Contrastive Event Extraction Using Video Enhancements

Anonymous ACL submission

Abstract

Event extraction aims to extract information of triggers associated with arguments from texts. Recent advanced methods consider the multi-modality to tackle the task by pairing the modalities without guaranteeing the alignment of event information across modalities, which negatively impacts on the model performances. To address the issue, we firstly constructed the Text Video Event Extraction (TVEE) dataset with an inner annotator agreement of 83.4%, containing 7,598 pairs of text-videos, each of which is connected by event alignments. To the best of our knowledge, this is the first multimodal dataset with aligned event information in each sentence and video pair. Secondly, we present a **C**ontrastive **L**earning based **E**vent **E**xtraction model with enhancements from the **V**ideo modality (CLEEV) to pair videos and texts using event information. CLEEV constructs negative samples by measuring event weights based on occurrences of event types to enhance the contrast. We conducted experiments on the TVEE and VM2E2 datasets by incorporating modalities to assist the event extraction, outperforming SOTA methods with 1.0 and 1.2 point percentage improvements in terms of F-score, respectively. Our experimental results show that the multimedia information improves the event extraction from the textual modality[1].

1 Introduction

Event Extraction (EE) aims to identify triggers and associated arguments, playing crucial role in downstream tasks such as timeline summarization (Li et al., 2021; Martschat and Markert, 2018) and text summarization (Daiya, 2020; Chen et al., 2021b). Most research focuses on textual modality of EE (Chen et al., 2015; Nguyen et al., 2016; Du et al., 2021), leaving event information across additional modalities such as image, video under investigation (Zhang et al., 2017; Li et al., 2020; Chen et al., 2021a). Multi-modal data, any combination of texts, images and videos, most often contains more information clues for event understanding than single modality. For example, as shown in Figure 1 (a), the rocket launching event is described in both text and video, the trajectory of the rocket depicted in the video makes it easy to understand that this is a *Movement.Transport* event rather than others. However, it is difficult to obtain the event with the left image only, where the rocket is static, triggering the need of video modality in addition to images for better event understanding. Initial efforts on multi-modal EE mainly consider image modality only without the video modality (Zhang et al., 2017; Tong et al., 2020; Li et al., 2020). Contrative learning methods (Zolfaghari et al., 2021; **?**;

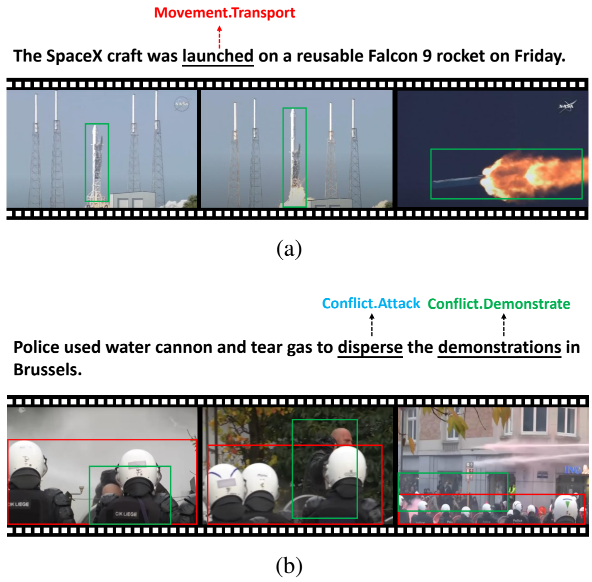


Figure 1: Two examples from the TVEE dataset. Entities from videos are annotated by boxes. Events (i.e., "launched", "disperse", "demonstrations") from the sentences are highlighted using the underscores.

[1] The dataset and code will be released based on acceptance.

Zhang et al., 2021b) have been proven to be successful on cross-modality representation learning. Recent methods (Chen et al., 2021a) propose to pre-train on videos with their auto-generated ASR transactions in a contrastive learning manner to pair the modalities of texts and videos and use the text video pairs for further event extraction. However, those multi-modal contrastive methods pair across modalities without aligning the event information on the sentence level. This inevitably introduces mis-alignments of events for paired instances, negatively impacting the EE models. Furthermore, they construct negative samples without differentiating their event-specific contribution. This limits the learning ability of the contrastive methods since events composed in the video of negative samples carry different information, resulting into the different contributions. For example, in Figure 1 (b), the *Conflict.Attack* event weights more than the *Conflict.Demonstrate* event.

To address this issue, we firstly construct a novel dataset named TVEE, which is composed of pairs of sentences and videos with aligned event information, i.e. sentence and video in a pair are describing the same events. To encode the task-specific (i.e., EE) multi-modal representation, we present a **C**ontrastive **L**earning based **E**vent **E**xtraction model enhanced by **V**ideo modality (CLEEV) with two modules: Event Extractor (EvE) and Video enhanced Event Contrastive Learner (ViECL). The EvE responds for the extraction of event triggers and arguments from the textual modality with a stack of a BERT model and two CRF layers. The ViECL assigns the weights of the event information when learning the representation across modalities on top of the contrastive learning.

We summarize our contributions as follows:

- To the best of our knowledge, we provide a benchmark dataset named TVEE, which is the first dataset that pairs texts and videos using same event descriptions to guarantee the event alignment. The dataset consists of 7,598 pairs, which are annotated with 33 event types.
- We present a contrastive model that weighs event information based on their occurrences to extract events by incorporating the video modality as assistance.
- We conducted experiments on two benchmark datasets TVEE and VM2E2 (Chen et al., 2021a) and improved the SOTA results with 1.0 and 1.2 point percentage improvements on event extraction in terms of F-score, showing the effectiveness of the video modality for event extraction in comparison with unimodal.

2 Proposed Model

We present the proposed model in Figure 2, which contains two modules: (1) EvE is a stack of the BERT model (Kenton and Toutanova, 2019) and two CRF layers for labeling the input sequence with event types and argument roles. (Section 2.2) (2) ViECL contrasts pairs between videos and texts by weighing event information based on event occurrences when constructing negative samples(Section 2.3). We present the notations in the model followed by the module details.

2.1 Notation

Inputs to the model are K pairs of sentences and videos $\{(x_i, v_i)\}_{i=1}^{K}$, where the k^{th} sentence is denoted as $x_k = \{w_1, w_2, ..., w_n\}$ with ground-truth labels $y_k = \{y_1, y_2, ..., y_n\}$ and the corresponding video is presented as $v_k = \{f_1, f_2, ..., f_m\}$ with m frames. For simplicity, we omit the subscript k. In addition, we use $r \in R$ and $e \in E$ to represent each trigger and event type, respectively.

2.2 Event Extractor

The EvE deals with the extraction of triggers and arguments from the textual modality using the trigger extractor and argument extractor, respectively.

Trigger Extractor Given an input sentence x, we firstly feed the sentence to the BERT model (i.e., text encoder) to produce the contextualized representation $\mathbf{s} \in \mathbb{R}^{n \times d}$, where d is the dimension. Then a CRF layer is stacked on top of the text encoder to label triggers with the following loss equation:

$$\mathcal{L}_t = -\sum_{i=1}^{K}\sum_{j=1}^{n} \log P(y_j|\mathbf{s}_i)$$

Argument Extractor Given a trigger r and its event type e, we obtain the trigger vector representation $\mathbf{r}$ using the span vector in $\mathbf{s}$ and embed e with an Embedding Layer to get its representation $\mathbf{e}$. Then $\mathbf{r}$ and $\mathbf{e}$ are concatenated with the sequence representation $\mathbf{s}$. The argument entities are labeled by another CRF layer:

$$\mathcal{L}_a = -\sum_{i=1}^{K}\sum_{j=1}^{n} \log P(y_j|\mathbf{s}_i; \mathbf{r}^i; \mathbf{e}^i)$$

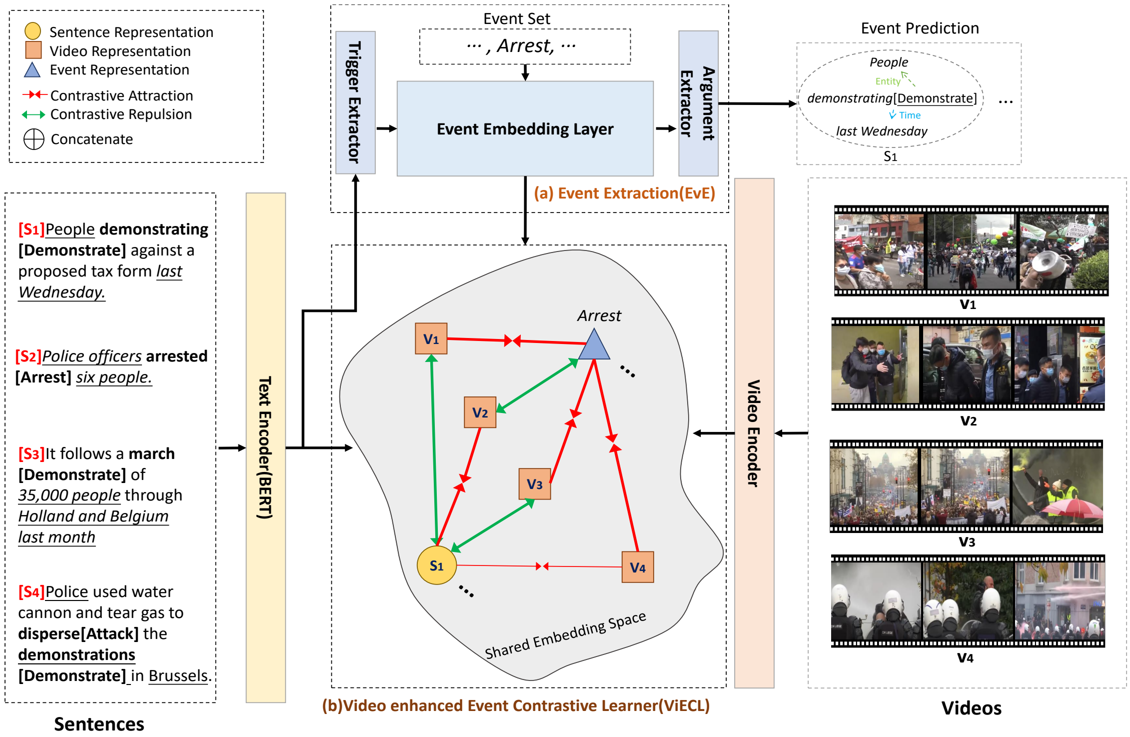


Figure 2: An overview of the proposed CLEEV model consisting of the event extractor EvE) (shown in (a)) and video enhanced event contrasive learner (ViECL) (shown in (b)). (a) presents the event extraction including trigger extractor and argument extractor. (b) presents the contrastive learning by incorporating event information. For simplicity, we use the pair (V_1, S_1) as the positive instance, where V_1 and S_1 are paired with the rest sentences and videos to construct negative examples.

2.3 Video Enhanced Event Contrastive Learner

The ViECL aims to enhance event extraction using the additional video modality by contrasting their event information. Specifically, we design two loss functions to enhance sentence and event representations respectively and incorporate event content to weigh negative samples. For a video v, we use a 3D-CNN based pre-trained model as video encoder and obtain its vector representation $\mathbf{v} \in \mathbb{R}^{m \times d}$ using a mean pooling layer.

Contrastive losses Intuitively, the distance of representations between $\mathbf{s}$ and video $\mathbf{v}$ describing the same events should be closer in the shared embedding space than the distance between s and $\mathbf{v}'$ with unrelated events. Based on this intuition, a text-video contrastive loss function is defined, which leverages videos to enhance text representation by matching texts and videos conditioned on their event content. Considering that event triggers of a specific event type may be diverse, it is not reasonable to represent events with their triggers. For example, *parade* and *march* are two triggers of the *Demonstrate* event type, however, the semantic and video descriptions of these two triggers are the same. Therefore, we use the event type to present a specific event.

Specifically, we set samples whose event type sets are different from the anchor sample as negative samples, and others are positive. In this way, vectors of text-video pairs with the same events are pulled together, and pairs with different events are pushed apart:

$$\mathcal{L}_T(\mathbf{s},\mathbf{v}) = \mathbb{E}_{\mathbf{s}'}[\mu_T(k,l)S(\mathbf{s}',\mathbf{v}) - S(\mathbf{s},\mathbf{v}) + \epsilon]_+ \\ + \mathbb{E}_{\mathbf{v}'}[\mu_T(i,j)S(\mathbf{s},\mathbf{v}') - S(\mathbf{s},\mathbf{v}) + \epsilon]_+$$

where i, j, k, l are the indexes of samples with $\mathbf{s}$, $\mathbf{v}'$, $\mathbf{s}'$ and $\mathbf{v}$ respectively. $S(\cdot,\cdot)$ is the distance function and $\mu(\cdot,\cdot)$ is the negative sample weighting function which will be introduced in detail in the following content.

Argument extraction relies on representations of both texts and events, where text is refined by $\mathcal{L}_T$. Similar to contrastive text learning, representations of an event and the video depicting it are tend to be closer than the videos do not contain this event. We employ contrastive event learning by matching event-video pairs to enhance event representations in this work. Specifically, for a specific event type e, we push apart its representation from the unmatched video representation $\mathbf{v}^{'}$ and bridge the distance with the matched video $\mathbf{v}$. The match judgement principle is defined as: a video v and an event type e are matched if e is in the event type set of v, and meanwhile, the significance weight w_e of e in this video should be larger than η, otherwise they are mis-matched. The conrastive event learning loss is defined as:

$$\mathcal{L}_E(\mathbf{e},\mathbf{v}) = \mathbb{E}_{\mathbf{e}'}[\mu_E(\mathbf{e}^{'},i)S(\mathbf{e}^{'},\mathbf{v}) - S(\mathbf{e},\mathbf{v}) + \epsilon]_+ \\ + \mathbb{E}_{\mathbf{v}'}[\mu_E(\mathbf{e},j)S(\mathbf{e},\mathbf{v}^{'}) - S(\mathbf{e},\mathbf{v}) + \epsilon]_+$$

where i,j are the indexes of the samples with v and $v^{'}$.

The overall loss of ViECL is defined as:

$$\mathcal{L}_{VCL} = \sum_{(s,v)\in D} \lambda_1 \mathcal{L}_T(s,v) + \sum_{v\in D}\sum_{e\in E_{all}} \lambda_2 \mathcal{L}_E(e,v)$$

where λ_1 and λ_2 are learnable parameters to balance weights of $\mathcal{L}_T$ and $\mathcal{L}_E$ and D is the training set.

Negative Sample Weighting As mentioned above, treating negative samples chosen based on events equally is not reasonable because negative samples have various events with different significance levels. To address this problem, we firstly weigh different event types in a sample: as the significance of events is more intuitive describing in videos than texts, we use videos to measure event significance by passing video features to a linear model with a *Softmax* layer. The weight of significance corresponding to the k^{th} event type e_k in the o^{th} sample is presented as:

$$w_{e_k^o} = \frac{exp(\phi(\mathbf{v}_o)_k)}{\sum_{l=1}^{|E|} exp(\phi(\mathbf{v}_o)_l)}$$

$$\phi(\mathbf{v}_o) = W\mathbf{v}_o + b$$

Then we assign weight scores to the negative sample with index j by measuring the difference between its event type set and the anchor sample with index i. For $\mathcal{L}_T$, the weighting function can be presented as:

$$\mu_T(i,j) = \frac{\sum_{e\in E_i \setminus E_j} w_{e^i} + \sum_{e\in E_j \setminus E_i} w_{e^j}}{\sum_{e\in E_i} w_{e^i} + \sum_{e\in E_j} w_{e^j} + \delta}$$

where δ is used to avoid the denominator to be 0. For $\mathcal{L}_E$, the weighting function is calculated by:

$$\mu_E(e_k^i, j) = \frac{\sum_{e\in E_j - e_k^i} w_{e^j}}{w_{e^i} + \sum_{e\in E_j - e_i} w_{e^j} + \delta}$$

2.4 Training and Inference

During the training phase, parameters of the video encoder are frozen and a linear layer is appended to project video vectors to the shared embedding space. We jointly optimize parameters of Trigger Extractor and Argument Extractor with an Adam optimizer to learn the EvE:

$$\mathcal{L}_{EvE} = \mathcal{L}_t + \mathcal{L}_a$$

The EvE loss and ViECL loss are jointly optimized:

$$\mathcal{L} = \mathcal{L}_{EvE} + \sigma \mathcal{L}_{ViECL}$$

where σ is a hyper-parameter to balance the losses.

In the inference phrase, only the sentences are used to predict the most likely event:

$$e^* = arg\ max\ \mathrm{P}(e|s)$$

where e^* is the event results predicted in a Sequential Labeling manner.

3 TVEE Dataset

3.1 Data Collection

Event schema We borrow the event schema from the benchmark ACE2005 (Walker et al., 2006) as our event schema, which contains 8 event types and 33 subtypes.

Data Source We collect data from the On Demand News[2] channel that contains international news videos with a wide coverage of event types. In addition, news from this channel generally have multiple sentences describing events, which are depicted in the videos at the same time. As a result, a total of 24,129 news videos are collected and further split into frames per second. As textual sentences are binding with pictures, we therefore employ the OCR tool [3] to extract sentences. Then we drop the frames without sentence descriptions and keep the video segments longer than 2 frames. The rest 7,598 instances are kept as our sentence-video pairs.

3.2 Data Annotation

We follow the ACE2005 (Walker et al., 2006) annotation guideline to annotate triggers, event types, entities and argument roles in the sentences with a two-stage iterative annotation method. To speed up the annotation, we adopt the state-of-the-art information extraction model ONEIE (Lin et al., 2020)

[2]https://www.youtube.com/c/ondemandnews

[3]https://cloud.tencent.com/product/ocr-catalog

Item	Statistics	
	Sentence	Video
# Instances	7,598	7,598
# Events	6584	-
# Average Events / Instance	0.87	-
Average Length	17.0	6.7
Max Length	43	7
Min Length	12	4

Table 1: Statistics of TVEE. Lengths corresponding to texts and videos are token and time second, respectively. “-" means absent.

to obtain pseudo event annotations from raw sentences. In the first annotation stage, we employ ten expert annotators to correct the pseudo labels and supplement event annotations missed by the ONEIE model. To guarantee the annotation quality, three experienced annotators are invited to double check the annotations. Then we employed another two annotators to evaluate 100 sentences sampled from the dataset at random to calculate the Inter-Annotator Agreement(IAA), which is 83.4% . The statistics of TVEE is listed in Table 1.

4 Experimental Settings

4.1 Dataset

We conduct experiments on the TVEE and VM2E2 datasets. The TVEE is split into training, development and testing sets with a ratio of 8:1:1. VM2E2 is a text-video multimodal event extraction dataset, where most sentences and videos are paired without strict event alignments. VM2E2 contains 13,239 sentences and 860 videos. We follow the splitting setting from Chen et al. (2021a) to divide the data into training and testing sets.

4.2 Evaluation

We evaluate the model with *Precision(P)*, *Recall(R)* and *F-score(F1)* for event extraction, where a trigger prediction is considered correctly extracted on the condition of the offset and event type are same with the corresponding golden triggers; an argument is considered correctly extracted when the offset, argument role and event type are same with the corresponding golden arguments (Li et al., 2013).

4.3 Baselines

For event extraction, we adopt two SOTA models as our baselines: (1) We compare against the EEQA (Du and Cardie, 2020) model that performs SOTA on event extraction with the setting without considering external entity information. Because EEQA can not leverage videos as input, we trained EEQA only on the text data of both TVEE and VM2E2. (2) We compare the SOTA model of text-video event extraction JMMT (Chen et al., 2021a) on both TVEE and VM2E2. In particular, we use JMMT to extract events only from text data for fair comparison with our model.

4.4 Implementation Details

For texts, we use *bert base* model[4] to produce contextualized representations, which are further processed with mean pooling to calculate the sentence representation. For videos, we adopt the ResNexT-101 16 frames (Hara et al., 2018) model pre-trained on Kinetics (Carreira and Zisserman, 2017) to calculate the video representation with the same mean pooling strategy. In our experiments, we set the parameters λ_1, λ_2 to be 1.0 and σ as 1000.

4.5 Main Results

Table 2 presents the overall results of our model in comparison with related work on TVEE and VM2E2 test sets. Our model outperforms related work in extracting both triggers and arguments in terms of F1, thus achieving the best results for event extraction. Compared with EEQA, our model gains consistent improvements in terms of precision, recall and F1, indicating the effectiveness of the model for extracting events. In addition, the comparison with JMMT over F1 indicates the effectiveness of for improving event extraction.

4.6 Ablation Study

To verify the contribution of the contrastive module, we conduct ablation studies with the following six settings: (1) Text-only setting that trained without videos using BERT+CRF structure; (2) plain contrastive learning (PCL) contrasts representation learning by pairing the anchor sentence with the corresponding video as positive sample while the rest videos as the negative samples; (3) text contrastive learning (TCL) that contrasts learning by appending the contrastive text learning loss $mathcalL_T$; (4) event contrastive learning (ECL) that contrasts learning by appending the contrastive event learning loss $mathcalL_E$; (5) text and event contrastive learning (TECL) that trained with both contrastive text and event learning losses;

[4] https://huggingface.co/bert-base-uncased

MODEL	TVEE						VM2E2					
	Trigger			Argument			Trigger			Argument		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
EEQA	79.8	85.7	82.6	72.8	59.2	65.3	43.4	37.8	40.3	19.2	15.4	17.0
JMMT	**81.7**	83.6	82.6	**82.0**	64.9	72.5	39.7	**56.3**	46.6	17.9	**24.3**	20.6
Ours	81.0	**86.3**	**83.6**	79.7	**69.8**	**74.4**	**49.3**	46.4	**47.8**	**23.1**	20.7	**21.8**

Table 2: The results of our model on test sets in comparison with related work. Best results are highlighted in bold.

MODEL	Trigger			Argument		
	P	R	F1	P	R	F1
Text-only	78.3	83.9	81.0	77.8	62.7	69.4
PCL	**82.1**	80.7	81.4	**81.1**	61.7	70.1
TCL	81.6	83.0	82.3	77.3	65.9	71.1
ECL	80.9	83.0	81.9	80.8	67.3	73.4
TECL	81.7	83.6	82.6	79.6	67.9	73.3
WTECL	81.0	**86.3**	**83.6**	79.7	**69.8**	**74.4**

Table 3: The results of ablation studies on the TVEE test set.

Event Type	Text-only	WTECL	Event Type	Text-only	WTECL
Business	**4.0**	0.0	Justice	**83.5**	82.3
Conflict	77.7	**82.9**	Contact	**88.1**	86.1
Personnel	64.5	**68.5**	Transaction	16.6	**28.6**
Life	94.4	**95.1**	Movement	76.0	**79.5**

Table 4: F1 scores of trigger extraction on different event types with Text-only and WTECL settings.

(6) weighted text and event contrastive learning (WTECL) that introduces weights of negative samples with contrastive text and event losses.

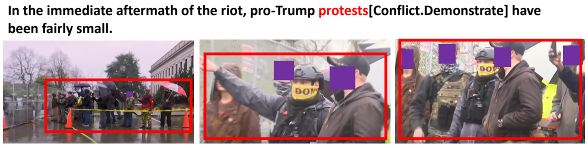
Effects of event information on contrastive learning Results of the settings with contrastive learning outperform the Text-only setting, demonstarting that learning event extraction by contrasting text and videos has better performance than extracting events that only consider the text modality. In comparison with the PCL setting, the introduce of event information based contrastive learning helps the model to extract events on both triggers and arguments. Benefit on the event information, the TCL setting, which is learning text representation by contrasting event information obtains improvement on Trigger extraction and argument extraction in terms of F1 than PCL. Compared with TCL and PCL, the ECL improves much on argument extraction performance, which shows the effectiveness of learning event types by contrasting with videos for further argument extraction. Results of TECL compared with TCL and ECL shows that the combination of contrastive text and event learning can benefit both trigger extraction and argument extraction than only considering one learning object. When introducing the negative sample weighting function, the WTECL model increases the performances on F1 scores of trigger and argument extraction compared with TECL, which shows the necessity of weighting negative samples and measuring various event weights.

Effects of ViECL on different event types. We compare the performance of Text-only setting and WTECL setting on the 8 event types which is shown in Table 4. The F-scores are improved with WTECL on five event types, where *Transaction* and *Conflict* events obtain the most improvement and *Business* declines the most. By observing videos of these event types, it turns out that that it is easier to judge events from the videos corresponding the improved event types than the declined ones. We list two examples from TVEE in Figure 3, the crowd gathered in (a) is the main content in the video, which indicates a *Conflict.Demonstrate* event, however, in (b) the *Business.Start-Org* event can only be identified by the red rope from the third frame. Therefore, we can conclude that the performance of video enhancement is based on the intuition level of event contents: the more intuitional, the better it performs.

5 Related Work

5.1 Event Extraction

Most event extraction research focuses on the sentence level. Early efforts on event extraction mainly used common CNN, RNN and their variants (Chen et al., 2015; Nguyen and Grishman, 2015; Nguyen et al., 2016) to tackle the extraction of triggers and arguments. With the success of pretrained language models (PLMs), research has employed transformers-based models such as BERT to improve the task Yang et al. (2019); Wadden et al. (2019); Kenton and Toutanova (2019). To learn better representation, Wang et al. (2021) leverage contrastive learning to pre-train on the Automatic Speech Recognition (AMR) of massive unsupervised data. To utilize knowledge from other modalities, some studies introduce multimedia data to



(a)

(b)

Figure 3: Two examples from the TVEE with event types *Conflict.Demonstrate* and *Business.Start-Org* with triggers marked in red color and arguments are underlined. The main objects trigger the corresponding events are labeled by red and green boxes. We mask faces with purple boxes for privacy.

learn multi-modal event extraction. Zhang et al. (2017) demonstrates the effectiveness of extracting events with visually based entity data. Tong et al. (2020) proposes a dual recurrent multimodal model to improve text event detection with external news images. Li et al. (2020) extract events from both text and image data jointly by projecting them into a common embedding space in a unsupervised way. Most similar work to ours is Chen et al. (2021a), it propose a Transformer based model to jointly extract events from text and video data. Chen et al. (2021a) leverage a pretrained video-text retrieval model to match the most relevant text video clip pairs as the coreferential sentence and video segment. Our work are different from Chen et al. (2021a) in many aspects. Firstly, we also target text and video pairs data but they are describing the same events content originally, so it doesn't depend on the capacity of retrieval model. Furthermore, we argue that the supplementary arguments in videos are negligible, so we soly focus on extracting events from texts and videos are used to enhance learning in contrastive way.

5.2 Contrastive Learning

Contrastive learning methods have shown the effectiveness in representation learning via pulling together positive samples with anchor samples and push apart negative samples in the representation space (Oord et al., 2018; Chen et al., 2020; He et al., 2020). Many specific tasks in NLP domain also have impressive performance based on contrastive learning such as question answering (Yeh and Chen, 2019) and information extraction (Peng et al., 2020; Wang et al., 2021).

Constrastive learning also has been demonstrated to perform greatly in multimodal domain tasks. Zhang et al. (2021a) introduced a contrastive learning based modal not only learn inter-modal similarities but also take intra-modal representation into account. Zhang et al. (2021b) propose a video-text match model exploiting rich information in videos to learn better textual constituents representation for unsupervised grammar induction. However, Zhang et al. (2021b) only focus on leveraging videos to learn text representations. Meanwhile, they treat every negative sample equally that don't take the difference of negative samples into account. Different from their work, in this paper, we construct negative samples and weigh them by measuring the difference between their event types. Moreover, event representations are also learnt by contrasting videos to improve argument extraction.

6 Conclusion and Future Work

In this work, we introduce the video modality to assist event extraction by considering their events information. We introduce a new dataset called TVEE which is consists of pairs of sentence and video which are describing the same events and is annotated with event labels in sentences. We publicly release the dataset to stimulate further research on multimodal event extraction and other tasks. Meanwhile, We proposed a contrastive learning based model composed of two contrastive losses and a negative sample weighting function. Experiments on two multimodal event extraction datasets shows that our model can improve event extraction and outperforms the baselines on this task. Our current did not consider other modalities such as the auIn the future, we will consider more modalities such as audio to enhance event extraction.

References

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308.

Brian Chen, Xudong Lin, Christopher Thomas, Manling Li, Shoya Yoshida, Lovish Chum, Heng Ji, and Shih-Fu Chang. 2021a. Joint multimedia event extraction from video and article. In *Proceedings of* the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 74–88.

Hong Chen, Raphael Shu, Hiroya Takamura, and Hideki Nakayama. 2021b. Graphplan: Story generation by planning with event graph.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning (ICML)*, pages 1597–1607. PMLR.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 167–176.

Divyanshu Daiya. 2020. Combining temporal event relations and pre-trained language models for text summarization. In *IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 641–646.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683.

Xinya Du, Alexander M Rush, and Claire Cardie. 2021. Grit: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of Meeting of the Association for Computational Linguistics (ACL)*, pages 634–644.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 6546–6555.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738.

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

Manling Li, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen McKeown. 2021. Timeline summarization based on event graph compression via time-aware optimal transport. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6443–6456, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 73–82.

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

Sebastian Martschat and Katja Markert. 2018. A temporally sensitive submodularity framework for timeline summarization. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 230–240, Brussels, Belgium. Association for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of Meeting of the Association for Computational Linguistics (ACL)*, pages 300–309.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 365–371.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from context or names? an empirical study on neural relation extraction. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672.

Meihan Tong, Shuai Wang, Yixin Cao, Bin Xu, Juanzi Li, Lei Hou, and Tat-Seng Chua. 2020. Image enhanced event detection in news articles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9040–9047.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou. 2021. CLEVE: Contrastive Pre-training for Event Extraction. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, volume 1, pages 6283–6297, Online. Association for Computational Linguistics.

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

Yi-Ting Yeh and Yun-Nung Chen. 2019. Qainfomax: Learning robust question answering system by mutual information maximization. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3370–3375.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021a. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–842.

Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, and Jiebo Luo. 2021b. Video-aided unsupervised grammar induction. In *Proceedings of Meeting of the Association for Computational Linguistics (ACL)*, pages 1513–1524.

Tongtao Zhang, Spencer Whitehead, Hanwang Zhang, Hongzhi Li, Joseph Ellis, Lifu Huang, Wei Liu, Heng Ji, and Shih-Fu Chang. 2017. Improving event extraction via multimodal integration. In *Proceedings of ACM International Conference on Multimedia (MM)*, MM '17, page 270–278, New York, NY, USA. Association for Computing Machinery.

Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. 2021. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1450–1459.