# EventPlus: A Temporal Event Understanding Pipeline

**Mingyu Derek Ma**[1*] **Jiao Sun**[2*] **Mu Yang**[3] **Kung-Hsiang Huang**[2]
**Nuan Wen**[2] **Shikhar Singh**[2] **Rujun Han**[2] **Nanyun Peng**[1,2]
[1] Computer Science Department, University of California, Los Angeles
[2] Information Sciences Institute, University of Southern California
[3] Texas A&M University
{ma,violetpeng}@cs.ucla.edu
{jiaosun,kunghsia,nuanwen,ssingh43,rujunhan}@usc.edu
yangmu@tamu.edu

## Abstract

We present EventPlus, a temporal event understanding pipeline that integrates various state-of-the-art event understanding components including *event trigger and type detection*, *event argument detection*, *event duration* and *temporal relation extraction*. Event information, especially event temporal knowledge, is a type of common sense knowledge that helps people understand how stories evolve and provides predictive hints for future events. EventPlus as the first comprehensive temporal event understanding pipeline provides a convenient tool for users to quickly obtain annotations about events and their temporal information for any user-provided document. Furthermore, we show EventPlus can be easily adapted to other domains (e.g., biomedical domain). We make EventPlus publicly available to facilitate event-related information extraction and downstream applications.

## 1 Introduction

Event understanding is intuitive for humans and important for daily decision making. For example, given the raw text shown in Figure 1, a person can infer lots of information including *event trigger and type*, *event related arguments* (e.g., agent, patient, location), *event duration* and *temporal relations* between events based on the linguistic and common sense knowledge. These understandings help people comprehend the situation and prepare for future events. The event and temporal knowledge are helpful for many downstream applications including question answering (Meng et al., 2017; Huang et al., 2019), story generation (Peng et al., 2018; Yao et al., 2019; Goldfarb-Tarrant et al., 2019, 2020), and forecasting (Wang et al., 2017; Granroth-Wilding and Clark, 2016; Li et al., 2018).

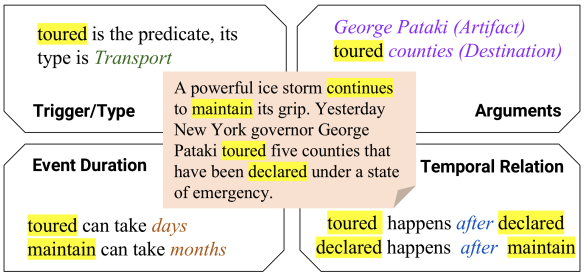


Figure 1: Event understanding components. We highlight events triggers in yellow, and mark the predicted task-related information in *Italic*.

Despite the importance, there are relatively few tools available for users to conduct text-based temporal event understanding. Researchers have been building natural language processing (NLP) analysis tools for "core NLP" tasks (Gardner et al., 2018; Manning et al., 2014; Khashabi et al., 2018). However, systems that target at semantic understanding of events and their temporal information are still under-explored. There are individual works for event extraction, temporal relation detection and event duration detection, but they are separately developed and thus cannot provide comprehensive and coherent temporal event knowledge.

We present EventPlus, the *first* pipeline system integrating several high-performance temporal event information extraction models for comprehensive temporal event understanding. Specifically, EventPlus contains event extraction (both on defined ontology and for novel event triggers), event temporal relation prediction, event duration detection and event-related arguments and named entity recognition, as shown in Figure 2.[1]

[*] Equal contribution.

[1] The system is publicly accessible at https://kairos-event.isi.edu. The source code is available at https://github.com/PlusLabNLP/EventPlus. We also provide an introductory video at https://pluslabnlp.github.io/eventplus.



*Proceedings of NAACL-HLT 2021: Demonstrations*, pages 56–65
June 6–11, 2021. 

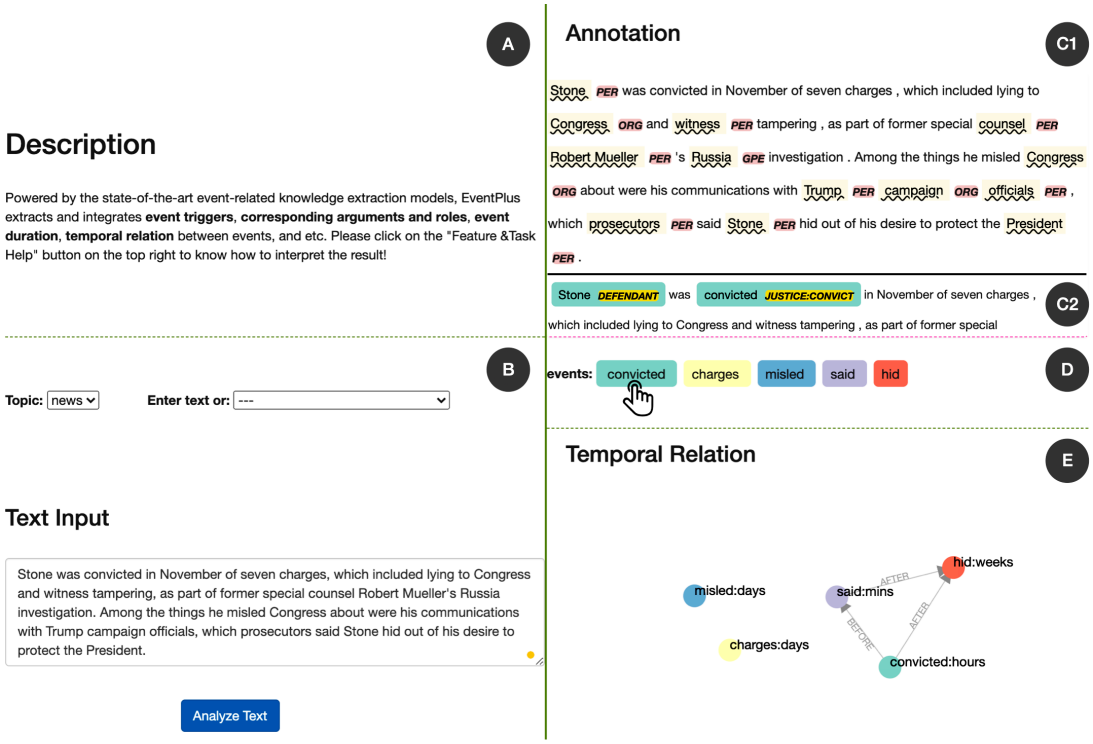


Figure 2: The interface of EventPlus. Users can either choose examples or freely input text which matches with their selected topic in $B$. $C$ shows the Name Entity Recognition (NER) results, which serve as argument candidates for events. When clicking on an event trigger in $D$, we show the selected event trigger and its corresponding *arguments* in $C2$. We show *temporal-related information* of all events in $E$, where nodes represent event triggers and edges represent their relations; we further indicate the *event duration* as labels of nodes.

EventPlus is designed with multi-domain support in mind. Particularly, we present an initial effort to adapt EventPlus to the biomedical domain. We summarize the contributions as follows:

- We present the first event pipeline system with comprehensive event understanding capabilities to extract *event triggers and argument*, *temporal relations* among events and *event duration* to provide an event-centric natural language understanding (NLU) tool to facilitate downstream applications.
- Each component in EventPlus has comparable performance to the state-of-the-art, which assures the quality and efficacy of our system for temporal event reasoning.

## 2 Component

In this section, we introduce each component in our system, as shown in Figure 3. We use a multi-task learning model for event trigger and temporal relation extraction (§ 2.1). The model introduced in § 2.2 extracts semantic-rich events following the ACE ontology, and the model introduced in § 2.3 predicts the event duration. Note that our system handles two types of event representations: one represents an event as the trigger word (Pustejovsky et al., 2003a) (as the event extraction model in § 2.1), the other represents event as a complex structure including trigger, type and arguments (Ahn, 2006) (as the event extraction model in § 2.2). The corpus following the former definition usually has a broader coverage while the latter can provide richer information. Therefore, we develop models to combine the benefits of both worlds. We also introduce a speculated and negated events handling component in § 2.4 to further identify whether an event happens or not.

### 2.1 Multi-task Learning of Event Trigger and Temporal Relation Extraction

The event trigger extraction component takes the input of raw text and outputs single-token event triggers. The input to the temporal relation extraction model is raw text and a list of detected event triggers. The model will predict temporal relationships between each pair of events. In previous literature (Han et al., 2019b), multi-task learning of these two tasks can significantly improve performance on both tasks following the intuition that

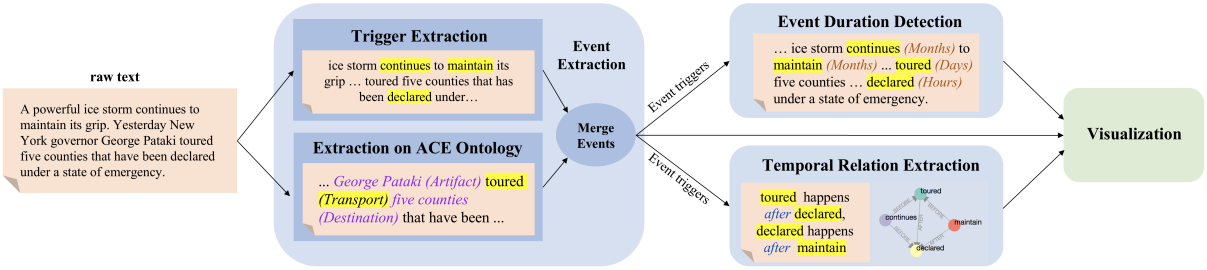


Figure 3: Overall system design of EventPlus. The raw text is first fed into two event extraction components, and then we pass the event triggers of the merged event list to event duration detection and temporal relation extraction models. Finally outputs from all models are combined for visualization.

event relation signals can be helpful to distinguish event triggers and non-event tokens.

The model feeds BERT embedding (Devlin et al., 2019) of the input text to a shared BiLSTM layer for encoding task-specific contextual information. The output of the BiLSTM is passed to an event scoring function and a relation scoring function which are MLP classifiers to calculate the probability of being an event (for event extraction) or a probability distribution over all possible relations (for temporal relation extraction). We train the multi-task model on MATRES (Ning et al., 2018a) containing temporal relations BEFORE, AFTER, SIMULTANEOUS and VAGUE. Though the model performs both tasks during training, it can be separately used for each individual task during inference.

## 2.2 Event Extraction on ACE Ontology

Although event triggers present the occurrence of events, they are not sufficient to demonstrate the semantic-rich information of events. ACE 2005[2] corpus defines an event ontology that represents an event as a structure with triggers and corresponding event arguments (participants) with specific roles (Doddington et al., 2004).[3] Our system is trained with ACE 2005 corpus, thus it is capable of extracting events with the complex structure. ACE focuses on events of a particular set of types including LIFE, MOVEMENT, TRANSACTION, BUSINESS, CONFLICT, CONTACT, PERSONNEL and JUSTICE, where each type has corresponding sub-types. Following prior works (Wadden et al., 2019; Lin et al., 2020), we keep 7 entity types (person, organization, location, geo-political entity, facility, vehicle, weapon), 33 event sub-types, and 22 argument roles that are associated with sub-types.

Similar to Han et al. (2019b), we build our event extraction component for ACE ontology upon a multi-task learning framework that consists of trigger detection, argument role detection and entity detection. These tasks share the same BERT encoder, which is fine-tuned during training. The entity detector predicts the argument candidates for all events in an input sentence. The trigger detector labels the input sequence with the event sub-types at the token level. The argument role detector finds the argument sequence[4] for each detected trigger via attention mechanism. For example, for the sentence in Figure 1, its target trigger sequence has MOVEMENT:TRANSPORT label at the position of "toured" token, and its argument sequence for this MOVEMENT:TRANSPORT event has B-ARTIFACT, I-ARTIFACT labels at the position of "George Pataki" and B-DESTINATION label at the position of "counties" respectively. The entire multi-task learning framework is jointly trained.

During inference, our system detects arguments solely based on triggers. To make our system better leverage information from argument candidates, we developed the following constraints during decoding based on the predicted entities (argument candidates) and other specific definitions in ACE:

- Entity-Argument constraint. The argument role label for a token can take one of the 22 argument roles if and only if the token at this position belongs to a predicted entity.
- Entity-Trigger constraint. The trigger label for a token can take one of the 33 event sub-types if and only if the token at this position does not belong to a predicted entity.
- Valid Trigger-Argument constraint. Based on the definitions in ACE05, each event sub-type takes

[2] https://www.ldc.upenn.edu/collaborations/past-projects/ace

[3] The ACE program provides annotated data for five kinds of extraction targets: entities, times, values, relations and events. We only focus on events and entities data in this paper.

[4] Argument sequences are presented using BIO encoding.

certain types of argument roles. We enforce that given the predicted trigger label, the argument roles in this sequence can only take those that are valid for this trigger.

To account for these constraints, we set the probability of all invalid configurations to be 0 during decoding.

### 2.3 Event Duration Detection

This component classifies event triggers into duration categories. While many datasets have covered time expressions which are explicit timestamps for events (Pustejovsky et al., 2003b; Cassidy et al., 2014; Reimers et al., 2018; Bethard et al., 2017), they do not target categorical event duration. To supplement this, Vashishtha et al. (2019) introduces the UDS-T dataset, where they provide 11 duration categories which we adopt for our event pipeline: INSTANT, SECONDS, MINUTES, HOURS, DAYS, WEEKS, MONTHS, YEARS, DECADES, CENTURIES and FOREVER. Pan et al. (2006) also present a news domain duration annotation dataset containing 58 articles developed from TimeBank corpus (we refer as Typical-Duration in the following), it provides 7 duration categories (a subset of the 11 categories in UDS-T from SECONDS to YEARS).

We developed two models for the event duration detection task. For a sentence, along with predicate root and span, the models perform duration classification. In the first method, we fine-tune a BERT language model (Devlin et al., 2019) on single sentences and take hidden states of event tokens from the output of the last layer, then feed into a multi-layer perceptron for classification.

The second model is adapted from the UDS-T baseline model, which is trained under the multi-task objectives of duration and temporal relation extraction. The model computes ELMo embeddings (Peters et al., 2018) followed by attention layers to compute the attended representation of the predicate given sentence. The final MLP layers extract the duration category. Even though this model can detect temporal relations, it underperforms the model we described in § 2.1, so we exclude the temporal relation during inference.

### 2.4 Negation and Speculation Cue Detection and Scope Resolution

The event extraction components described above are designed to extract all possible events, but we identify events that are indicated by speculation (e.g., *would*) or negation (e.g., *not*) keywords (Konstantinova et al., 2012). Since those events do not happen, we mark them with special labels. For example, in the sentence "The United States is *not* considering sending troops to Mozambique", we identify "send" will not happen.

We adapt the BERT-based negation and speculation cue detection model and the scope resolution model introduced by Khandelwal and Sawant (2020). To fine-tune these models, we use the SFU Review dataset with negation and speculation annotations (Taboada et al., 2006; Taboada and Grieve, 2004; Konstantinova et al., 2012), and we feed ground truth negation and speculation cues as input for the scope resolution model. We evaluate the two models on a separate testing set of the SFU Review dataset. The cue detection model yields a 0.92 F1 score, and the scope resolution model yields a 0.88 F1 score for token-level prediction, given ground truth cues as input. In EventPlus, we input cues detected by the cue detection model to the scope resolution model.

## 3 System

We design a pipeline system to enable the interaction among components with state-of-the-art performance introduced in § 2 and provide a comprehensive output for events and visualize the results. Figure 3 shows the overall system design.

### 3.1 Pipeline Design

**Event Extraction** EventPlus takes in raw text and feeds the tokenized text to two event extraction modules trained on ACE ontology-based datasets and free-formatted event triggers. The ACE ontology extraction modules will produce the output of event triggers ("toured" is a trigger), event type (it is a MOVEMENT:TRANSPORT event), argument and its role (the ARTIFACT is "George Pataki" and DESTINATION is "counties") and NER result ("New York" and "counties" are GEO-POLITICAL ENTITY and "governer" and "George Pataki" are PERSON). The trigger-only extraction model will produce all event triggers ("continues", "maintain" and "declared" are also event triggers but we do not have arguments predicted for them). Then trigger-only events will be merged to ACE-style events list and create a combined event list from the two models. For each extracted event, if it is in the negation or speculation scope predicted by the cue detection and scope resolution component, then we add a

"speculation or negation" argument to that event.

**Duration Detection and Temporal Relation Extraction** The combined events list will be passed to the event duration detection model to detect duration for each of the extracted events ("tours" will take DAYS etc.) and passed to temporal relation extraction component to detect temporal relations among each pair of events ("toured" is after "declared" etc.). Note that duration and temporal relation extraction are based on the context sentence besides the event triggers themselves and they are designed to consider contextualized information contained in sentences. Therefore "take (a break)" can take MINUTES in the scenario of "Dr. Porter is now taking a break and will be able to see you soon" but take DAYS in the context of "Dr. Porter is now taking a Christmas break" (Ning, 2019).

**Visualization** To keep the resulted temporal graph clear, we remove predicted VAGUE relations since that indicates the model cannot confidently predict temporal relations for those event pairs. Finally, all model outputs are gathered and pass to the front-end for visualization.

### 3.2 Interface Design

Figure 2 shows the interface design of EventPlus.[5] We display the NER result with wavy underlines and highlight event triggers and corresponding arguments with the same color upon clicks. Besides, we represent the temporal relations among events in a directed graph using d3 [6] if there are any, where we also indicate each event's duration in the label for each event node.

## 4 Evaluation

Each capability in the pipeline has its own input and output protocol, and they require various datasets to learn implicit knowledge independently. In this section, we describe the performance for each capability on corresponding labeled datasets.

### 4.1 Event Trigger Extraction

We report the evaluation about event triggers extraction component on TB-Dense (Cassidy et al., 2014) and MATRES (Ning et al., 2018a), two event extraction datasets in the news domain (Han et al., 2019b). We show the result in Table 1. Comparing the performance on TB-Dense with CAEVO (Chambers et al., 2014), DEER (Han et al., 2020a) and MATRES performance with Ning et al. (2018b), the model we use achieves best F1 scores and yields the state-of-the-art performance.

| Corpus | Model | F1 |
|---|---|---|
| TB-Dense | Chambers et al. (2014) | 87.4 |
| | Han et al. (2020a) | 90.3 |
| | Ours | **90.8** |
| MATRES | Ning et al. (2018b) | 85.2 |
| | Ours | **87.8** |

Table 1: Evaluation for event trigger extraction

### 4.2 Event Extraction on ACE Ontology

We evaluate our event extraction component on the test set of ACE 2005 dataset using the same data split as prior works (Lin et al., 2020; Wadden et al., 2019). We follow the same evaluation criteria:

- Entity: An entity is correct if its span and type are both correct.
- Trigger: A trigger is correctly **identified** (Trig-I) if its span is correct. It is correctly **classified** (Trig-C) if its type is also correct.
- Argument: An argument is correctly **identified** (Arg-I) if its span and event type are correct. It is correctly **classified** (Arg-C) if its role is also correct.

In Table 2, we compare the performance of our system with the current state-of-the-art method OneIE (Lin et al., 2020). Our system outperforms OneIE in terms of entity detection performance. However our trigger and argument detection performance is worse than it. We leave the improvements for triggers and arguments for future work.

| Model | Entity | Trig-I | Trig-C | Arg-I | Arg-C |
|---|---|---|---|---|---|
| OneIE | 90.2 | **78.2** | **74.7** | **59.2** | **56.8** |
| Ours | **91.3** | 75.8 | 72.5 | 57.7 | 55.7 |

Table 2: Test set performance on ACE 2005 dataset. Following prior works, we use the same evaluation criteria: *-I represent Trigger or Argument Identification. *-C represent Trigger or Argument Classification.

### 4.3 Event Duration Detection

We evaluate the event duration detection models on Typical-Duration and newly annotated ACE-Duration dataset to reflect the performance on

[5] We have a walk-through instruction available to help first-time end users get familiar with EventPlus. Please see our video for more information.

[6] https://d3js.org/

generic news domain for which our system is optimized. Since UDS-T dataset (Vashishtha et al., 2019) is imbalanced and has limited samples for some duration categories, we do not use it as an evaluation benchmark but we sample 466 high IAA data points as training resources. We split Typical-Duration dataset and use 1790 samples for training, 224 for validation and 224 for testing.

To create ACE-Duration, we sample 50 unique triggers with related sentences from the ACE dataset, conduct manual annotation with three annotators and take the majority vote as the gold duration category. Given natural ordering among duration categories, the following metrics are employed: accuracy over 7 duration categories (Acc), coarse accuracy (Acc-c, if the prediction falls in categories whose distance to the ground truth is 1, it is counted as correct) and Spearman correlation (Corr).

| | Typical-Duration | | | ACE-Duration | | |
|---|---|---|---|---|---|---|
| Model | Acc | Acc-c | Corr | Acc | Acc-c | Corr |
| UDS-T (U) | 0.20 | 0.54 | 0.59 | 0.38 | 0.68 | 0.62 |
| UDS-T (T) | 0.52 | 0.79 | 0.71 | 0.47 | 0.67 | 0.50 |
| UDS-T (T+U) | 0.50 | 0.76 | 0.68 | **0.49** | 0.74 | 0.66 |
| BERT (T) | **0.59** | **0.81** | **0.75** | 0.31 | 0.67 | 0.64 |
| BERT (T+U) | 0.56 | **0.81** | 0.73 | 0.45 | **0.79** | **0.70** |

Table 3: Event duration detection experimental result. Typical-Duration results are from testing subset. Notations in the bracket of model names indicate resources for training, U: 466 UDS-T high IAA samples, T: Typical-Duration training set

Experimental results in Table 3 show the BERT model is better than UDS-T ELMo-based model in general and data augmentation is especially helpful to improve performance on ACE-Duration. Due to the limited size of ACE-Duration, we weight more on the Typical-Duration dataset and select BERT (T) as the best configuration. To the best of our knowledge, this is the state-of-the-art performance on the event duration detection task.

### 4.4 Temporal Relation Extraction

We report temporal relation extraction performance on TB-Dense and MATRES datasets. TB-Dense consider the duration of events so the labels are INCLUDES, INCLUDED IN, BEFORE, AFTER, SIMULTANEOUS and VAGUE, while MATRES uses start-point as event temporal anchor and hence its labels exclude INCLUDES and INCLUDED IN. In EventPlus, we augment extracted events from multiple components, so we report temporal relation extraction result given golden events as relation candidates to better reflect single task performance.

| Corpus | Model | F1 |
|---|---|---|
| TB-Dense | Vashishtha et al. (2019) | 56.6 |
| | Meng and Rumshisky (2018) | 57.0 |
| | Ours | **64.5** |
| MATRES | Ning et al. (2018b) | 65.9 |
| | Ning et al. (2018a) | 69.0 |
| | Ours | **75.5** |

Table 4: Experimental result for temporal relation extraction given golden event extraction result

Table 4 shows the experimental results.[7] Our model in § 2.1 achieves the best result on temporal relation extraction and is significantly better than (Vashishtha et al., 2019) mentioned in § 2.3.[8]

## 5 Extension to Biomedical Domain

With our flexible design, each component of EventPlus can be easily extended to other domains with little modification. We explore two approaches to extend the event extraction capability (§ 2.2) to the biomedicine domain: 1) multi-domain training (MDT) with GENIA (Kim et al., 2009), a dataset containing biomolecular interaction events from scientific literature, with shared token embeddings, which enables the model to predict on both news and biomedical text; 2) replace the current component with an in-domain event extraction component **SciBERT-FT** (Huang et al., 2020) which is a biomedical event extraction system based on fine-tuned SciBERT (Beltagy et al., 2019).

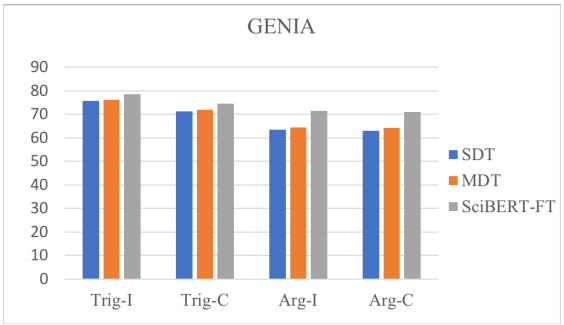


Figure 4: Performance comparison of single-domain training (SDT), multi-domain training (MDT) and SciBERT-FT on the Dev set of GENIA

[7]The MATRES experiment result in Table 4 uses 183 documents for training and 20 for testing developed from the entire TempEval-3 dataset. Han et al. (2019a) reports higher F1 score but it uses a subset of MATRES (22 documents for train, 5 for dev and 9 for test) and has different setting.

[8]The latest state-of-the-art work (Han et al., 2020a) only reports end-to-end event extraction and temporal relation extraction result, pure temporal relation extraction result given ground-truth events are not provided. We are not able to compare with it directly.

While MDT on ACE and GENIA datasets from different domains improves the performance on GENIA, it is still lower than **SciBERT-FT** (Figure 4). Therefore, we decide to pursue the second extension approach to incorporate **SciBERT-FT** and extend EventPlus to the biomedical domain.

## 6 Related Works

Existing NLP toolkits (Manning et al., 2014; Khashabi et al., 2018) provide an interface for a set of useful models. Some tools integrate several models in a pipeline fashion (Peng et al., 2015; Noji and Miyao, 2016). The majority of these systems focus on token-level tasks like tokenization, lemmatization, part-of-speech tagging, or sentence-level tasks like syntactic parsing, semantic role labeling etc. There are only a few systems that can provide capabilities of event extraction and temporal information detection (Tao et al., 2013; Ning, 2019).

For event extraction, some systems only provide results within a certain defined ontology such as AIDA (Li et al., 2019), there are also some works utilizing data from multiple modalities (Li et al., 2020a,b). Some works could handle novel events (Xiang and Wang, 2019; Ahmad et al., 2021; Han et al., 2020b; Huang and Peng, 2020), but they are either restricted to a certain domain (Yang et al., 2018) or lack of performance superiority because of their lexico-syntactic rule-based algorithm (Valenzuela-Escárcega et al., 2015). For temporal information detection, Ning et al. (2019) proposes a neural-based temporal relation extraction system with knowledge injection. Most related to our work, Ning et al. (2018b) demonstrates a temporal understanding system to extract time expression and implicit temporal relations among detected events, but this system cannot provide event-related arguments, entities and event duration information.

These previous works either are not capable of event understanding or just focus on one perspective of event-related features. There is no existing system that incorporates a comprehensive set of event-centric features, including event extraction and related arguments and entities, temporal relations, and event duration.

## 7 Conclusion and Future Work

We represent EventPlus, a pipeline system that takes raw texts as inputs and produces a set of temporal event understanding annotations, including *event trigger and type*, *event arguments*, *event duration* and *temporal relations*. To the best of our knowledge, EventPlus is the first available system that provides such a comprehensive set of temporal event knowledge extraction capabilities with state-of-the-art components integrated. We believe EventPlus will provide insights for understanding narratives and facilitating downstream tasks.

In the future, we plan to further improve EventPlus by tightly integrating event duration prediction and temporal relation extraction modules. We also plan to improve the performance for triggers and arguments detection under the ACE ontology and develop joint training models to optimize all event-related features in an end-to-end fashion.

## Acknowledgments

Many thanks to Yu Hou for the quality assessment annotations, to Fred Morstatter and Ninareh Mehrabi for feedback on the negation and speculation event handling, and to the anonymous reviewers for their feedback. This material is based on research supported by DARPA under agreement number FA8750-19-2-0500. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

## References

Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Gate: Graph attention transformer encoder for cross-lingual relation and event extraction. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*.

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

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

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the*

*11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

G Doddington, A Mitchell, M Przybocki, L Ramshaw, S Strassel, and R Weischedel. 2004. Automatic content extraction (ace) program: task definitions and performance measures. In *Proceedings of the Fourth International Language Resources and Evaluation Conference (LREC'04)*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338.

Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. 2019. Plan, write, and revise: an interactive system for open-domain story generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 89–97.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019a. Deep structured neural network for event temporal relation extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 666–106, Hong Kong, China. Association for Computational Linguistics.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019b. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 434–444, Hong Kong, China. Association for Computational Linguistics.

Rujun Han, Xiang Ren, and Nanyun Peng. 2020a. Deer: A data efficient language model for event temporal reasoning. *arXiv preprint arXiv:2012.15283*.

Rujun Han, Yichao Zhou, and Nanyun Peng. 2020b. Domain knowledge empowered structured neural net for end-to-end event temporal relation extraction. In *the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5717–5729. Association for Computational Linguistics.

Kung-Hsiang Huang and Nanyun Peng. 2020. Efficient end-to-end learning of cross-event dependencies for document-level event extraction. *ArXiv*, abs/2010.12787.

Kung-Hsiang Huang, Mu Yang, and Nanyun Peng. 2020. Biomedical event extraction with hierarchical knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1277–1285, Online. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401.

Aditya Khandelwal and Suraj Sawant. 2020. Negbert: A transfer learning approach for negation detection and scope resolution. *ArXiv*, abs/1911.04211.

Daniel Khashabi, Mark Sammons, Ben Zhou, Tom Redman, Christos Christodoulopoulos, Vivek Srikumar, Nick Rizzolo, Lev Ratinov, Guanheng Luo, Quang Do, et al. 2018. Cogcompnlp: Your swiss army knife for nlp. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado. Association for Computational Linguistics.

Natalia Konstantinova, Sheila CM De Sousa, Noa P Cruz Díaz, Manuel J Mana López, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Lrec*, pages 3190–3195.

Manling Li, Ying Lin, Joseph Hoover, Spencer Whitehead, Clare Voss, Morteza Dehghani, and Heng Ji. 2019. Multilingual entity, relation, event and human value extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 110–115, Minneapolis, Minnesota. Association for Computational Linguistics.

Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Daniel Napierski, and Marjorie Freedman. 2020a. GAIA: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, Online. Association for Computational Linguistics.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020b. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online. Association for Computational Linguistics.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. *arXiv preprint arXiv:1805.05081*.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Yuanliang Meng and Anna Rumshisky. 2018. Context-aware neural model for temporal information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 527–536, Melbourne, Australia. Association for Computational Linguistics.

Yuanliang Meng, Anna Rumshisky, and Alexey Romanov. 2017. Temporal information extraction for question answering using syntactic dependencies in an lstm-based architecture. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 887–896.

Qiang Ning. 2019. *Understanding time in natural language text*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6203–6209, Hong Kong, China. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018a. A multi-axis annotation scheme for event temporal relations. In *ACL*.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018b. CogCompTime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77, Brussels, Belgium. Association for Computational Linguistics.

Hiroshi Noji and Yusuke Miyao. 2016. Jigg: A framework for an easy natural language processing pipeline. In *Proceedings of ACL-2016 System Demonstrations*, pages 103–108, Berlin, Germany. Association for Computational Linguistics.

F. Pan, Rutu Mulkar-Mehta, and J. Hobbs. 2006. Learning event durations from event descriptions. In *ACL*.

Nanyun Peng, Francis Ferraro, Mo Yu, Nicholas Andrews, Jay DeYoung, Max Thomas, Matthew R. Gormley, Travis Wolfe, Craig Harman, Benjamin Van Durme, and Mark Dredze. 2015. A concrete Chinese NLP pipeline. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 86–90, Denver, Colorado. Association for Computational Linguistics.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003a. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003b. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2018. Event time extraction with a decision tree of neural classifiers. *Transactions of the Association for Computational Linguistics*, 6:77–89.

Maite Taboada, Caroline Anthony, and Kimberly D Voll. 2006. Methods for creating semantic orientation dictionaries. In *LREC*, pages 427–432.

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Re# port SS# 04# 07), Stanford University, CA, pp. 158q161. AAAI Press*.

Fangbo Tao, Kin Hou Lei, Jiawei Han, ChengXiang Zhai, Xiao Cheng, Marina Danilevsky, Nihit Desai, Bolin Ding, Jing Ge Ge, Heng Ji, et al. 2013. Eventcube: multi-dimensional search and mining of structured and text data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1494–1497.

Marco A. Valenzuela-Escárcega, Gus Hahn-Powell, Mihai Surdeanu, and Thomas Hicks. 2015. A domain-independent rule-based framework for event extraction. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 127–132, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy. Association for Computational Linguistics.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Zhongqing Wang, Yue Zhang, and Ching Yun Chang. 2017. Integrating order information and event relation for script event prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 57–67.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. DCFEE: A document-level Chinese financial event extraction system based on automatically labeled training data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55, Melbourne, Australia. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Weischedel Ralph, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*.