DocEE: A Large-Scale and Fine-grained Benchmark for Document-level Event Extraction

Meihan Tong[1], **Bin Xu**[1]*, **Shuai Wang**[2], **Meihuan Han**[1], **Yixin Cao**[3]
Jiangqi Zhu[1], **Siyu Chen**[1], **Lei Hou**[1], **Juanzi Li**[1]
[1]Knowledge Engineering Laboratory, Tsinghua University, Beijing, China
[2]SLP Group, AI Technology Department, JOYY Inc, China
[3]Singapore Management University, Singapore
tongmeihan@gmail.com, xubin@tsinghua.edu.cn

Abstract

Event extraction aims to identify an event and then extract the arguments participating in the event. Despite the great success in sentence-level event extraction, events are more naturally presented in the form of documents, with event arguments scattered in multiple sentences. However, a major barrier to promote document-level event extraction has been the lack of large-scale and practical training and evaluation datasets. In this paper, we present DocEE, a new document-level event extraction dataset including 27,000+ events, 180,000+ arguments. We highlight three features: large-scale manual annotations, fine-grained argument types and application-oriented settings. Experiments show that there is still a big gap between state-of-the-art models and human beings (41% Vs 85% in F1 score), indicating that DocEE is an open issue. DocEE is now available at https://github.com/tongmeihan1995/DocEE.git.

1 Introduction

Event Extraction (EE) aims to detect events from text, including event classification and event argument extraction. EE is one of the fundamental tasks in text mining (Feldman and Sanger, 2006) and has many applications. For instance, it can monitor political or military crises to generate real-time notifications and alerts (Dragos, 2013), and dig the links and connections (e.g., Who Met Whom and When) between dignitaries for portrait analysis (Zhan et al., 2020).

Most existing datasets (e.g., ACE2005 [1] and KBP2017 [2]) focus on sentence-level event extraction, while events are usually described at the document level, and event arguments are typically scattered across different sentences (Hamborg et al., 2019). Figure 1 shows an *Air Crash* event. To extract argument *Date*, we need to read sentence [1], while to extract argument *Cause of the Accident*, we need to integrate information in sentences [6] and [7]. Clearly, this requires reasoning over multiple sentences and modeling long-distance dependency, intuitively beyond the reach of sentence-level EE. Therefore, it is necessary to move EE forward from sentence-level to document-level.

Only a few datasets are curated for document-level EE. MUC-4(Grishman and Sundheim, 1996) provides 1,700 news articles annotated with 4 event types and 5 argument types. The 5 arguments are shared among different event types without further refinement. WikiEvents(Li et al., 2021) consists of only 246 documents with very few (22% of total) cross-sentences argument annotations. RAMS(Ebner et al., 2020) limits the scope of the arguments in a 5-sentence window around its event trigger, which is not in line with the actual application, and the number of the argument types in RAMS is only 65, which is quite limited. Doc2EDAG, TDJEE and GIT (Zheng et al., 2019; Wang et al., 2021; Xu et al., 2021) contain only 5 event types and 35 argument types in financial domain. In summary, existing datasets for document-level EE fail in the following aspects: small scale of data, limited coverage of domain and insufficient refinement of argument types. Therefore, it is urgent to develop a manually labeled, large-scale dataset to accelerate the research in document-level EE.

In the paper, we present DocEE, a large-scale human-annotated document-level EE dataset. Figure 1 illustrates an example of DocEE. DocEE focus on the extraction of the main event, that is *one-event-per-document*. We regard news headlines as the main event trigger and focus on main event arguments extraction throughout the article. We highlight the following three contributions of DocEE to this field: 1) Large-scale Manual Annotations. Do-

*Corresponding author.

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

[2] https://tac.nist.gov/2017/KBP/

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3970 - 3982
July 10-15, 2022 ©2022 Association for Computational Linguistics

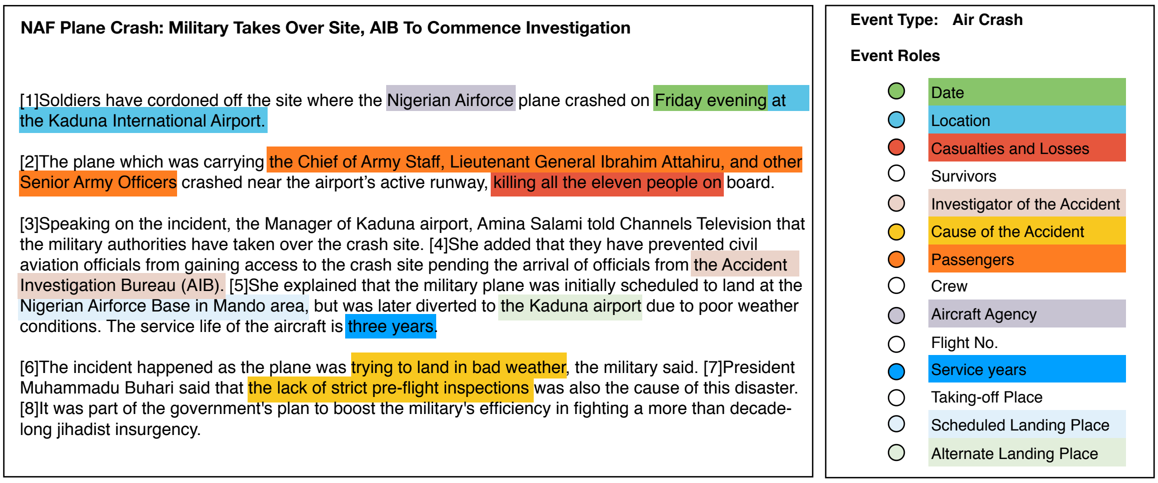

Figure 1: An example from DocEE. Each document in DocEE is annotated with event type and involved event arguments. In the example, the document mainly describes a *Air Crash* event which contains the following arguments: *Data*,*Location*, *Causality and Losses* and etc. We use different colors to distinguish event arguments.

cEE contains 27,485 document-level events with 180,528 arguments, far exceeding the scale of existing document-level EE datasets. The large-scale annotations of DocEE can provide sufficient training and testing data, to fairly evaluate EE models. 2) Fine-grained argument types. DocEE has a total of 356 argument types, which is much more than the number of argument types in existing dataset (5 in MUC-5 and 65 in RAMS). Besides the general arguments, such as time and location, we design more personalized event arguments for each event type, such as *Water Level* for *Flood* event and *Magnitude* for *Earthquake* event. These fine-grained roles can bring more detailed semantics and pose a higher challenge to the semantic disambiguation ability of existing models. 3) Application-oriented settings. In the actual application, event extraction often face the problems of how to quickly adapt from the rich-resource domains to new domains. Therefore, we have added a cross-domain setting to better test the transfer capability of the EE models. In addition, DocEE removes the limitation that the arguments range should be within a certain window in RAMS, to better cope with realistic scenarios where the length of the article will be particularly long, and the argument of the event may appear in any corner of the article. With more scattered event arguments (see Table 1), DocEE poses a higher challenge to the long text processing capability of existing models.

To assess the challenges of DocEE, we implement 9 recent state-of-the-art EE models on DocEE along with human evaluation. Experiments demonstrate the high-quality of DocEE and show that even the performance of SOTA model is far lower than human performance, showing that the faintness of existing technology in processing document-level EE.

2 Related Datasets

Sentence-level Event Extraction Dataset Automatic Content Extraction (ACE2005)[1] consists of 599 documents with 8 event types and 33 subtypes. Text Analysis Conference (TAC-KBP)[2] also releases three benchmarks: TAC-KBP 2015/2016/2017, with 9/8/8 event types and 38/18/18 event subtypes. RED[3] annotates events from 95 English newswires. Chinese Emergency Corpus (CEC) focuses on Chinese breaking news, with a total of 332 articles in 5 categories. MAVEN (Wang et al., 2020) and LSEE (Chen et al., 2017) only annotate event triggers, with 168/21 types of trigger instances in 11,832/72,611 sentences. Based on them, various superior models have been proposed to improve the sentence-level EE and have achieved great success (Orr et al., 2018; Nguyen and Grishman, 2018; Tong et al., 2020).

Document-level Event Extraction Dataset Most of the existing document-level event datasets only focus on event classification, but lack event argument labelings, such as 20news [4] and THUC-News [5]. There are a few datasets annotated with cross-sentences event arguments. MUC-4 (Nguyen

[3] https://catalog.ldc.upenn.edu/LDC2016T23
[4] https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups
[5] http://thuctc.thunlp.org

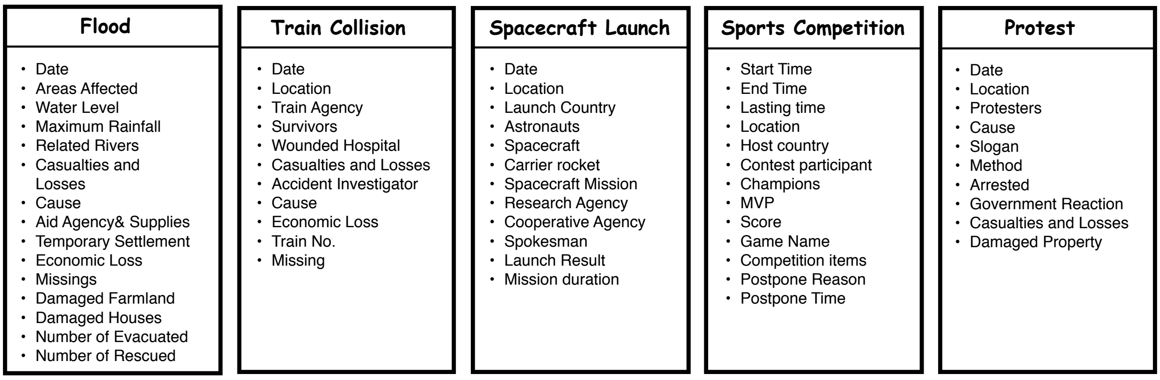

Figure 2: Five examples of event schema in DocEE.

et al., 2016) only contains 4 event types and 5 argument types, and the 4 event types are close to each other and limited to the terrorist attack topic[6]. WikiEvents (Li et al., 2021) and RAMS(Ebner et al., 2020) consist of 246/9,124 documents with only 59/65 argument types, and most of the arguments in the two datasets are shared among different event types without further refinement. Doc2EDAG, TDJEE and GIT (Zheng et al., 2019; Wang et al., 2021; Xu et al., 2021) only define 5 event types and 35 argument types in financial domain. Cancer Genetics, EPM, GENIA2011, GENIA2013, Pathway Curation and MLEE (Pyysalo et al., 2013; Ohta et al., 2011; Kim et al., 2011, 2013; Ohta et al., 2013; Van Landeghem et al., 2013) are limited to the biological domain. In summary, these datasets are either limited to specific domains, or have very limited data scale, or have not carefully refined event argument schema.

Open-domain Event Extraction Dataset To collect EE dataset in open domain, one way is to leverage semi-structured resources (Wikipedia) or existing knowledge bases (Freebase). The representative works are EventKG (Gottschalk and Demidova, 2018), Event Wiki (Ge et al., 2018) and Historical Wiki (Hienert and Luciano, 2012). The other way is to exploit open IE tools, such as dependency parsing, to extract events from unstructured text. The representative works are Event Logic Graph (Ding et al., 2019) and Giveme5W1H (Hamborg et al., 2019). The advantage of the open-domain EE dataset is its large scale, but the disadvantage is that it lacks manual review, and thus the quality cannot be guaranteed.

[6] https://www-nlpir.nist.gov/related_projects/muc/muc_data/muc_data_index.html

3 Constructing DocEE

Our main goal is to collect a large-scale dataset to promote the development of event extraction from sentence-level to document-level. In the following sections, we will first introduce how to construct the event schema, and then how to collect candidate data and how to label them through crowdsourcing.

3.1 Event Schema Construction

News is the first-hand source of hot events, so we focus on extracting events from news. Previous event schemas, such as FrameNet (Baker, 2014) and HowNet (Dong and Dong, 2003), pay more attention to trivial actions such as *eating* and *sleeping*, and thus is not suitable for document-level news event extraction.

To construct event schema, we gain insight from journalism. Journalism typically divides events into hard news and soft news (Reinemann et al., 2012; Tuchman, 1973). Hard news is a social emergency that must be reported immediately, such as earthquakes, road accidents and armed conflicts. Soft news refers to interesting incidents related to human life, such as celebrity deeds, sports events and other entertainment-centric reports. Based on the hard/soft news theory and the category framework in (Lehman-Wilzig and Seletzky, 2010), we define a total of 59 event types, with 31 hard news event types and 28 soft news event types. Detailed information is shown in Appendix Table 1. Our schema covers influential events of human concern, such as earthquakes, floods and diplomatic summits, which cannot be extracted at the sentence level and require multiple sentences to describe.

To construct argument schema, we leverage infobox in Wikipedia. As shown in Figure 3(a), the Wikipedia page describes an event, and the keys in the infobox, such as *Date* and *Total fatalities*, can be regarded as the prototype arguments of the event. Based on this observation, we manually collect 20 wiki pages for each event type, and use their shared keys in infobox as our basic set of argument types. After that, we further expand the basic set. Specifically, for event type e, we first collect 20 news stories from New York Times, and then invited 5 students (native English-speaking, major in journalism) to summarize the key facts the public would like to learn from the news of e. For instance, in *Flood* event news, *Water Level* is a key fact, because it is an important factual basis for flood cause analysis and disaster relief decision-making, and can arouse widespread concern. Finally, by merging the key facts of the 5 students, we complete the argument types expansion. To ensure the quality, we further invite the above 5 students to make a trial labeling on the collected news, and filter argument types that appear less frequently in the article.

In total, we define 356 event arguments types for 59 event types. On average, there are 6.0 event arguments per class. Figure 2 illiterates some examples of event arguments types we defined. The complete schema and corresponding examples can be found *Event Schema.md* in the supplementary materials.

3.2 Candidate Data Collection

In this section, we introduce how to collect candidate document-level events. We choose wiki as our annotation source. Wiki contains two kinds of events: historical events and timeline events (Hienert and Luciano, 2012). Historical events refer to the events that have their own wiki page, such as *1922 Picardie mid-air collision*. Timeline events refer to the news events organized in chronological order, such as *A heatwave strikes India and South Asia* in wiki page *Portal:Current_events/June_2010*.[7] Figure 3 shows examples of two events. We adopt both kinds of events as our candidate data, because only using historical events will lead to uneven data distribution under our event schema, and timeline events can be a good supplement.

For a historical event, we adopt its Wikipedia article as the document of the event arguments to be annotated. For a timeline event, we use the URL to download the original news article as the document of the event arguments to be annotated. Because 22% of the timeline events do not have URLs (Wikipedia editors do not provide the URL when editing the entry), so we use Scale SERP[8] to find news articles and manually confirm their authenticity. For historical event, we adopt *templates+event type* as the query key to retrieve candidate events. The templates includes *"List of"+event type*, *event type+"in"+year*, *"Category:"+event type+"in"+country*, etc. More templates show in Appendix Table 7. For timeline event, we choose events between 1980 and 2021 as candidates, because there are very few events before 1980.

In order to balance the length of the article, we filtered out articles less than 5 sentences, and also truncated articles that were too long (more than 50 sentences). Finally, we select 44,000 candidate events from Wikipedia.

3.3 Crowdsourced Labeling

Given the candidate events and the predefined event schema, we now introduce how to annotate them through crowdsourcing. To ensure the quality of annotations, all annotators are either native English speakers or English-major students with TOEFL higher than 100 or IELTS higher than 7.5. The crowdsourced labeling process consists of two stages.

3.3.1 Stage 1: Event Classification

At this stage, annotators are required to classify candidate events into predefined event types. Following (Hamborg et al., 2018; Hsi, 2018), we focus on main event classification, so Stage 1 is a single-label classification task. Specifically, the main event refers to the event reflected in the title and mainly described in the article. Formally, given the candidate event $e =< t, a >$, where t represents the title and a represents the article, Stage 1 aims to obtain label y for each e, where y belongs to the 59 event types defined in subsection 3.1.

In total, we invite about 60 annotators to participate in Stage 1 annotation. The online annotation page is displayed in Appendix Figure 5. We first manually label 100 articles as standard answers to *pre-test* annotators, and weed out annotators with an accuracy rate of less than 70%, which left us 48 valid annotators. Then, we ask two independent annotators to annotate each candidate event. If the results of the two annotators are inconsistent

[7] en.wikipedia.org/wiki/Portal:Current_events/June_2010

[8] https://app.scaleserp.com/playground

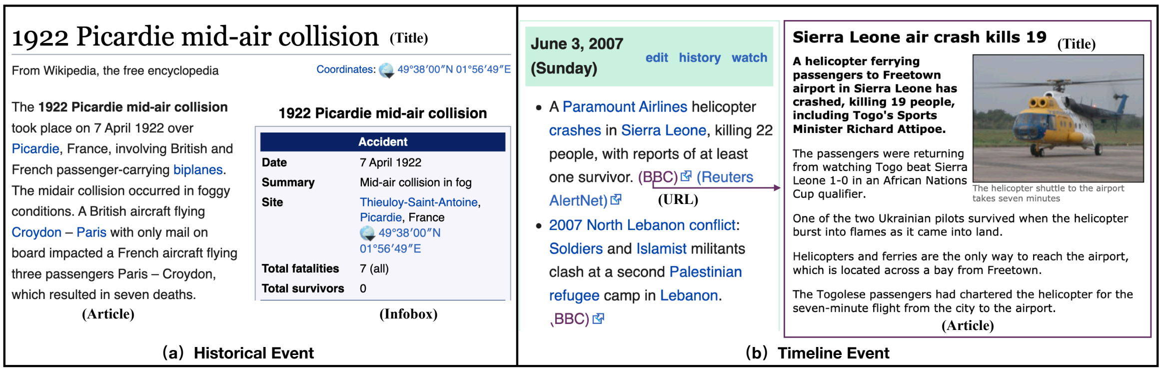

Figure 3: Two sources of candidate events in DocEE. The left is a historical event, which has its own wiki page, and the right are two timeline events arranged in a wiki page by time unit. Each timeline event consists of a brief description and a URL pointing to original news.

(32.8% in this case), a third annotator will be the final judge. Due to the variety of event types in reality, a candidate event may not belong to any predefined class. We classify such event into the other class, which accounts for 23.6% of the total data.

3.3.2 Stage 2: Event argument Extraction

At this stage, annotators are required to extract event arguments from the whole article. Formally, given the candidate event $e =< t, a >$, its event type y and the predefined argument types R of y, Stage 2 aims to find all the arguments from the article a.

Due to the heavy workload in Stage 2, we invite more than 90 annotators. An example of the online annotation page is shown in Appendix Figure 6. We use a *preliminary annotation - multiple rounds inspection* method for labeling. In the preliminary annotation step, each article will be labeled by an annotator. We distribute no more than two event types to each annotator in this step to make the annotators more focused. Then, in the step of multiple rounds inspection, we first select high-precision annotators based on inter-annotator agreement to form a reviewer team (44.4% of the total), and then each article will go through three rounds of error correction by three independent annotators in the reviewer team. After each round, we will feed back annotation issues to the reviewers so that they can correct them in the next round of annotation. The accuracy rate has steadily increased from 56.24%, 76.83% to 85.96% after each round, which shows the effectiveness of our labeling method. We take the third round results as the final annotations.

We clarify some annotation details here. We do not include articles, prepositions in our annotations. For instance, we select "damaged car" among "damaged car", "damaged car belonging to the victim" and "the damaged car". For event arguments with multiple mentions in the document, for example, *Cause of the Accident* in Figure 1 that has two mentions, we will label all mentions to ensure the completeness of the extraction. For repeated mentions that refer to the same entity, we only label once.

3.3.3 Annotation Quality & Remuneration

Following (Artstein and Poesio, 2008; McHugh, 2012), we use Cohen's kappa coefficient to measure the Inter-Annotator Agreement(IAA). The IAA scores are 94% and 81% for State 1 Event Classification and Stage 2 Event Argument Extraction respectively, which are relatively high. The annotators spend an average of 0.5 minutes labeling a piece of data in Stage 1, so we pay them 0.1$ for each piece of data. It takes about 5 minutes to label a piece of data in Stage 2 , so we pay 0.8$ for each piece of data.

4 Data Analysis of DocEE

In the section, we analyze various aspects of DocEE to provide a deep understanding of the dataset and the task of document-level event extraction.

Overall Statistic In total, DocEE labels 27,485 valid document-level events and 180,528 event arguments. Each article is annotated with 6.6 event arguments on average. Event *Famous Person - Divorce* has the highest average number of event arguments per article (18.1), while event *Regime Change* has the lowest average number of event arguments per article (3.8). We compare DocEE to various representative event extraction datasets in Table 1, including sentence-level EE datasets

Datasets	#isDocEvent	#EvTyp.	#ArgTyp.	#Doc.	#Tok.	#Sent.	#ArgInst.	#ArgScat.
ACE2005	✗	33	35	599	290k	15,789	9,590	1
KBP2016	✗	18	20	169	94k	5,295	7,919	1
KBP2017	✗	18	20	167	86k	4,839	10,929	1
MUC-4	✓	4	5	1,700	495k	21,928	2,641	4.0
WikiEvents	✓	50	59	246	190k	8,544	5,536	2.2
RAMS	✓	139	65	9,124	957k	34,536	21,237	4.8
DocEE(ours)	✓	59	356	27,485	16,268k	749,568	180,528	10.2

Table 1: Statistics of EE datasets (isDocEvent: whether the event in the corpus at the document-level, EvTyp.: event type, ArgTyp.: event argument type, Doc.: document, Sent.: sentence, ArgInst.: event arguments, ArgScat.: the number of sentences in which event arguments of the same event are scattered)

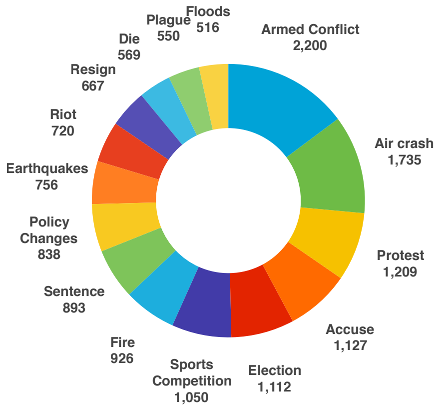

Figure 4: Top 15 event types in DocEE.

ACE2005, KBP and document-level EE dataset MUC-4, Wikievents, RAMS. We find that DocEE is much larger than existing datasets in many aspects, including the documents number and argument instances number. Compared to MUC-4, DocEE has far more event arguments (180,528 compared to 2,641). The reason is that among the 1,700 documents in MUC-4, 47.4% of articles are not labeled with any event argument, while DocEE guarantees that each article contains at least three event argument labels in crowdsourcing process, which greatly solves the problem of data scarcity of the event arguments in document-level EE.

Event Type Statistic Figure 4 shows the distribution of the top 15 frequent event types that have the most number of instances in DocEE. DocEE covers a variety of event types, including Armed Conflict (8.0%), Air crash (6.3%), Protest (4.4%), CommitCrime - Accuse (4.1%), Election (4.0%), Sports Competition (3.8%), Fire (3.4%), etc. Our annotated data follows a long-tailed distribution, which is due to the uneven distribution of class in the real data. According to statistics, there are 30.5% of classes with more than 500 instances and 83.1% of classes with more than 200 instances. More detailed information is shown in Appendix Table 6.

Event Arguments Statistic We randomly sample 1000 articles from DocEE for manual analysis, which contains a total of 4962 event arguments instances. We first classify event arguments based on their mention numbers. As shown in Table 2, 78.6% of the event arguments have a unique mention, and 21.4% of the event arguments have multiple mentions, which poses a great challenge to the model's recall capability. Then, we classify event arguments based on their mention lengths. 60.8% of the event arguments have no more than 3 words, and most of them are named entities such as person and location. While 30.8% event arguments have less than 10 words and 8.4% event arguments are answered by more than 10 words, such event arguments mainly include *Cause of the Accident*, *Investigation Results* , etc.

5 Experiments on DocEE

In this section, we show the challenges of DocEE by conducting comprehensive experiments on various SOTA models. We first introduce two benchmark settings, and then we conduct experiments on both event classification task and event argument extraction task. Finally, we discuss possible future research directions for document-level event extraction.

Benchmark Settings We design two benchmark settings for evaluation: normal setting and cross-domain setting. In the normal setting, we hope the training set and test set to be identically distributed. Specifically, for each event type, we randomly select 80% of the data as the training set, 10% of the data as the validation set, and the remaining 10% of the data as the test set.

Table 2: Answer types of event arguments in DocEE.

Answer Types	%	Examples
Single Answer	78.6	A masked man in a black hoodie showed a **gun** and was handed money before running east on Warren Street, according to the initial report. Argument Type: Bank Robbery Argument: Weapon Used
Multiple Answers	21.4	At around 6:20 a.m. **a lorry**, driven by David Fairclough of Wednesfield, rammed into the rear of **a tanker**, which then struck **a car** in front and exploded. The ensuing pile-up involved **160 vehicles** on a 400-yard (370 m) stretch of the motorway. Argument Type: Road Crash Argument: Number of Vehicles involved in the Crash

In order to be application-oriented, we design cross-domain setting to test the transfer capability of the SOTA models. We choose the event type under the subject of natural disasters as the target domain, including Floods, Droughts, Earthquakes, Insect Disaster, Famine, Tsunamis, Mudslides, Hurricanes, Fire and Volcano Eruption, and adopt the remaining 49 event types as source domains. The division reduces the overlap of argument types between the source domain and the target domain. In this setting, the models will first be pre-trained on the source domain, and then conduct 5-shot fine-tuning on the target domain. The detailed data split for each setting is shown in Table 3.

Method	Normal			Cross-Domain		
	Train	Dev	Test	Train	Dev	Test
#EvTyp.	59	59	59	59	10	10
#Doc.	22k	2.7k	2.7k	23.7k	1.6k	2.0k
#ArgInst.	141k	19k	19k	156k	11k	13k

Table 3: Statistics for two benchmark settings (Sec.5): normal and cross-domain.

Hyperparameters We use base version of pre-trained model for all the transformer-based methods, and set the learning rate to 2e-5. The batch size is 128 and the maximum document length is 512. All baselines are implemented by HuggingFace[9] with default parameters and all models can be fit into eight V100 GPUs with 16G memory. The training procedure lasts for about a few hours. For all the experiments, we report the average result of five runs as the final result. In human evaluation, we randomly select 1,000 document-level events and invite three students to label them. The final result is the average of their labeling accuracy.

5.1 Event Classification

Baselines We adopt a CNN-based method and various pre-trained transformer-based methods as our baselines, including: 1) **TextCNN** (Kim, 2014) uses different sizes CNN kernels to extract key information in text for classification. 2) **BERT** (Devlin et al., 2018) exploits unsupervised objective functions, masking language model (MLM) and next sentence prediction for pre-training. 3) **ALBERT** (Lan et al., 2020) proposes a self-supervised loss to improve inter-sentence coherence in BERT. 4) **DistilBert** (Sanh et al., 2019) combines language modeling, knowledge distillation and cosine-distance losses to improve BERT. 5) **RoBERTa** (Liu et al., 2019) is built on BERT and trains with much larger mini-batches and learning rates. Following (Kowsari et al., 2019), we use Precision(P), Recall(R) and Macro-F1 score as the evaluation metrics.

Method	Normal			Cross-Domain		
	P	R	F	P	R	F
TextCNN	78.6	75.4	76.2	8.5	2.1	2.6
BERT	89.3	89.8	89.5	72.6	78.4	75.4
ALBERT	88.8	88.9	88.1	25.5	21.3	23.2
DistilBert	89.6	90.7	90.1	79.8	79.6	79.7
RoBERTa	90.0	91.1	90.5	86.0	86.7	86.3
Human	98.4	97.7	98.0	-	-	-

Table 4: Overall Performance on Event Classification.

Overall Performance Table 4 shows the experimental results under the normal and cross-domain settings, from which we have the following observations: 1) Compared with TextCNN, transformer based models (BERT, ALBERT, DistillBert, RoBERTa) perform better, which are pre-trained on a large-scale unsupervised corpus and have more background semantic knowledge to rely on. 2) Humans have achieved high scores on DocEE, verifying the high quality of our annotated data sets. 3) There is still a gap between the performance of the current SOTA models and human beings, which indicates that more technological advances are needed in future work. Humans can connect and merge key information to form a knowledge

[9] https://huggingface.co/models

Methods	Normal Setting						Cross-domain Setting					
	EM			HM			EM			HM		
	P	R	F	P	R	F	P	R	F	P	R	F
BERT_Seq(sent)	23.8	28.2	25.8	38.1	45.1	41.3	7.0	7.6	7.2	18.4	20.0	19.1
BERT_Seq(chunk)	27.9	30.8	29.2	43.9	48.4	46.0	19.2	22.9	20.9	34.9	41.7	38.0
BERT_Seq(doc)	35.3	35.9	35.6	52.6	53.5	53.1	20.3	25.0	22.4	35.5	42.2	38.6
MG-Reader	30.3	35.9	32.9	45.6	50.8	48.1	19.9	23.2	21.4	37.1	39.7	38.4
Doc2EDAG	37.1	36.1	36.6	54.2	53.7	53.9	21.6	26.3	23.7	36.2	45.7	40.4
BERT_QA	41.9	28.1	33.5	75.8	50.6	60.7	29.4	20.3	24.0	68.1	46.9	55.5
Ontology_QA	51.3	34.2	41.0	80.3	53.6	64.3	36.6	25.2	29.8	69.7	48.0	56.9
Human	87.8	84.2	85.9	80.9	87.2	89.0	-	-	-	-	-	-

Table 5: Overall Performance on Event argument Extraction(%).

network to help them understand the main event, while deep learning models typically fail in long text perception. 4) There is a performance degradation from the normal setting to the cross-domain setting, which shows that domain migration is still a huge challenge for current SOTA models. Among the pre-trained baselines, ALBERT's performance drops the most. The reason may be that the parameter scale in ALBERT is relatively small, and the reserved source domain knowledge is limited.

5.2 Event argument Extraction

Baselines We introduce the following mainstream baselines for evaluation: 1) **BERT_Seq** (one of the baseline in Du and Cardie (2020a)) uses the pre-trained BERT model to sequentially label words in the article. Given the input article $A = \{w_1, w_2, \ldots, w_n\}$, the output of Sequence Labeling Methods is $O = \{r_1, r_2, \ldots, r_n\}$, where $r \in R$ and R is the set of the argument types. 2) **MG-Reader** (Du and Cardie, 2020a) improves document-level EE by proposing a novel multi-granularity reader to dynamically aggregate information in sentence and paragraph-level. 3) **DocEDAG** (Zheng et al., 2019) generates an entity-based directed acyclic graph for document-level EE. 4) **BERT_QA**(Du and Cardie, 2020c) uses the argument type as question to query the article for answer. Given the input article A, the argument type $r \in R$ as the question, the output is $O = \{start_r, end_r\}$. We give -1 for these not mentioned event arguments. 5) **Ontology_QA**. Following Vargas-Vera and Motta (2004), we refine the initial query in BERT_QA with argument ontology knowledge obtained from Oxford dictionary (Dictionary, 1989).

Considering the length limitation of pre-trained models, we split the article in three different ways. (Sent) means to split the article by sentence[10]. (Chunk) means to split the article by every 128 tokens (default). (Doc) means no splitting. We adopt Longformer (Beltagy et al., 2020) as encoder for the *(doc)* baseline, and BERT-base for the other baselines.

Following prior work (Du and Cardie, 2020b), we use Head noun phrase Match (HM) and Exact Match (EM) as two evaluation metrics. HM is a relatively relaxed metric. As long as the head noun of the predicted result is consistent with the golden label, it will be judged as correct. While EM requires that the prediction result is exactly the same as the gold label, which is relatively stricter.

Overall Performance As shown in Table 5, there is a big gap between the performance of SOTA models and human performance (41.0% Vs 85.9% in F score), indicating that document-level event argument extraction remains a challenge task.

The failure of existing baselines may be due to two reasons. One possible reason is the catastrophic forgetting in neural networks. Compared to NER and sentence-level EE, document-level EE (our task) highlights the model's capability to process long texts: the model has to read the entire text before determining the argument type of a span. Although a few models have been proposed to improve the long text capabilities of pre-trained models (such as longformer), and have achieved good results, (the performance of long-former (BERT_Seq(doc)) is superior to BERT_Seq(sent) and BERT_Seq(chunk) as shown in Table 5), but these models still have a big performance gap compared with human beings.

Another reason is that existing baselines suffer from the inferior capability in semantic understanding, which is reflected in two aspects: 1) EE models

[10] https://www.nltk.org/api/nltk.tokenize.html

fail to distinguish arguments of similar events. For instance, the article mainly describes *the 2021 U.S. Alaska Peninsula earthquake*, and also briefly mentions *2008 Wenchuan earthquake*. When asking the *Date* of the main event, EE models are easy to confuse the correct answer 2021 with the wrong answer 2008. 2) EE models often mistake unrelated entities for event arguments. For example, when extracting the event argument *Attack Target* in the *the 911 terrorist attack on the Pentagon* event, except to the correct answer *the New York Pentagon*, EE models often mistake other unrelated location entities in the article (such as *Mount Sinai Hospital*) as one of the answers.

We believe that the following research directions are worthy of attention: 1) Exploring pre-trained models with stronger long text processing capabilities. 2) Exploiting ontology and commonsense knowledge to improve the semantic understanding of EE models. In the future, we will focus on promote event extraction to a higher level, such as cross-document level.

6 Conclusion

In this paper, we present DocEE, a large-scale document-level EE dataset to promote event extraction from sentence-level to document-level. Comparing to existing datasets, DocEE greatly expands the data scale, with more than 27,000+ events and 180,000+ arguments, and contains more refined event arguments. Experiments show that DocEE remains an open issue.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (2018YFB1005100) and (2018YFB1005101), National Engineering Laboratory for Cyberlearning and Intelligent Technology, Beijing Key Lab of Networked Multimedia, the Institute for Guo Qiang, Tsinghua University (2019GQB0003), the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant, and Beijing Academy of Artificial Intelligence.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Collin F. Baker. 2014. FrameNet: A knowledge base for natural language processing. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 1–5, Baltimore, MD, USA. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

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)*, pages 409–419, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Oxford English Dictionary. 1989. Oxford english dictionary. *Simpson, Ja & Weiner, Esc*.

Xiao Ding, Zhongyang Li, Ting Liu, and Kuo Liao. 2019. ELG: an event logic graph. *CoRR*, abs/1907.08015.

Zhendong Dong and Qiang Dong. 2003. Hownet - a hybrid language and knowledge resource. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 820–824.

Valentina Dragos. 2013. Developing a core ontology to improve military intelligence analysis. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 17(1):29–36.

Xinya Du and Claire Cardie. 2020a. Document-level event role filler extraction using multi-granularity contextualized encoding. *CoRR*, abs/2005.06579.

Xinya Du and Claire Cardie. 2020b. Document-level event role filler extraction using multi-granularity contextualized encoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8010–8020, Online. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020c. Event extraction by answering (almost) natural questions. *arXiv preprint arXiv:2004.13625*.

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*, pages 8057–8077, Online. Association for Computational Linguistics.

Ronen Feldman and James Sanger. 2006. *Information Extraction*, page 94–130. Cambridge University Press.

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

Simon Gottschalk and Elena Demidova. 2018. Eventkg: A multilingual event-centric temporal knowledge graph. *CoRR*, abs/1804.04526.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Felix Hamborg, Corinna Breitinger, and Bela Gipp. 2019. Giveme5w1h: A universal system for extracting main events from news articles. In *Proceedings of the 13th ACM Conference on Recommender Systems, 7th International Workshop on News Recommendation and Analytics (INRA 2019)*.

Felix Hamborg, Soeren Lachnit, Moritz Schubotz, Thomas Hepp, and Bela Gipp. 2018. Giveme5w: Main event retrieval from news articles by extraction of the five journalistic w questions.

Daniel Hienert and Francesco Luciano. 2012. Extraction of historical events from wikipedia. *CoRR*, abs/1205.4138.

Andrew Hsi. 2018. *Event Extraction for Document-Level Structured Summarization*. Ph.D. thesis, Carnegie Mellon University.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of genia event task in bionlp shared task 2011. In *Proceedings of BioNLP shared task 2011 workshop*, pages 7–15.

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. The genia event extraction shared task, 2013 edition-overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, and Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Sam N. Lehman-Wilzig and Michal Seletzky. 2010. Hard news, soft news, 'general' news: The necessity and utility of an intermediate classification. *Journalism*, 11(1):37–56.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. *CoRR*, abs/2104.05919.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besançon. 2016. A dataset for open event extraction in English. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1939–1943, Portorož, Slovenia. European Language Resources Association (ELRA).

T. Nguyen and R. Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *AAAI*.

Tomoko Ohta, Sampo Pyysalo, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Sophia Ananiadou, and Jun'ichi Tsujii. 2013. Overview of the pathway curation (pc) task of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 67–75.

Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. 2011. Overview of the epigenetics and post-translational modifications (epi) task of bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 16–25.

Walker Orr, Prasad Tadepalli, and Xiaoli Fern. 2018. Event detection with neural networks: A rigorous empirical evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 999–1004, Brussels, Belgium. Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, and Sophia Ananiadou. 2013. Overview of the cancer genetics (CG) task of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 58–66, Sofia, Bulgaria. Association for Computational Linguistics.

Carsten Reinemann, James Stanyer, Sebastian Scherr, and Guido Legnante. 2012. Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism*, 6(2):221–239.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Meihan Tong, Bin Xu, Shuai Wang, Yixin Cao, Lei Hou, Juanzi Li, and Jun Xie. 2020. Improving event detection via open-domain trigger knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5887–5897, Online. Association for Computational Linguistics.

Gaye Tuchman. 1973. Making news by doing work: Routinizing the unexpected. *American journal of Sociology*, 79(1):110–131.

Sofie Van Landeghem, Jari Björne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, et al. 2013. Large-scale event extraction from literature with multi-level gene normalization. *PloS one*, 8(4):e55814.

Maria Vargas-Vera and Enrico Motta. 2004. Aqua–ontology-based question answering system. In *Mexican International Conference on Artificial Intelligence*, pages 468–477. Springer.

Peng Wang, Zhenkai Deng, and Ruilong Cui. 2021. Tdjee: A document-level joint model for financial event extraction. *Electronics*, 10(7):824.

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. *CoRR*, abs/2004.13590.

Runxin Xu, Tianyu Liu, Lei Li, and Baobao Chang. 2021. Document-level event extraction via heterogeneous graph-based interaction model with a tracker. *CoRR*, abs/2105.14924.

Ge Zhan, Ming Wang, and Meiyi Zhan. 2020. Public opinion detection in an online lending forum: Sentiment analysis and data visualization. In *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pages 211–213.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2edag: An end-to-end document-level framework for chinese financial event extraction. *arXiv preprint arXiv:1904.07535*.

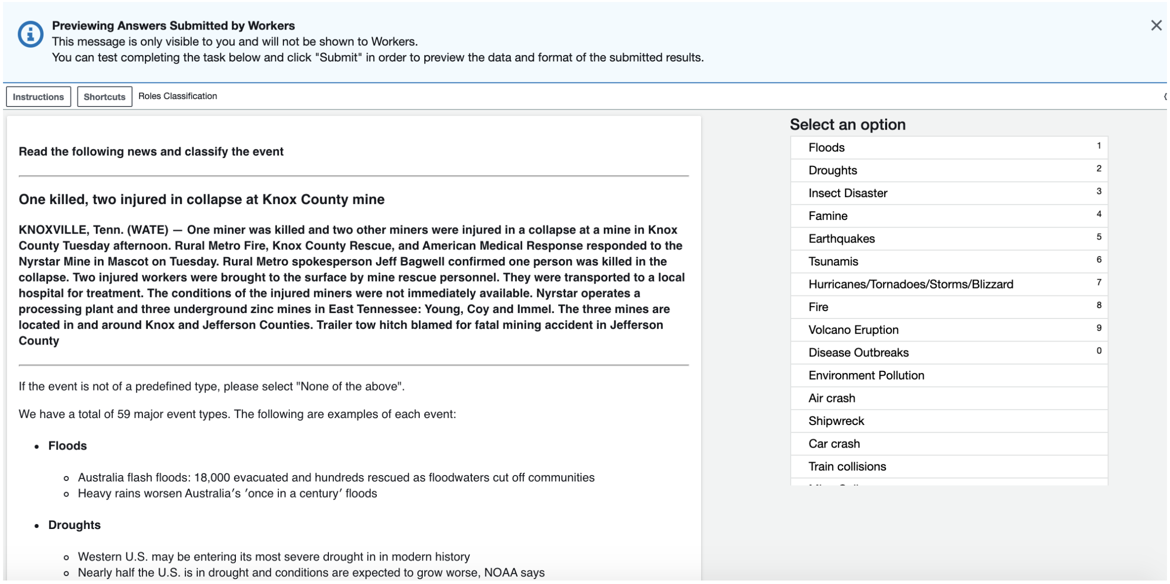

Figure 5: The online annotation page for Stage 1 event classification. The annotator needs to read the news title and the article to decide the event type of the main event. We provide examples for each event type to facilitate understanding.

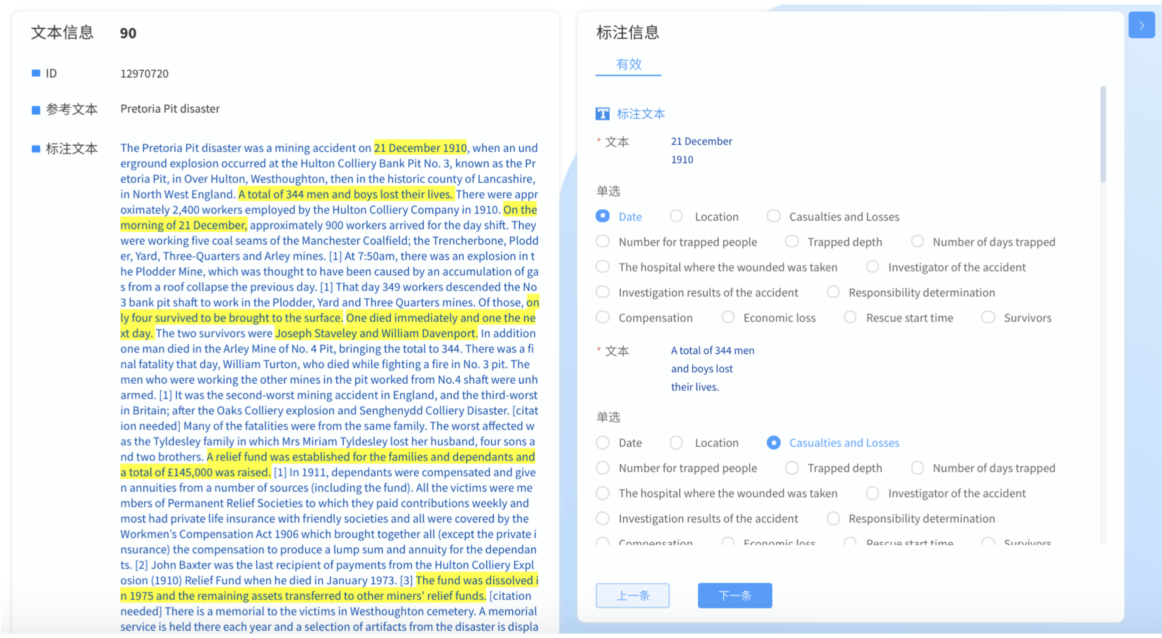

Figure 6: The online annotation page for Stage 2 event role extraction. The annotator first draws the answer from text on the left, and then selects the appropriate event role label from the label column on the right. The final drawn result will be displayed on the right.

Table 6: Event Type Statistic

Event Type	Documents	Event Type	Documents
Armed Conflict	2200	Air crash	1735
Protest	1209	CommitCrime - Accuse	1127
Election	1112	Sports Competition	1050
Fire	926	CommitCrime - Sentence	893
Government Policy Changes	838	Earthquakes	756
Riot	720	Resign	667
Famous Person - Death	569	Disease Outbreaks	550
Floods	516	Appoint	508
Storm	506	Diplomatic Talks	502
Strike	471	CommitCrime - Arrest	439
Road Crash	418	Environment Pollution	385
Organization Closed	382	Gas Explosion	370
Bank Robbery	363	Break Historical Records	359
Sign Agreement	341	Awards Ceremony	331
Famous Person - Give a speech	327	Shipwreck	321
Mine Collapses	318	CommitCrime - Release	313
Military Excercise	307	Mass Poisoning	303
Financial Crisis	302	New Achievements in Aerospace	299
Train Collisions	298	Withdraw from an Orgnization	292
Diplomatic Visit	269	Organization Merge	265
Insect Disaster	265	Organization Fine	260
CommitCrime - Investigate	260	New Wonders in Nature	255
Volcano Eruption	247	Famine	247
Famous Person - Sick	244	New Archeological Discoveries	240
Tear Up Agreement	222	Famous Person - Marriage	194
Financial Aid	189	Organization Established	167
Famous Person - Divorce	152	Tsunamis	140
Droughts	129	Mudslides	123
Famous Person - Recovered	118	Join in an Orgnization	117
Regime Change	69		

Table 7: Template for Collecting Candidate Data

Template
"List of"+event type
event type+"in"+year
"Category:"+event type+"in"+country
"Category:"+event type+"in"+year
"List of"+event type+"in"+country
"List of"+event type+"by"+property
"List of"+event type+"in"+year
"List of historical"+event type year+event type
location+event type