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 Contrastive Learning based Event Extraction model with enhancements from the Video 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\(^1\).

\(^1\)The dataset and code will be released based on acceptance.

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). Contrastive learning methods (Zolfaghari et al., 2021; ?;
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 Contrastive Learning based Event Extraction model enhanced by Video 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 using the trigger extractor and argument extractor, respectively.

2 Proposed Model

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 $s \in \mathbb{R}^{d \times d}$, 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:

$$
L_t = - \sum_{i=1}^{K} \sum_{j=1}^{n} \log P(y_j | s_i)
$$

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

$$
L_a = - \sum_{i=1}^{K} \sum_{j=1}^{n} \log P(y_j | s_i; r^i; e^i)
$$

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.

2.3 Model Details

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.
2.3 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 \( v \in \mathbb{R}^{d} \) using a mean pooling layer.

**Contrastive losses** Intuitively, the distance of representations between \( s \) and video \( v \) describing the same events should be closer in the shared embedding space than the distance between \( s \) and \( v' \) 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:

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

where \( i, j, k, l \) are the indexes of samples with \( s, v', s', v \) 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 \( L_T \). 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 $\mathbf{v}$ and an event type $e$ are matched if $e$ is in the event type set of $\mathbf{v}$, and meanwhile, the significance weight $w_e$ of $e$ in this video should be larger than $\eta$, otherwise they are mis-matched. The contrastive event learning loss is defined as:

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

$$
+ \mathbb{E}_e'[\mu_E(\mathbf{e}, j) S(\mathbf{e}, \mathbf{v}') - S(\mathbf{e}, \mathbf{v}) + \epsilon]_+
$$

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

The overall loss of ViECL is defined as:

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

where $\lambda_1$ and $\lambda_2$ 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:

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

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

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:

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

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

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

### 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)
We evaluate the model with Ann-\text{to-Video} (Chen et al., 2021a) to extract events from text data for fair comparison with our model.

4.4 Implementation Details

For texts, we use \textit{bert base} model\footnote{https://huggingface.co/bert-base-uncased} 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 $\lambda_1, \lambda_2$ to be 1.0 and $\sigma$ 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 $\mathcal{L}_T$; (4) event contrastive learning (ECL) that contrasts learning by appending the contrastive event learning loss $\mathcal{L}_E$; (5) text and event contrastive learning (TECL) that trained with both contrastive text and event learning losses;

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

<table>
<thead>
<tr>
<th>Item</th>
<th>Statistics</th>
<th>Sentence</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td># Instances</td>
<td>7,598</td>
<td>7,598</td>
<td></td>
</tr>
<tr>
<td># Events</td>
<td>6,584</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td># Average Events / Instance</td>
<td>0.87</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Average Length</td>
<td>17.0</td>
<td>6.7</td>
<td></td>
</tr>
<tr>
<td>Max Length</td>
<td>43</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Min Length</td>
<td>12</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>MODEL</th>
<th>Trigger</th>
<th>Argument</th>
<th>Trigger</th>
<th>Argument</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>EEQA</td>
<td>79.8</td>
<td>85.7</td>
<td>82.6</td>
<td>72.8</td>
</tr>
<tr>
<td>JMMT</td>
<td>81.7</td>
<td>83.6</td>
<td>82.6</td>
<td>82.0</td>
</tr>
<tr>
<td>Ours</td>
<td>81.0</td>
<td>86.3</td>
<td>83.6</td>
<td>79.7</td>
</tr>
<tr>
<td></td>
<td>23.1</td>
<td>20.7</td>
<td>21.8</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Text-only</th>
<th>WTECL</th>
<th>Event Type</th>
<th>Text-only</th>
<th>WTECL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>82.1</td>
<td>88.5</td>
<td>Conflict</td>
<td>83.5</td>
<td>82.3</td>
</tr>
<tr>
<td>Conflict</td>
<td>82.9</td>
<td>Contact</td>
<td>88.1</td>
<td>86.1</td>
<td></td>
</tr>
<tr>
<td>Personnel</td>
<td>64.5</td>
<td>68.5</td>
<td>Transaction</td>
<td>16.6</td>
<td>28.6</td>
</tr>
<tr>
<td>Life</td>
<td>94.4</td>
<td>95.1</td>
<td>Movement</td>
<td>76.0</td>
<td>79.5</td>
</tr>
</tbody>
</table>

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, demonstrating 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 introduction 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 show 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 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 intuitive, 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
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 an 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 only 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).

Contrastive 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 aumni the future, we will consider more modalities such as audio to enhance event extraction.

References


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


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.


