) Different from the above token-level classification models, **BiLSTM+CRF** and **BERT+CRF** are sequence labeling models. To verify the effectiveness of modeling multiple event correlations, the two models both adopt the conditional random field (CRF) (Lafferty et al., 2001) as their output layers, which can model structured output dependencies. And they use BiLSTM and $\text{BERT}_{\text{BASE}}$ as their feature extractors respectively.

As we manually tune hyperparameters and some training details, the results of reproduced models may be different from the original papers. Please refer to Appendix A for reproduction details.

Evaluation Following the widely-used setting introduced in Section 2, we report the micro precision, recall, and F-1 scores for trigger classification as our evaluation metrics. For direct comparisons with the token-level classification models, we use span-based metrics for the sequence labeling base-

Method	ACE 2005			MAVEN		
	P	R	F-1	P	R	F-1
DMCNN	73.7 ± 2.42	63.3 ± 3.30	68.0 ± 1.95	$\mathbf{66.3 \pm 0.89}$	55.9 ± 0.50	60.6 ± 0.20
BiLSTM	71.7 ± 1.70	$\mathbf{82.8 \pm 1.00}$	$\mathbf{76.8 \pm 1.01}$	59.8 ± 0.81	67.0 ± 0.76	62.8 ± 0.82
BiLSTM+CRF	$\mathbf{77.2 \pm 2.08}$	74.9 ± 2.62	75.4 ± 1.64	63.4 ± 0.70	64.8 ± 0.69	64.1 ± 0.13
MOGANED	70.4 ± 1.38	73.9 ± 2.24	72.1 ± 0.39	63.4 ± 0.88	64.1 ± 0.90	63.8 ± 0.18
DMBERT	70.2 ± 1.71	78.9 ± 1.64	74.3 ± 0.81	62.7 ± 1.01	$\mathbf{72.3 \pm 1.03}$	67.1 ± 0.41
BERT+CRF	71.3 ± 1.77	77.1 ± 1.99	74.1 ± 1.56	65.0 ± 0.84	70.9 ± 0.94	$\mathbf{67.8 \pm 0.15}$

Table 5: The overall trigger classification performance of various models on ACE 2005 and MAVEN.

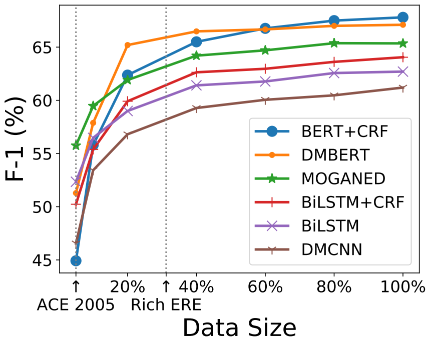

Figure 4: Model performance (F-1) change along with the training data size.

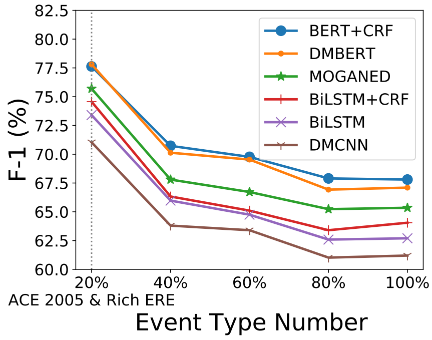

Figure 5: Model performance (F-1) change along with the number of event types.

lines. On ACE 2005[1], we use 40 newswire articles for test, 30 random documents for development, and 529 documents for training following previous work (Chen et al., 2015; Wang et al., 2019c), and sample all the unlabeled words as negative instances. To get stable results, we run each model 10 times on both datasets and report the averages and standard deviations for each metric.

5.3 Overall Experimental Results

The overall experimental results are in Table 5, from which we have the following observations:

(1) Although the models perform well on ACE 2005, their performances are significantly lower and not satisfying on MAVEN. It indicates that our MAVEN is challenging and the general domain ED still needs more research efforts. (2) The result deviations of various models on MAVEN are typically significantly lower than on the small-scale ACE 2005, which suggests that the small-scale datasets cannot stably benchmark sophisticated ED methods, while MAVEN alleviates this problem with its massive annotated data. (3) It is surprising to find that the BiLSTM-based models achieve remarkably high performance on ACE 2005, even outperform the BERT models. We guess this is because the small-scale dataset cannot stably train and benchmark large models. The results on MAVEN are intuitive. (4) From the comparison between the **BiLSTM+CRF** and **BiLSTM**, we can observe that the CRF-based method achieves obvious improvement on MAVEN, but cannot outperform the vanilla **BiLSTM** on ACE 2005. **BERT+CRF** also outperforms **DMBERT** on MAVEN even without the effective dynamic multi-pooling mechanism. Considering the key advantage of the CRF output layer in ED is to model multiple event correlations, the results are consistent with our observations in Section 4.3 that the multiple events in one sentence phenomenon is much more common in MAVEN. This suggests how to better modeling multiple events is worth exploring.

5.4 Analyses on Data Size and #Event Types

MAVEN contains more data and covers more event types compared with existing benchmarks. In this section, we analyze the benefits of a larger data scale and the challenge of more event types.

We randomly choose different proportions of documents from the MAVEN training set and compare the model performances trained with different sizes of data in Figure 4. We can observe that

[1] catalog.ldc.upenn.edu/LDC2006T06

Method	ACE 2005 Trigger Classification		
	P	R	F-1
DMBERT	70.2 ± 1.71	$\mathbf{78.9 \pm 1.64}$	74.3 ± 0.81
+aug	68.7 ± 1.21	76.4 ± 1.16	72.4 ± 0.75
+pretrain	$\mathbf{71.9 \pm 1.12}$	78.7 ± 1.44	$\mathbf{75.1 \pm 0.56}$

Table 6: The performance of DMBERT with two simple knowledge transfer methods on ACE 2005.

MAVEN can sufficiently train the models and stably benchmark them, and we will get unreliable comparison results at the existing datasets' scale.

We also randomly choose different proportions of event types and train the models to only classify the chosen types. The model performances are shown in Figure 5. With the increase in the number of event types, we can observe significant performance drops, which demonstrates the challenge brought by the high coverage of MAVEN.

5.5 Analyses on Transferability

As MAVEN annotates a large range of general domain events, an intuitive question is whether the general ED knowledge learned on MAVEN can transfer to other ED tasks that do not have sufficient data. We examine the transferability of MAVEN with experiments on ACE 2005.

We explore two simple transfer learning methods on **DMBERT** model. (1) Data augmentation (**+aug**) is to add $18,729$ MAVEN instances into ACE 2005 training set and directly train the model. As the event schema of ACE 2005 and MAVEN is different, we manually build an incomplete mapping of event types, which is shown in Appendix B. (2) Intermediate pre-training (**+pretrain**), which is to first train the model on MAVEN and then finetune it on ACE 2005. This method has been shown to be effective on some natural language inference tasks (Wang et al., 2019a).

The results are shown in Table 6, from which we can observe that as MAVEN focuses on different event types and a different text domain (Wikipedia), direct data augmentation harms ED performances while tested on ACE 2005 (newswire data). However, intermediate pre-training can improve ED on ACE 2005 with the general event knowledge learned on MAVEN, which indicates MAVEN's high coverage of event types can benefit other ED tasks. It is worth to explore how to apply more advanced transfer learning methods to improve the performance on low-resource ED scenarios.

Method	Identification Mistakes		Event Type Mistakes		
	FP	FN	Parent -Children	Between Siblings	Into Top 50%
DMCNN	27.3%	55.9%	15.5%	19.8%	89.2%
BiLSTM	26.9%	52.9%	14.5%	14.6%	90.3%
MOGANED	44.5%	31.3%	15.5%	17.8%	86.8%
DMBERT	48.5%	27.2%	13.1%	19.0%	87.0%

Table 7: The proportions of different kinds of mistakes in various models' predictions on MAVEN dev set. The numbers of positive and negative instances are $18,904$ and $79,699$, respectively.

5.6 Error Analysis

To analyze the abilities required by MAVEN, we conduct error analyses on the prediction results of various token-level classification ED models (the sequence labeling methods have span prediction errors, hence cannot be analyzed with misclassifying types as here). The results are shown in Table 7, from which we can observe:

(1) "Identification Mistakes" indicates misclassifying negative instances into positive types (FP) or misclassifying positive instances into N/A (FN), which is the most common mistake. It indicates that identifying event semantics from various and complicated language expressions is still challenging and needs further efforts.

(2) "Event Type Mistakes" indicates misclassifying between the 168 event types. The percentages of the three subtype mistakes are all calculated within "Event Type Mistakes". "Parent-Children" indicates misclassifying instances into their parent or children types in the tree-structure hierarchical event type schema, and "Between Siblings" indicates misclassifying instances into their sibling types. Considering each event type only has one parent type and 9.96 sibling types on average, the percentages of these two kinds of mistakes are significantly higher than misclassifying into other distant types. It suggests that existing models typically cannot well distinguish subtle differences between event types, and our hierarchical event type schema may help models to this point.

(3) "Into Top 50%" indicates misclassifying into event types with top 50% amounts of data. It shows that ED models should develop the ability to resist the influence of the inherent data imbalance problem. Hence, further explorations on handling these problems may bring more effective ED models. To this end, our hierarchical event schema may also be helpful in developing data balancing and data augmentation methods.

6 Related Work

As stated in Section 2, we follow the ED task definition specified in the ACE challenges, especially the ACE 2005 dataset (Doddington et al., 2004) in this paper, which requires ED models to generally detect the event triggers and classify them into specific event types. The ACE event schema is simplified into Light ERE and further extended to Rich ERE (Song et al., 2015) to cover more but still a limited number of event types. Rich ERE is used to create various datasets and the TAC KBP challenges (Ellis et al., 2014, 2015, 2016; Getman et al., 2017). Nowadays, the majority of ED and event extraction models (Ji and Grishman, 2008; Li et al., 2013; Chen et al., 2015; Feng et al., 2016; Liu et al., 2017; Zhao et al., 2018; Yan et al., 2019) are developed on these datasets. Our MAVEN follows the effective framework and extends it to numerous general domain event types and data instances.

There are also various datasets defining the ED task in different ways. The early MUC series datasets (Grishman and Sundheim, 1996) define event extraction as a slot-filling task. The TDT corpus (Allan, 2012) and some recent datasets (Minard et al., 2016; Araki and Mitamura, 2018; Sims et al., 2019; Liu et al., 2019) follow the open-domain paradigm, which does not require models to classify events into pre-defined event types for better coverage but limits the downstream application of the extracted events. Some datasets are developed for ED on specific domains, like the biomedical domain (Pyysalo et al., 2007; Kim et al., 2008; Thompson et al., 2009; Buyko et al., 2010; Nédellec et al., 2013), literature (Sims et al., 2019), Twitter (Ritter et al., 2012; Guo et al., 2013) and breaking news (Pustejovsky et al., 2003). These datasets are also typically small-scale due to the inherent complexity of event annotation, but their different settings are complementary to our work.

7 Conclusion and Future work

In this paper, we present a massive general domain event detection dataset (MAVEN), which significantly alleviates the data scarcity and low coverage problems of existing datasets. We conduct a thorough evaluation of the state-of-the-art ED models on MAVEN. The results indicate that general domain ED is still challenging and MAVEN may facilitate further research. We also explore some promising directions with analytic experiments, including modeling multiple event correlations (Section 5.3), utilizing the hierarchical event schema to distinguish close types (Section 5.6), and improving other ED tasks with transfer learning (Section 5.5). In the future, we will extend MAVEN to more event-related tasks like event argument extraction, event sequencing, etc.

Acknowledgement

We thank the anonymous reviewers for their insightful comments and suggestions. This work is supported by NSFC Key Projects (U1736204, 61533018), grants from Institute for Guo Qiang, Tsinghua University (2019GQB0003) and Beijing Academy of Artificial Intelligence (BAAI2019ZD0502). This work is also supported by the Pattern Recognition Center, WeChat AI, Tencent Inc. Xiaozhi Wang is supported by Tsinghua University Initiative Scientific Research Program.

References

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53.

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

James Allan. 2012. *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media.

Jun Araki and Teruko Mitamura. 2015. Joint event trigger identification and event coreference resolution with structured perceptron. In *Proceedings of EMNLP*, pages 2074–2080.

Jun Araki and Teruko Mitamura. 2018. Open-domain event detection using distant supervision. In *Proceedings of COLING*, pages 878–891.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of ACL-COLING*, pages 86–90.

P Basile, A Caputo, G Semeraro, and L Siciliani. 2014. Extending an information retrieval system through time event extraction. In *Proceedings of DART*, pages 36–47.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2010. The GeneReg corpus for gene expression regulation events — an overview of the corpus and its in-domain and out-of-domain interoperability. In *Proceedings of LREC*.

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically labeled data generation for large scale event extraction. In *Proceedings of ACL*, pages 409–419.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of ACL-IJCNLP*, pages 167–176.

Yubo Chen, Hang Yang, Kang Liu, Jun Zhao, and Yantao Jia. 2018. Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms. In *Proceedings of EMNLP*, pages 1267–1276.

Pengxiang Cheng and Katrin Erk. 2018. Implicit argument prediction with event knowledge. In *Proceedings of ACL*, pages 831–840.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Linguistic Data Consortium. 2005. ACE (Automatic Content Extraction) English annotation guidelines for events. *Version*, 5(4).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Ning Ding, Ziran Li, Zhiyuan Liu, Haitao Zheng, and Zibo Lin. 2019. Event detection with trigger-aware lattice neural network. In *Proceedings of EMNLP-IJCNLP*, pages 347–356.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of LREC*.

Zhendong Dong and Qiang Dong. 2003. HowNet - a hybrid language and knowledge resource. In *Proceedings of NLP-KE*, pages 820–824.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M Strassel. 2015. Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results. In *TAC*.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M Strassel. 2016. Overview of linguistic resources for the TAC KBP 2016 evaluations: Methodologies and results. In *TAC*.

Joe Ellis, Jeremy Getman, and Stephanie M Strassel. 2014. Overview of linguistic resources for the TAC KBP 2014 evaluations: Planning, execution, and results. In *TAC*.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Bing Qin, Heng Ji, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceedings of ACL*, pages 66–71.

James Ferguson, Colin Lockard, Daniel Weld, and Hannaneh Hajishirzi. 2018. Semi-supervised event extraction with paraphrase clusters. In *Proceedings of NAACL*, pages 359–364.

Charles Fillmore. 2006. Frame semantics. *Cognitive linguistics: Basic readings*, 34:373–400.

Charles J Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280, pages 20–32.

Tao Ge, Lei Cui, Baobao Chang, Zhifang Sui, Furu Wei, and Ming Zhou. 2018. EventWiki: A knowledge base of major events. In *Proceedings of LREC*.

Jeremy Getman, Joe Ellis, Zhiyi Song, Jennifer Tracey, and Stephanie M Strassel. 2017. Overview of linguistic resources for the TAC KBP 2017 evaluations: Methodologies and results. In *TAC*.

Reza Ghaeini, Xiaoli Fern, Liang Huang, and Prasad Tadepalli. 2016. Event nugget detection with forward-backward recurrent neural networks. In *Proceedings of ACL*, pages 369–373.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference- 6: A brief history. In *Proceedings of COLING*.

Weiwei Guo, Hao Li, Heng Ji, and Mona Diab. 2013. Linking tweets to news: A framework to enrich short text data in social media. In *Proceedings of ACL*, pages 239–249.

Prashant Gupta and Heng Ji. 2009. Predicting unknown time arguments based on cross-event propagation. In *Proceedings of ACL-IJCNLP*, pages 369–372.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R. Voss, Jiawei Han, and Avirup Sil. 2016. Liberal event extraction and event schema induction. In *Proceedings of ACL*, pages 258–268.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-shot transfer learning for event extraction. In *Proceedings of ACL*, pages 2160–2170.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL*, pages 254–262.

Jin-Dong Kim, Tomoko Ohta, Kanae Oda, and Jun'ichi Tsujii. 2008. From Text to Pathway: Corpus annotation for knowledge acquisition from biomedical literature. *Series on Advances in Bioinformatics and Computational Biology*, 6:165–176.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, page 282289.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of ACL*, pages 73–82.

Wei Li, Dezhi Cheng, Lei He, Yuanzhuo Wang, and Xiaolong Jin. 2019. Joint event extraction based on hierarchical event schemas from framenet. *IEEE Access*, 7:25001–25015.

Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2019. Cost-sensitive regularization for label confusion-aware event detection. In *Proceedings of ACL*, pages 5278–5283.

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of ACL*, pages 1789–1798.

Xiao Liu, Heyan Huang, and Yue Zhang. 2019. Open domain event extraction using neural latent variable models. In *Proceedings of ACL*, pages 2860–2871.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of EMNLP*, pages 1247–1256.

Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. 2019. Distilling discrimination and generalization knowledge for event detection via delta-representation learning. In *Proceedings of ACL*, pages 4366–4376.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of LREC*, pages 4417–4422.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL*, pages 1003–1011.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of NAACL*, pages 300–309.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of ACL*, pages 365–371.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

James Pustejovsky, Patrick Hanks, Roser Saur, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The TimeBank corpus. *Proceedings of Corpus Linguistics*.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bjrne, Jorma Boberg, Jouni Jrvinen, and Tapio Salakoski. 2007. BioInfer: A corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8:50.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392.

Alan Ritter, Mausam Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. *Proceedings of the ACM SIGKDD*, 1104-1112.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of ACL*, pages 3623–3634.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *arXiv preprint arXiv:1706.09528*.

Paul Thompson, Syed Iqbal, John McNaught, and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC bioinformatics*, 10:349.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019a. Can You Tell Me How to Get Past Sesame Street? Sentence-Level Pretraining Beyond Language Modeling. In *Proceedings of ACL*, pages 4465–4476.

Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun, and Peng Li. 2019b. Adversarial training for weakly supervised event detection. In *Proceedings of NAACL*, pages 998–1008.

Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. 2019c. HMEAE: Hierarchical modular event argument extraction. In *Proceedings of EMNLP-IJCNLP*, pages 5777–5783.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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 EMNLP-IJCNLP*, pages 5766–5770.

Hui Yang, Tat-Seng Chua, Shuguang Wang, and Chun-Keat Koh. 2003. Structured use of external knowledge for event-based open domain question answering. In *Proceedings of SIGIR*, pages 33–40.

Yang Yang, Deyu Zhou, Yulan He, and Meng Zhang. 2019. Interpretable relevant emotion ranking with event-driven attention. In *Proceedings of EMNLP-IJCNLP*, pages 177–187.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of EMNLP*, pages 2369–2380.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of ACLs*, pages 764–777.

Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. *ArXiv*, abs/1212.5701.

Ying Zeng, Yansong Feng, Rong Ma, Zheng Wang, Rui Yan, Chongde Shi, and Dongyan Zhao. 2018. Scale up event extraction learning via automatic training data generation. In *Proceedings of AAAI*.

Yue Zhao, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng. 2018. Document embedding enhanced event detection with hierarchical and supervised attention. In *Proceedings of ACL*, pages 414–419.

A Hyperparameter Settings and Training Details

In this section, we introduce the hyperparameter settings and training details of various ED models that we implemented for experiments.

A.1 BERT-based Models

For both **DMBERT** and **BERT-CRF**, we use the $\text{BERT}_{\texttt{BASE}}$ model and the released pre-trained checkpoints[2], and implement them with HuggingFace's Transformers library (Wolf et al., 2019). The two models are both trained with the AdamW[3] optimizer and share most of the hyperparameters. Their hyperparameters are shown in Table 8.

For the **DMBERT** model, we insert special tokens (`[unused0]` and `[unused1]`) around the trigger candidates to indicate their positions and use a much larger batch size, hence the results are higher than the original implementation (Wang et al., 2019b).

For the **BERT+CRF** model, we use the widely-used "BIO" tagging schema, where "B-EventType", "I-EventType" and "O" stand for "Begin Event Type", "Inside Event Type" and "Others" respectively.

Learning Rate	5×10^{-5}
Adam ϵ	1×10^{-8}
Warmup Rate	0.0
DMBERT Batch Size	336
BERT-CRF Batch Size	256
DMBERT Validation Steps	500
BERT-CRF Validation Steps on MAVEN	100
BERT-CRF Validation Steps on ACE 2005	50

Table 8: Hyperparameter settings for the BERT-based models.

A.2 MOGANED Model

MOGANED model is implemented by ourselves since the official codes are not released. Compared with the original paper, our reproduction

[2] https://github.com/google-research/bert

[3] https://www.fast.ai/2018/07/02/adam-weight-decay/#adamw

uses Adam optimizer and does not use the L2 norm, while other model details are the same as Yan et al. (2019). We set most hyperparameters same as Yan et al. (2019) but the hyperparameter λ to be 1 rather than 5 since we find it can achieve better performances on both datasets. For the hyperparameters not mentioned in the original paper, we tune them manually. All hyperparameters are shown in Table 9.

K	3
λ	1
Batch Size	30
Leaky Alpha	0.2
Dropout Rate	0.3
Learning Rate	1×10^{-3}
Dimension of Pos-Tag Feature	50
Dimension of NER-Tag Feature	50
Dimension of Word Embedding	100
Dimension of Position Embedding	50
Dimension of Hidden Feature	100
Dimension of Graph Feature	150
Dimension of W_{att} Feature	100
Dimension of Aggregation Feature	100

Table 9: Hyperparameter settings for MOGANED.

A.3 DMCNN model

DMCNN model is implemented by ourselves since the official codes are not released. Compared with Chen et al. (2015), we use Adam optimizer instead of the ADADELTA (Zeiler, 2012) optimizer. We set all the hyperparameters the same as Chen et al. (2015) except the word embedding dimension and learning rate, which are not mentioned in the original paper. As the pre-trained word embeddings used in the original paper are not publicly released, we use the pre-trained word embeddings released by Chen et al. (2018) instead. The hyperparameters are shown in Table 10.

Batch Size	170
Dropout Rate	0.5
Learning Rate	1×10^{-3}
Adam ϵ	1×10^{-8}
Kernel Size	3
Dimension of PF	5
Number of Feature Map	200
Dimension of Word Embedding	100

Table 10: Hyperparameter settings for DMCNN.

A.4 BiLSTM-based Models

For both **BiLSTM** and **BiLSTM-CRF**, we use the pre-trained word embeddings released by Chen et al. (2018) and train them with the Adam (Kingma and Ba, 2014) optimizer. Similar with **BERT-CRF**, we use “BIO” tagging schema in **BiLSTM-CRF**. Their hyperparameters are shown in Table 11.

Batch Size	200
Dropout Rate	0.3
Learning Rate	1×10^{-3}
Adam ϵ	1×10^{-8}
Dimension of Hidden Layers	200
Dimension of Word Embedding	100

Table 11: Hyperparameter settings for the BiLSTM-based models.

A.5 Overall Training Details

For reproducibility, we report the training details of various models in this section. Table 12 shows the used computing infrastructures, the numbers of model parameters as well as the average running time of various models.

We mostly follow the original hyperparameter settings but also manually tune some hyperparameters. We select the models with the F-1 scores on the development sets of the both datasets. The validation performances of various models are shown in Table 13.

Method	**Computing Infrastructure**	**#para.**	**Runtime**	
			ACE 2005	**MAVEN**
DMCNN	1× RTX 2080 Ti	2M	3 min	5.5 min
BiLSTM	1× RTX 2080 Ti	2M	18 min	29 min
BiLSTM+CRF	1× RTX 2080 Ti	3M	21 min	67 min
MOGANED	1× RTX 2080 Ti	40M	55 min	90 min
DMBERT	8× RTX 2080 Ti	110M	110 min	201 min
BERT+CRF	1× RTX 2080 Ti	110M	32 min	97 min

Table 12: Training details of various models, including the computing infrastructures, the numbers of parameters, and the average runtimes.

Method	**ACE 2005**			**MAVEN**		
	P	**R**	**F-1**	**P**	**R**	**F-1**
DMCNN	73.3	53.5	61.8	66.5	55.5	60.5
BiLSTM	72.3	67.6	69.8	60.3	66.9	63.4
BiLSTM+CRF	75.9	60.8	67.5	64.1	64.6	64.3
MOGANED	72.4	66.2	69.1	63.7	63.7	63.7
DMBERT	71.4	72.4	71.9	64.6	70.1	67.2
BERT+CRF	75.4	76.8	76.1	65.7	68.8	67.2

Table 13: Validation performance of various models.

B Event Type Mapping for ACE and MAVEN

In Table 14, we present the event type mapping between parts of ACE 2005 and MAVEN event types, which is used in the data augmentation experiments in Section 5.5.

C Hierarchical Event Type Schema

We present the tree-structure hierarchical event type schema used by MAVEN in Figure 6. The eight red types are virtual types without annotated

ACE Types	MAVEN Types
Injure	Bodily_harm
Die	Death
Transport	Traveling
Transfer-Ownership	Getting,Receving, Commerce_buy, Giving, Submitting_documents, Supply, Commerce_sell, Renting, Exchange
Transfer-Money	Commerce_pay, Expensiveness, Earnings_and_losses
Attack	Attack
Demonstrate	Protest
Meet	Come_together, Social_event
Phone-Write	Communication, Telling
Arrest-Jail	Arrest, Prison
Extradite	Extradition
Trial-Hearing	Justifying

Table 14: Mapping between parts of ACE 2005 event types and MAVEN event types.

instances, which are only used for organizing similar event types together. The virtual types do not participate in classification for all the models and when we say we have 168 event types we do not take them into account.

D Event Types and their corresponding frames

As stated in Section 3.1, we manually induce 168 event types from the 598 FrameNet event-related frames. We present the mapping between the event types and frames in Table 15 to help understand our event schema construction process. Note that the shown mapping is not a strict mapping, i.e., the semantic coverage of a MAVEN event type may be larger than the union of its corresponding frames.

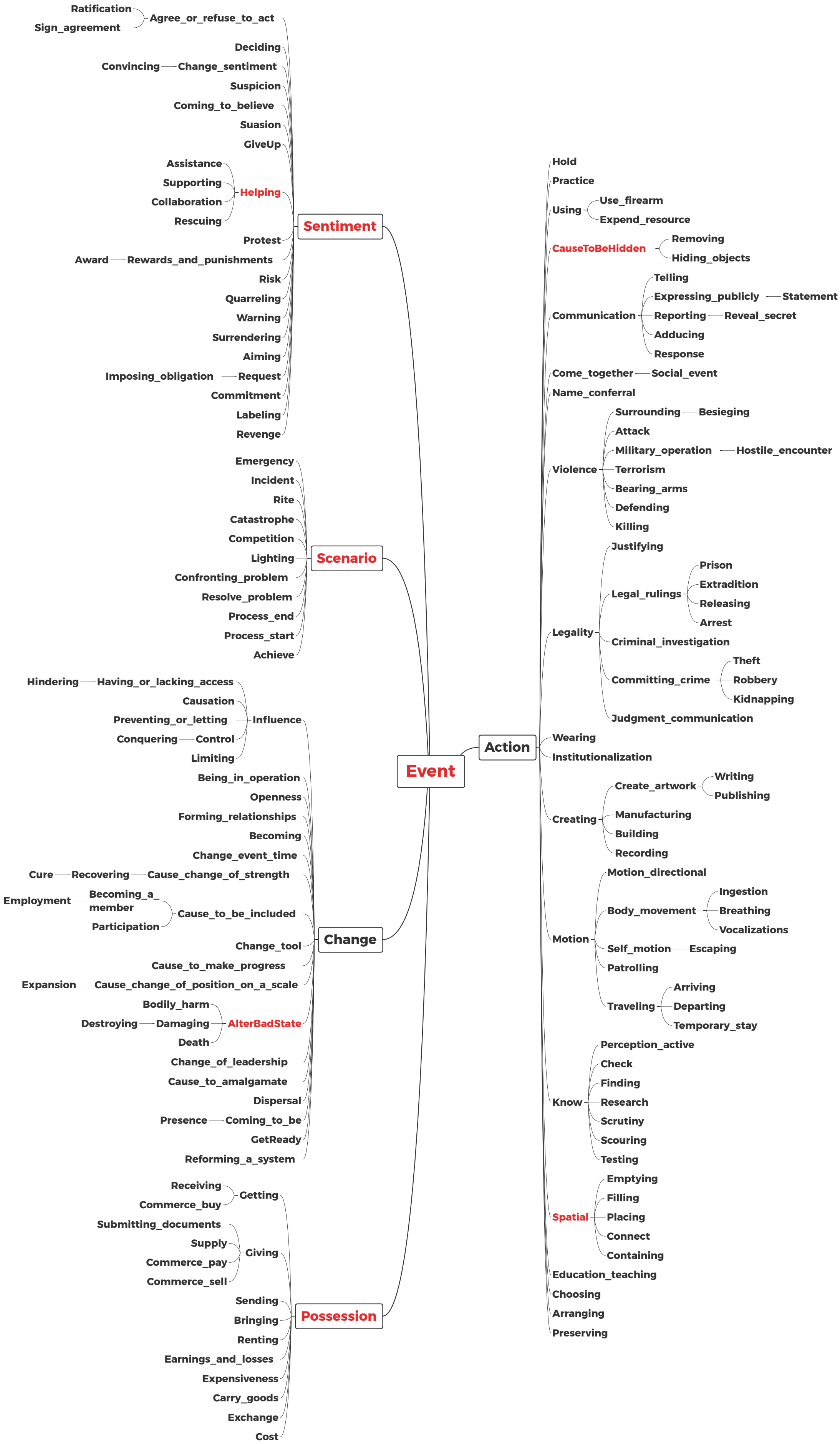

Figure 6: The hierarchical event type schema used in MAVEN. The red labels are virtual event types without annotated instances.

Event Type	Corresponding Frame(s)
Know	Becoming_aware
Warning	Warning
Catastrophe	Catastrophe
Placing	Placing_scenario, Placing, Being_located
Causation	Cause_to_start, Causation
Arriving	Drop_in_on, Visitor_arrival, Access_scenario, Visit_host_arrival, Arriving, Visiting_scenario_arrival
Sending	Commerce_money-transfer, Sending, Delivery, Product_delivery, Commerce_goods-transfer,Transfer, Post_transfer
Protest	Reasoning
Preventing_or_letting	Avoiding, Preventing, Prevent_from_having, Preventing_or_letting
Motion	Motion_scenario, Temporary_leave, Cause_motion, Cause_to_move_in_place, Motion, Cause_fluidic_motion, Fluidic_motion, Mass_motion
Damaging	Damaging
Destroying	Destroying, Cause_to_fragment, Render_nonfunctional
Death	Death, Losing_someone
Perception_active	Perception_active
Presence	Circumscribed_existence, Presence, Existence
Influence	Subjective_influence, Eventive_cognizer_affecting
Receiving	Post_receiving, Receiving
Check	Verification
Hostile_encounter	Hostile_encounter
Killing	Killing
Conquering	Conquering
Releasing	Releasing_from_custody, Bail_decision, Releasing, Freeing_from_confinement, Breaking_out_captive
Attack	Counterattack, Attack, Invading, Suicide_attack
Earnings_and_losses	Earnings_and_losses
Choosing	Adopt_selection, Choosing
Traveling	Visiting, Touring, Travel
Recovering	Rejuvenation
Using	Using
Coming_to_be	Coming_to_be
Cause_to_be_included	Cause_to_be_included
Process_start	Process_start, Activity_start
Change_event_time	Holding_off_on, Change_event_time, Change_event_duration
Reporting	Reporting

Bodily_harm	Cause_harm, Experience_bodily_harm
Suspicion	Suspicion
Statement	Statement, Claim_ownership
Cause_change_of_position_on_a_scale	Cause_change_of_position_on_a_scale
Coming_to_believe	Coming_to_believe
Expressing_publicly	Speak_on_topic, Expressing_publicly
Request	Request
Control	Being_in_control, Domination, Control, Self_control
Supporting	Supporting
Defending	Repel, Defending
Building	Building
Military_operation	Military_operation
Self_motion	Self_motion
GetReady	Activity_ready_state
Forming_relationships	Forming_relationships
Becoming_a_member	Becoming_a_member
Action	Enforcing, Execute_plan, Conduct, Intentionally_act
Removing	Removing, Removing_scenario
Surrendering	Surrendering, Surrendering_possession
Agree_or_refuse_to_act	Agree_or_refuse_to_act
Participation	Participation
Deciding	Deciding, Waver_between_options
Education_teaching	Education_teaching
Emptying	Emptying, Container_focused_removing
Getting	Getting, Post_getting
Besieging	Besieging
Creating	Intentionally_create, Creating, Coming_up_with
Process_end	Process_completed_state, Process_end, Activity_done_state, Cause_to_end, Activity_stop
Body_movement	Gesture, Body_movement
Expansion	Cause_expansion
Telling	Telling
Change	Cause_change, Cause_change_of_phase
Legal_rulings	Legal_rulings
Bearing_arms	Bearing_arms
Giving	Conferring_benefit, Offering, Giving, Post_giving
Name_conferral	Name_conferral
Arranging	Arranging, Making_arrangements
Use_firearm	Use_firearm
Committing_crime	Committing_crime, Misdeed, Offenses
Assistance	Assistance

Surrounding	Surrounding
Quarreling	Quarreling
Expend_resource	Expend_resource
Motion_directional	Motion_directional, Intentional_traversing, Traversing
Bringing	Bringing
Communication	Chatting, Talking_into, Communication_response, Encoding, Contacting, Discussion, Successfully_communicate_message, Communication
Containing	Containing, Containment
Manufacturing	Manufacturing
Social_event	Social_event_individuals, Social_event_collective, Social_event
Robbery	Robbery
Competition	Competition
Writing	Text_creation
Rescuing	Rescuing
Judgment_communication	Judgment_communication, Judgment_direct_address
Change_tool	Change_tool
Hold	Manipulation, Manipulate_into_doing
Being_in_operation	Being_in_operation, Being_operational
Recording	Recording
Carry_goods	Carry_goods
Cost	Expensiveness
Departing	Visitor_departure, Setting_out, Disembarking, Visit_host_departure, Visiting_scenario_departing, Departing
GiveUp	Abandonment
Change_of_leadership	Change_of_leadership
Escaping	Dodging, Fleeing, Escaping, Evading, Quitting_a_place
Aiming	Aiming
Hindering	Hindering
Preserving	Preserving
Create_artwork	Create_physical_artwork, Craft
Openness	Openness
Connect	Spatial_contact, Attaching
Reveal_secret	Reveal_secret
Response	Response, Respond_to_proposal, Response_scenario
Scrutiny	Court_examination, Scrutiny, Inspecting, Scrutinizing_for
Lighting	Light_movement
Criminal_investigation	Criminal_investigation
Hiding_objects	Hiding_objects
Confronting_problem	Confronting_problem, Difficulty
Renting	Renting

Breathing	Breathing
Patrolling	Patrolling
Arrest	Arrest, Detaining, Imprisonment, Being_incarcerated, Being_in_captivity
Convincing	Suasion, Attempt_suasion
Commerce_sell	Commerce_sell
Cure	Cure
Temporary_stay	Temporary_stay
Dispersal	Dispersal
Collaboration	Collaboration
Extradition	Extradition
Change_sentiment	Cause_to_experience
Commitment	Commitment
Commerce_pay	Commerce_pay
Filling	Filling, Container_focused_placing
Becoming	Becoming
Achieve	Accomplishment
Practice	Practice
Cause_change_of_strength	Cause_change_of_strength
Supply	Supply
Cause_to_amalgamate	Cause_to_amalgamate
Scouring	Scouring
Violence	Violence
Reforming_a_system	Reforming_a_system
Come_together	Gathering_up, Come_together
Wearing	Dressing, Clothing, Wearing
Cause_to_make_progress	Cause_to_make_progress
Legality	Legality
Employment	Being_employed
Rite	Rite
Publishing	Publishing
Adducing	Adducing
Exchange	Exchange, Exchange_currency
Ratification	Ratification
Sign_agreement	Sign_agreement
Commerce_buy	Shopping, Commerce_buy
Imposing_obligation	Imposing_obligation
Rewards_and_punishments	Fining, Execution, Rewards_and_punishments, Corporal_punishment
Institutionalization	Institutionalization
Testing	Operational_testing, Examination
Ingestion	Ingestion, Ingest_substance
Labeling	Labeling
Kidnapping	Kidnapping
Submitting_documents	Submitting_documents

Prison	Prison
Justifying	Justifying
Emergency	Emergency, Emergency_fire
Terrorism	Terrorism
Vocalizations	Vocalizations
Risk	Daring
Resolve_problem	Resolve_problem
Revenge	Revenge
Limiting	Limiting, Limitation
Research	Experimentation, Research
Having_or_lacking_access	Having_or_lacking_access
Theft	Theft
Incident	Coincidence
Award	Deserving

Table 15: The 168 event types in MAVEN and their corresponding frames in FrameNet.