Cross-media Structured Common Space for Multimedia Event Extraction

Manling Li[1]*, **Alireza Zareian**[2]*, **Qi Zeng**[1], **Spencer Whitehead**[1], **Di Lu**[3], **Heng Ji**[1], **Shih-Fu Chang**[2]
[1]University of Illinois at Urbana-Champaign
[2]Columbia University
[3]DataMiner
hengji@illinois.edu, sfchang@ee.columbia.edu

Abstract

We introduce a new task, **M**ulti**M**edia **E**vent **E**xtraction (M^{2}E^2), which aims to extract events and their arguments from multimedia documents. We develop the first benchmark and collect a dataset of 245 multimedia news articles with extensively annotated events and arguments.[1] We propose a novel method, **W**eakly **A**ligned **S**tructured **E**mbedding (**WASE**), that encodes structured representations of semantic information from textual and visual data into a common embedding space. The structures are aligned across modalities by employing a weakly supervised training strategy, which enables exploiting available resources without explicit cross-media annotation. Compared to uni-modal state-of-the-art (SotA) methods, our approach achieves 4.0% and 9.8% absolute F-score gains on text event argument role labeling and visual event extraction. Compared to SotA multimedia unstructured representations, we achieve 8.3% and 5.0% absolute F-score gains on multimedia event extraction and argument role labeling, respectively. We also extract 21.4% more events from images than traditional text-only event extraction methods.

1 Introduction

Traditional event extraction methods target a single modality, such as text (Wadden et al., 2019), images (Yatskar et al., 2016) or videos (Ye et al., 2015; Caba Heilbron et al., 2015; Soomro et al., 2012). However, the practice of contemporary journalism (Stephens, 1998) distributes news via multimedia. By randomly sampling 100 multimedia news articles from the Voice of America (VOA), we find that 33% of images in the articles contain visual objects that serve as event arguments that are not mentioned in the text. Take

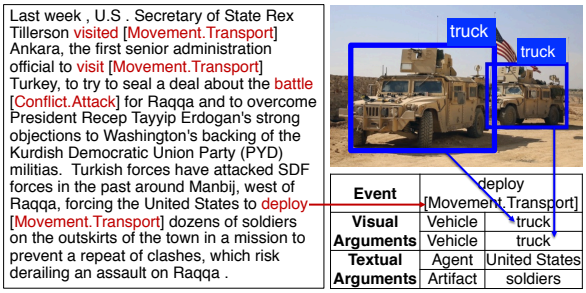


Figure 1: An example of Multimedia Event Extraction. An event mention and some event arguments (*Agent* and *Person*) are extracted from text, while the vehicle arguments can only be extracted from the image.

Figure 1 as an example, we can extract the *Agent* and *Person* arguments of the *Movement.Transport* event from text, but can extract the *Vehicle* argument only from the image. Nevertheless, event extraction is independently studied in Computer Vision (CV) and Natural Language Processing (NLP), with major differences in task definition, data domain, methodology, and terminology. Motivated by the complementary and holistic nature of multimedia data, we propose **M**ulti**M**edia **E**vent **E**xtraction (**M^{2}E^2**), a new task that aims to jointly extract events and arguments from multiple modalities. We construct the first benchmark and evaluation dataset for this task, which consists of 245 fully annotated news articles.

We propose the first method, **W**eakly **A**ligned **S**tructured **E**mbedding (**WASE**), for extracting events and arguments from multiple modalities. Complex event structures have not been covered by existing multi-modal representation methods (Wu et al., 2019b; Faghri et al., 2017; Karpathy and Fei-Fei, 2015), so we propose to learn a *structured* multi-modal embedding space. More specifically, given a multimedia document, we represent each image or sentence as a graph, where each node represents an event or entity and each

*These authors contributed equally to this work.
[1]All code, data and resources will be made publicly available for research purposes.

edge represents an argument role. The node and edge embeddings are represented in a multimedia common semantic space, as they are trained to resolve event co-reference across modalities and to match images with relevant sentences. This enables us to jointly classify events and argument roles from both modalities. A major challenge is the lack of multi-modal event argument annotations, which are costly to obtain due to the annotation complexity. Therefore, we propose a weakly supervised framework, which takes advantage of annotated uni-modal corpora to separately learn visual and textual event extraction, and uses an image-caption dataset to align the modalities.

We evaluate WASE on the new task of M^{2}E^2. Compared to the state-of-the-art uni-modal methods and multimedia flat representations, our method significantly outperforms on both event extraction and argument role labeling in all settings. Moreover, it extracts 21.4% more events than text-only baselines. In summary, this paper makes the following contributions:

- We propose a new task, MultiMedia Event Extraction, and construct the first annotated news dataset as a benchmark to support deep analysis of cross-media events.
- We develop a weakly supervised training framework, which utilizes existing single-modal annotated corpora, and enables joint inference without cross-modal annotation.
- Our proposed method, WASE, is the first to leverage structured representations and graph-based neural networks for multimedia common space embedding.

2 Task Definition

2.1 Problem Formulation

Each input document consists of a set of images $\mathcal{M} = \{m_1, m_2, \dots\}$ and a set of sentences $\mathcal{S} = \{s_1, s_2, \dots\}$. Each sentence s can be represented as a sequence of tokens $s = (w_1, w_2, \dots)$, where w_i is a token from the document vocabulary $\mathcal{W}$. The input also includes a set of entities $\mathcal{T} = \{t_1, t_2, \dots\}$ extracted from the document text. The objective of M^{2}E^{2}is twofold:

Event Extraction: Given a multimedia document, extract a set of events, where each event e has a type y_e and is grounded on a text trigger word w or an image m or both, i.e.,

$$e = (y_e, \{w, m\}).$$

Note that w and m can both exist, which means the visual event and the textual event are the same. For example in Figure 1, *deploy* indicates the same *Movement.Transport* event as the image. We consider the event e as **text-only** event if only w exists, and as **image-only** event if it only contains m, and as **multimedia** event if it has both w and m.

Argument Extraction: The second task is to extract a set of arguments of event e. Each argument a has an argument role type y_a, and is grounded on an entity t or an image object o (represented as a bounding box), or both,

$$a = (y_a, \{t, o\}).$$

We merge the arguments of visual and textual events if they are the same, as shown in Figure 1.

2.2 The M^{2}E^2 Dataset

We choose eight newsworthy and visually detectable event types from the widely used ACE[2] event ontology (Walker et al., 2006), and expand the role set by adding visual arguments such as *instrument*, as shown in Table 1.

Event Type	Argument Role
Movement.Transport (107|53)	Agent (20|64), Artifact (83|103), Vehicle (11|51), Destination (54|0), Origin (42|0)
Conflict.Attack (143|27)	Attacker (77|12), Target (84|19), Instrument (20|15), Place (61|0)
Conflict.Demonstrate (76|69)	Entity (52|184), **Police** (3|26), **Instrument** (0|118), Place (48|25)
Justice.ArrestJail (103|56)	Agent (42|119), Person (96|99), **Instrument** (0|11), Place (24|0)
Contact.PhoneWrite (33|37)	Entity (33|46), **Instrument** (0|43), Place (8|0)
Contact.Meet (57|79)	Participant (55|321), Place (35|0)
Life.Die (177|64)	Agent (39|0), Instrument (4|2), Victim (165|155), Place (54|0)
Transaction. TransferMoney (10|6)	Giver (8|3), Recipient (8|5), Money (4|8)

Table 1: Event types and argument roles in M^{2}E^2, with expanded ones in bold. Numbers in parentheses represent the counts of textual and visual events/arguments.

We collect 108,693 multimedia news articles from the Voice of America (VOA) website [3] 2006-2017. We select 245 documents as the annotation

[2] https://catalog.ldc.upenn.edu/ldc2006T06
[3] https://www.voanews.com/

set based on three criteria: (1) Informativeness: articles with more event mentions; (2) Illustration: articles with more images (> 4); (3) Diversity: articles that balance the event type distribution regardless of true frequency. The data statistics are shown in Table 2. Among all of these events, 95 textual events and 91 visual events can be aligned as 98 cross-media event pairs. The dataset can be divided into 611 text-only events, 300 image-only events, and 98 multimedia events.

Source		Event		Argument Role	
sentence	image	textual	visual	textual	visual
6,167	1,014	706	391	1,124	1,407

Table 2: M^2E^2 data statistics.

We follow the ACE event annotation guidelines (Walker et al., 2006) for textual event and argument annotation. For images, we design an annotation guideline as detailed in the Appendix. To localize arguments, we annotate *union bounding box* covering all constituents (e.g., a crowd); and *instance bounding box* which is the smallest region that covers an individual participant (e.g., one person in the crowd).

3 Method

3.1 Approach Overview

As shown in Figure 2, the training phase contains three tasks: text event extraction (Section 3.2), visual situation recognition (Section 3.3), and cross-media alignment (Section 3.4). We learn a cross-media shared encoder, a shared event classifier, and a shared argument classifier. In the testing phase (Section 3.5), given a multimedia news article, we encode the sentences and images into the structured common space, and jointly extract textual and visual events and arguments, followed by cross-modal coreference resolution.

3.2 Text Event Extraction

Text Structured Representation: As shown in Figure 3, we choose Abstract Meaning Representation (AMR) (Banarescu et al., 2013) to represent text because it includes a rich set of 150 fine-grained semantic roles. To encode each text sentence, we run the CAMR parser (Wang et al., 2015b,a, 2016) to generate an AMR graph, based on the named entity recognition and part-of-speech (POS) tagging results from Stanford CoreNLP (Manning et al., 2014). To represent each word w in a sentence s, we concatenate its pre-trained GloVe word embedding (Pennington et al., 2014), POS embedding, entity type embedding and position embedding. We then input the word sequence to a bi-directional long short term memory (Bi-LSTM) (Graves et al., 2013) network to encode the word order and get the representation of each word $\boldsymbol{w}$. Given the AMR graph, we apply a Graph Convolutional Network (GCN) (Kipf and Welling, 2016) to encode the graph contextual information following (Liu et al., 2018a):

$$\boldsymbol{w}_i^{(k+1)} = f(\sum_{j \in \mathcal{N}(i)} g_{ij}^{(k)}(\boldsymbol{W}_{E(i,j)}\boldsymbol{w}_j^{(k)} + \boldsymbol{b}_{E(i,j)}^{(k)})), \quad (1)$$

where $\mathcal{N}(i)$ is the neighbour nodes of w_i in the AMR graph, $E(i,j)$ is the edge type between w_i and w_j, g_{ij} is the gate following (Liu et al., 2018a), k represents GCN layer number, and f is the Sigmoid function. $\boldsymbol{W}$ and $\boldsymbol{b}$ denote parameters of neural layers in this paper. We take the hidden states of the last GCN layer for each word as the common-space representation $\boldsymbol{w}^{\mathbb{C}}$, where $\mathbb{C}$ stands for the common (multi-modal) embedding space. For each entity t, we obtain its representation $\boldsymbol{t}^{\mathbb{C}}$ by averaging the embeddings of its tokens.

Event and Argument Classifier: We classify each word w into event types y_e[4] and classify each entity t into argument role y_a:

$$P(y_e|w) = \frac{\exp\left(\boldsymbol{W}_e\boldsymbol{w}^{\mathbb{C}} + \boldsymbol{b}_e\right)}{\sum_{e'}\exp\left(\boldsymbol{W}_{e'}\boldsymbol{w}^{\mathbb{C}} + \boldsymbol{b}_{e'}\right)},$$
$$P(y_a|t) = \frac{\exp(\boldsymbol{W}_a[\boldsymbol{t}^{\mathbb{C}}; \boldsymbol{w}^{\mathbb{C}}] + \boldsymbol{b}_a)}{\sum_{a'}\exp(\boldsymbol{W}_{a'}[\boldsymbol{t}^{\mathbb{C}}; \boldsymbol{w}^{\mathbb{C}}] + \boldsymbol{b}_{a'})}. \quad (2)$$

3.3 Image Event Extraction

Image Structured Representation: To obtain image structures similar to AMR graphs, and inspired by *situation recognition* (Yatskar et al., 2016), we represent each image with a *situation graph*, that is a star-shaped graph as shown in Figure 3, where the central node is labeled as a verb v (e.g., *destroying*), and the neighbor nodes are arguments labeled as $\{(n, r)\}$, where n is a noun (e.g., *ship*) derived from WordNet synsets (Miller, 1995) to indicate the entity type, and r indicates the role (e.g., *item*) played by the entity in the event, based on FrameNet (Fillmore et al., 2003).

[4] We use BIO tag schema to decide trigger word boundary, i.e., adding prefix *B-* to the type label to mark the beginning of a trigger, *I-* for inside, and *O* for none.

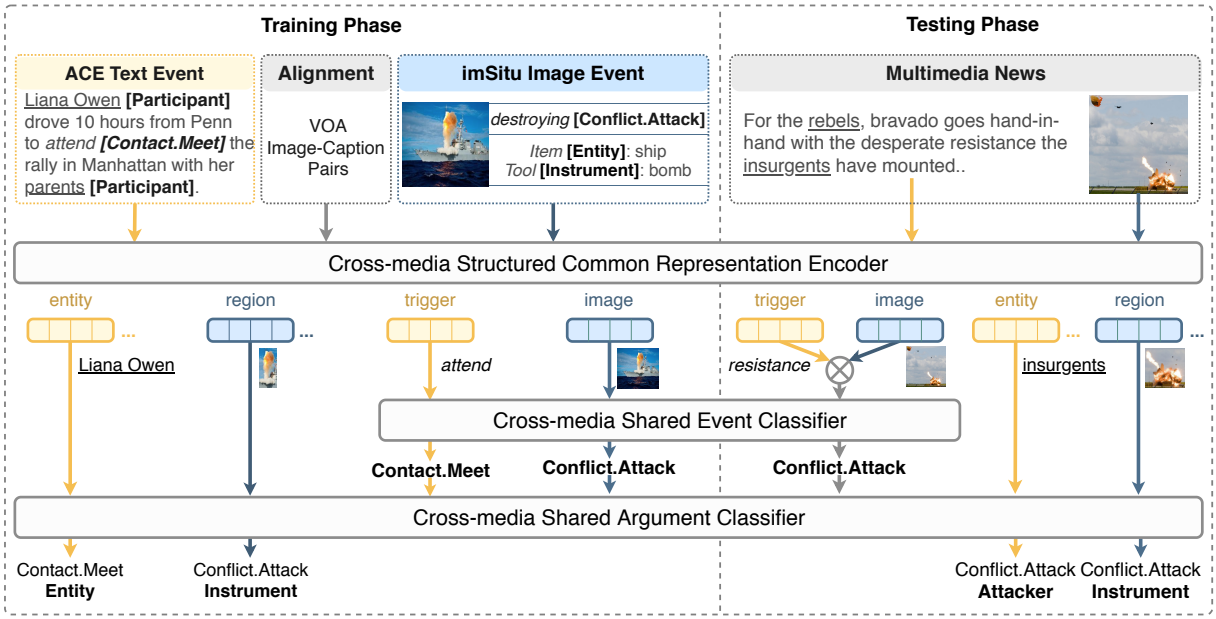


Figure 2: Approach overview. During training (left), we jointly train three tasks to establish a cross-media structured embedding space. During test (right), we jointly extract events and arguments from multimedia articles.

We develop two methods to construct situation graphs from images and train them using the imSitu dataset (Yatskar et al., 2016) as follows.

(1) Object-based Graph: Similar to extracting entities to get candidate arguments, we employ the most similar task in CV, object detection, and obtain the object bounding boxes detected by a Faster R-CNN (Region-based Convolutional Neural Network) (Ren et al., 2015) model trained on Open Images (Kuznetsova et al., 2018) with 600 object classes. We employ a VGG-16 CNN (Simonyan and Zisserman, 2014) to extract visual features of an image $\boldsymbol{m}$ and and another VGG-16 to encode the bounding boxes $\{\boldsymbol{o}_i\}$. Then we apply a Multi-Layer Perceptron (MLP) to predict a verb embedding from $\boldsymbol{m}$ and another MLP to predict a noun embedding for each $\boldsymbol{o}_i$.

$$\hat{\boldsymbol{m}} = \text{MLP}_\text{m}(\boldsymbol{m}) \ , \ \hat{\boldsymbol{o}}_i = \text{MLP}_\text{o}(\boldsymbol{o}_i).$$

We compare the predicted verb embedding to all verbs v in the imSitu taxonomy in order to classify the verb, and similarly compare each predicted noun embedding to all imSitu nouns n which results in probability distributions:

$$P(v|m) = \frac{\exp\left(\hat{\boldsymbol{m}}\boldsymbol{v}\right)}{\sum_{v'}\exp\left(\hat{\boldsymbol{m}}\boldsymbol{v}'\right)},$$

$$P(n|o_i) = \frac{\exp(\hat{\boldsymbol{o}}_i\boldsymbol{n})}{\sum_{n'}\exp(\hat{\boldsymbol{o}}_i\boldsymbol{n}')},$$

where $\boldsymbol{v}$ and $\boldsymbol{n}$ are word embeddings initialized with GloVE (Pennington et al., 2014). We use another MLP with one hidden layer followed by Softmax (σ) to classify role r_i for each object o_i:

$$P(r_i|o_i) = \sigma\big(\text{MLP}_\text{r}(\hat{\boldsymbol{o}}_i)\big).$$

Given verb v^* and role-noun (r_i^*, n_i^*) annotations for an image (from the imSitu corpus), we define the situation loss functions:

$$\mathcal{L}_v = -\log P(v^*|m),$$
$$\mathcal{L}_r = -\log(P(r_i^*|o_i) + P(n_i^*|o_i)).$$

(2) Attention-based Graph: State-of-the-art object detection methods only cover a limited set of object types, such as 600 types defined in Open Images. Many salient objects such as *bomb*, *stone* and *stretcher* are not covered in these ontologies. Hence, we propose an open-vocabulary alternative to the object-based graph construction model. To this end, we construct a role-driven attention graph, where each argument node is derived by a spatially distributed attention (heatmap) conditioned on a role r. More specifically, we use a VGG-16 CNN to extract a 7×7 convolutional feature map for each image m, which can be regarded as attention *keys* $\boldsymbol{k}_i$ for 7×7 local regions. Next, for each role r defined in the situation recognition ontology (e.g., *agent*), we build an attention *query* vector $\boldsymbol{q}_r$ by concatenating role embedding $\boldsymbol{r}$ with the image feature $\boldsymbol{m}$ as context and apply a fully connected layer:

$$\boldsymbol{q}_r = \boldsymbol{W}_q[\boldsymbol{r}; \boldsymbol{m}] + \boldsymbol{b}_q.$$

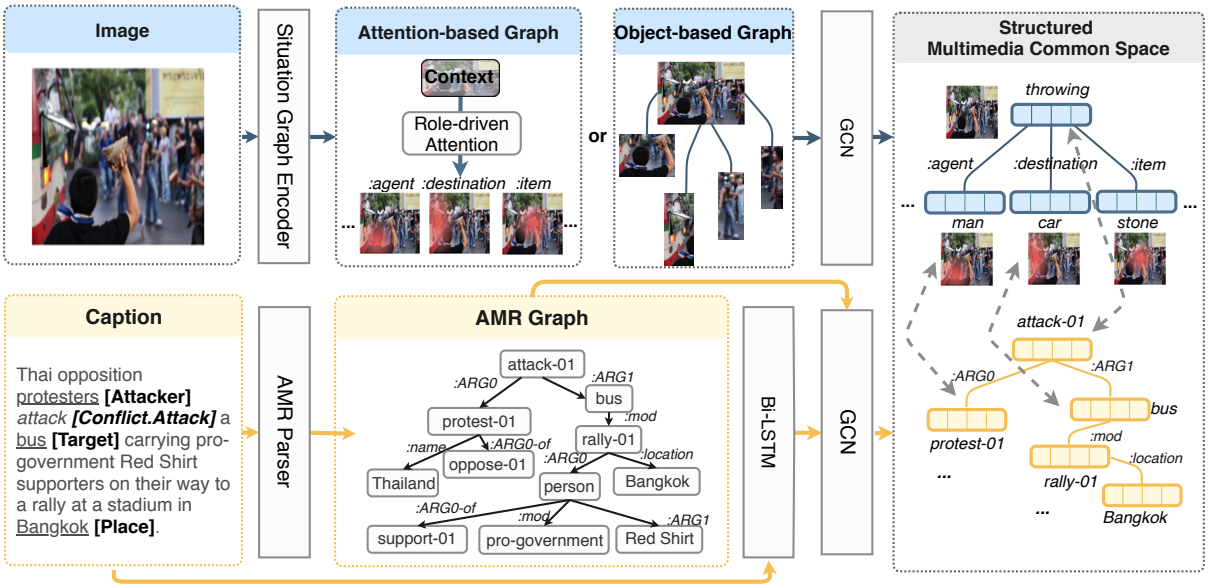


Figure 3: Multimedia structured common space construction. Red pixels stands for attention heatmap.

Then, we compute the dot product of each query with all keys, followed by Softmax, which forms a heatmap $\boldsymbol{h}$ on the image, i.e.,

$$h_i = \frac{\exp(\boldsymbol{q}_r \boldsymbol{k}_i)}{\sum_{j \in 7\times 7} \exp(\boldsymbol{q}_r \boldsymbol{k}_j)}.$$

We use the heatmap to obtain a weighted average of the feature map to represent the argument o_r of each role r in the visual space:

$$\boldsymbol{o}_r = \sum_i h_i \boldsymbol{m}_i.$$

Similar to the object-based model, we embed $\boldsymbol{o}_i$ to $\hat{\boldsymbol{o}}_i$, compare it to the imSitu noun embeddings to define a distribution, and define a classification loss function. The verb embedding $\hat{\boldsymbol{m}}$ and the verb prediction probability $P(v|m)$ and loss are defined in the same way as in the object-based method.

Event and Argument Classifier: We use either the object-based or attention-based formulation and pre-train it on the imSitu dataset (Yatskar et al., 2016). Then we apply a GCN to obtain the structured embedding of each node in the common space, similar to Equation 1. This yields $\boldsymbol{m}^{\mathbb{C}}$ and $\boldsymbol{o}_i^{\mathbb{C}}$. We use the same classifiers as defined in Equation 2 to classify each visual event and argument using the common space embedding:

$$P(y_e|m) = \frac{\exp(\boldsymbol{W}_e \boldsymbol{m}^{\mathbb{C}} + \boldsymbol{b}_e)}{\sum_{e'} \exp(\boldsymbol{W}_{e'} \boldsymbol{m}^{\mathbb{C}} + \boldsymbol{b}_{e'})},$$
$$P(y_a|o) = \frac{\exp(\boldsymbol{W}_a [\boldsymbol{o}^{\mathbb{C}}; \boldsymbol{m}^{\mathbb{C}}] + \boldsymbol{b}_a)}{\sum_{a'} \exp(\boldsymbol{W}_{a'} [\boldsymbol{o}^{\mathbb{C}}; \boldsymbol{m}^{\mathbb{C}}] + \boldsymbol{b}_{a'})}. \quad (3)$$

3.4 Cross-Media Joint Training

In order to make the event and argument classifier shared across modalities, the image and text graph should be encoded to the same space. However, it is extremely costly to obtain the parallel text and image event annotation. Hence, we use event and argument annotations in separate modalities (i.e., ACE and imSitu datasets) to train classifiers, and simultaneously use VOA news image and caption pairs to align the two modalities. To this end, we learn to embed the nodes of each image graph close to the nodes of the corresponding caption graph, and far from those in irrelevant caption graphs. Since there is no ground truth alignment between the image nodes and caption nodes, we use image and caption pairs for weakly supervised training, to learn a soft alignment from each words to image objects and vice versa.

$$\alpha_{ij} = \frac{\exp\left(\boldsymbol{w}_i^{\mathbb{C}} \boldsymbol{o}_j^{\mathbb{C}}\right)}{\sum_{j'} \exp\left(\boldsymbol{w}_i^{\mathbb{C}} \boldsymbol{o}_{j'}^{\mathbb{C}}\right)}, \beta_{ji} = \frac{\exp\left(\boldsymbol{w}_i^{\mathbb{C}} \boldsymbol{o}_j^{\mathbb{C}}\right)}{\sum_{i'} \exp\left(\boldsymbol{w}_{i'}^{\mathbb{C}} \boldsymbol{o}_j^{\mathbb{C}}\right)},$$

where w_i indicates the i^{th} word in caption sentence s and o_j represents the j^{th} object of image m. Then, we compute a weighted average of softly aligned nodes for each node in other modality, i.e.,

$$\boldsymbol{w}_i' = \sum_j \alpha_{ij} \boldsymbol{o}_j^{\mathbb{C}}, \ \boldsymbol{o}_j' = \sum_i \beta_{ji} \boldsymbol{w}_i^{\mathbb{C}}. \quad (4)$$

We define the alignment cost of the image-caption pair as the Euclidean distance between each node

to its aligned representation,

$$\langle s, m \rangle = \sum_i ||\boldsymbol{w}_i - \boldsymbol{w}'_i||_2^2 + \sum_j ||\boldsymbol{o}_j - \boldsymbol{o}'_j||_2^2$$

We use a triplet loss to pull relevant image-caption pairs close while pushing irrelevant ones apart:

$$\mathcal{L}_c = \max(0, 1 + \langle s, m \rangle - \langle s, m^- \rangle),$$

where m^- is a randomly sampled negative image that does not match s. Note that in order to learn the alignment between the image and the trigger word, we treat the image as a special object when learning cross-media alignment.

The common space enables the event and argument classifiers to share weights across modalities, and be trained jointly on the ACE and imSitu datasets, by minimizing the following objective functions:

$$\mathcal{L}_e = -\sum_w \log P(y_e|w) - \sum_m \log P(y_e|m),$$
$$\mathcal{L}_a = -\sum_t \log P(y_a|t) - \sum_o \log P(y_a|o),$$

All tasks are jointly optimized:

$$\mathcal{L} = \mathcal{L}_v + \mathcal{L}_r + \mathcal{L}_e + \mathcal{L}_a + \mathcal{L}_c$$

3.5 Cross-Media Joint Inference

In the test phase, our method takes a multimedia document with sentences $S = \{s_1, s_2, \dots\}$ and images $M = \{m_1, m_2, \dots, \}$ as input. We first generate the structured common embedding for each sentence and each image, and then compute pairwise similarities $\langle s, m \rangle$. We pair each sentence s with the closest image m, and aggregate the features of each word of s with the aligned representation from m by weighted averaging:

$$\boldsymbol{w}''_i = (1 - \gamma)\boldsymbol{w}_i + \gamma \boldsymbol{w}'_i, \quad (5)$$

where $\gamma = \exp(-\langle s, m \rangle)$ and $\boldsymbol{w}'_i$ is derived from m using Equation 4. We use $\boldsymbol{w}''_i$ to classify each word into an event type and to classify each entity into a role with multi-modal classifiers in Equation 2. To this end, we define $\boldsymbol{t}''_i$ similar to $\boldsymbol{t}_i$ but using $\boldsymbol{w}''_i$. Similarly, for each image m we find the closest sentence s, compute the aggregated multi-modal features $\boldsymbol{m}''$ and $\boldsymbol{o}''_i$, and feed into the shared classifiers (Equation 3) to predict visual event and argument roles. Finally, we merge the cross-media events of the same event type if the similarity $\langle s, m \rangle$ is higher than a threshold.

4 Experiments

4.1 Evaluation Setting

We conduct text-only evaluation, image-only evaluation, and multimedia evaluation on text-only, image-only, and multimedia events in M^2E^2 dataset in Section 2.2. We adopt the traditional event extraction measures, i.e., *Precision*, *Recall* and F_1. For textual events, we follow (Ji and Grishman, 2008; Li et al., 2013): a textual event is correct if its event type and trigger offsets match a reference trigger; and a textual event argument is correct if its event type, offsets, and role label match a reference argument. Similarly, for visual events: a visual event is correct if its event type and image match a reference visual event; and visual event argument is correct if its event type, localization, and role label match a reference argument. An argument is correctly localized if the Intersection over Union (IoU) of the predicted bounding box with the ground truth bounding box is over 0.5. An multimedia event is correct if its event type, and trigger offsets (or image) match a reference trigger (or image); its textual argument and visual argument evaluation is same as textual and visual events.

The baselines include: (1) **Text-only** models: We use the state-of-the-art model JMEE (Liu et al., 2018a) and GAIL (Zhang et al., 2019) for comparison. We also evaluate the effectiveness of cross media joint training by including a version of our model trained only on ACE, denoted as $\text{WASE}^{\mathbb{T}}$. (2) **Image-only** models: Since we are the first to extract newsworthy events, and the most similar work *situation recognition* can not localize arguments in images, we use our model trained only on image corpus as baselines. Our visual branch has two versions, object-based and attention-based, denoted as $\text{WASE}^{\mathbb{I}}_{\text{obj}}$ and $\text{WASE}^{\mathbb{I}}_{\text{att}}$. (3) **Multimedia** models: To show the effectiveness of structured embedding, we include a baseline by removing the text and image GCNs from our model, which is denoted as Flat. The Flat baseline ignores edges and treats images and sentences as sets of vectors. We also compare to the state-of-the-art cross-media common representation model, Contrastive Visual Semantic Embedding VSE-C (Shi et al., 2018), by training it the same way as WASE.

4.2 Quantitative Performance

As shown in Table 3, our complete methods (WASE_{att} and WASE_{obj}) outperform all baselines

Training	Model	Text-Only Evaluation						Image-Only Evaluation						Multimedia Evaluation					
		Event			Argument			Event			Argument			Event			Argument		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Text	JMEE	42.5	58.2	48.7	22.9	28.3	25.3	-	-	-	-	-	-	42.1	57.9	48.7	21.1	12.6	15.8
	GAIL	43.4	53.5	47.9	23.6	29.2	26.1	-	-	-	-	-	-	44.0	51.6	46.5	22.7	12.8	16.4
	WASE$^{\mathbb{T}}$	42.3	58.4	48.2	21.4	30.1	24.9	-	-	-	-	-	-	41.2	54.2	47.1	20.1	13.0	15.7
Image	WASE$^{\mathbb{I}}_{\text{att}}$	-	-	-	-	-	-	29.7	61.9	40.1	9.1	10.2	9.6	28.3	59.9	38.4	2.9	6.1	3.8
	WASE$^{\mathbb{I}}_{\text{obj}}$	-	-	-	-	-	-	28.6	59.2	38.7	13.3	9.8	11.2	26.1	58.3	36.1	4.7	5.0	4.9
Multimedia	VSE-C	33.5	47.8	39.4	16.6	24.7	19.8	30.3	48.9	26.4	5.6	6.1	5.7	33.3	48.2	39.3	11.1	14.9	12.8
	Flat$_{\text{att}}$	34.2	63.2	44.4	20.1	27.1	23.1	27.1	57.3	36.7	4.3	8.9	5.8	33.9	59.8	42.2	12.9	17.6	14.9
	Flat$_{\text{obj}}$	38.3	57.9	46.1	21.8	26.6	24.0	26.4	55.8	35.8	9.1	6.5	7.6	34.1	56.4	42.5	16.3	15.9	16.1
	WASE$_{\text{att}}$	37.6	66.8	48.1	27.5	33.2	**30.1**	32.3	63.4	42.8	9.7	11.1	10.3	38.2	67.1	49.1	18.6	21.6	**19.9**
	WASE$_{\text{obj}}$	42.8	61.9	**50.6**	23.5	30.3	26.4	43.1	59.2	**49.9**	14.5	10.1	**11.9**	43.0	62.1	**50.8**	19.5	18.9	19.2

Table 3: Event and argument extraction results (%). We compare three categories of baselines in three evaluation settings. The main contribution of the paper is joint training and joint inference on multimedia data (bottom right).

in three evaluation settings in terms of F_1. The comparison with other multimedia models demonstrates the effectiveness of our model architecture and training strategy. The advantage of structured embedding is shown by better performance over flat baseline. Our model outperforms its text-only and image-only variants in all settings, showing the inadequacy of single-modal information for complex news understanding and the effectiveness of knowledge transfer between modalities.

WASE$_{\text{obj}}$ preforms better on visual argument role labeling due to better bounding box localization. For example, WASE$_{\text{obj}}$ achieves 88.8% recall for *Attacker* , while WASE$_{\text{att}}$ only obtains 6.7%. In comparison, WASE$_{\text{att}}$ achieves higher recall, since WASE$_{\text{obj}}$ cannot extract arguments not from predefined object types. The existing object detection types are for daily life scenario, with only 52 out of 600 types in OpenImages (Kuznetsova et al., 2018) covering M^{2}E^2 argument roles.

Model	P (%)	R (%)	F_1 (%)
rule_based	10.1	100	18.2
VSE	31.2	74.5	44.0
Flat$_{\text{att}}$	33.1	73.5	45.6
Flat$_{\text{obj}}$	34.3	76.4	47.3
WASE$_{\text{att}}$	39.5	73.5	51.5
WASE$_{\text{obj}}$	40.1	75.4	52.4

Table 4: Cross-media event coreference performance.

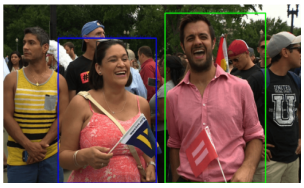
We predict cross-media event alignment by pairing textual and image events in the same document, and calculate *Precision*, *Recall* and F_1 to compare with 98 ground truth event pairs[5]. As shown in Table 4, WASE$_{\text{obj}}$ outperforms all multimedia embedding models, as well as the rule-based baseline using event type matching. It demonstrates the effectiveness of learning object-level cross-media soft alignment.

[5]We do not use coference clustering metrics because we only focus on mention-level cross-media event coreference instead of the full coreference in all documents.

4.3 Qualitative Analysis

Our cross-media joint training approach successfully boosts both event extraction and argument role labeling performance. For example, in Figure 4 (a), the text-only model can not extract *Justice.Arrest* event, but the joint model can use the image as background to detect the event type. In Figure 4 (b), the image-only model detects the image as *Conflict.Demonstration*, but the sentences in the same document help our model not to label it as *Conflict.Demonstration*. Compared with multimedia flat embedding in Figure 5, WASE can learn structures such as *Artifact* is on top of *Vehicle*, and the person in the middle of *Justice.Arrest* is *Entity* instead of *Agent*.

Iraqi security forces *search* ***[Justice.Arrest]*** a civilian in the city of Mosul.



People celebrate Supreme Court ruling on Same Sex Marriage in front of the Supreme Court in Washington.

Figure 4: Image helps textual event extraction, and surrounding sentence helps visual event extraction.

4.4 Remaining Challenges

One of the biggest challenges of M^{2}E^{2}is localizing arguments in images. Object-based models suffer from the limited object types. Attention-based method is not able to precisely localize the

Flat	Event	Movement.Transport	Flat	Event	Justice:ArrestJail
	Role	Artifact = none		Role	Agent = man
Ours	Event	Movement.Transport	Ours	Event	Conflict.Attack
	Role	Artifact = man		Role	Entity = man

Figure 5: Comparison with multimedia flat embedding.

objects for each argument, since there is no supervision on attention extraction during training. For example, in Figure 6, the *Entity* argument in the *Conflict.Demonstrate* event is correctly predicted as *troops*, but its localization is incorrect because *Place* argument share similar attention. When one argument targets at too many instances, attention heatmap tend to lose focus and cover the whole image, as shown in Figure 7.

Figure 6: Argument labeling error examples: correct entity name but wrong localization.

Figure 7: Attention heatmaps lose focus due to large instance candidate number.

5 Related Work

Text Event Extraction Text event extraction has been extensively studied for general news domain (Ji and Grishman, 2008; Liao and Grishman, 2011; Huang and Riloff, 2012; Li et al., 2013; Chen et al., 2015; Nguyen et al., 2016; Hong et al., 2018; Liu et al., 2018b; Chen et al., 2018; Zhang et al., 2019; Liu et al., 2018a; Wang et al., 2019; Yang et al., 2019; Wadden et al., 2019). Multimedia features has been proven to effectively improve text event extraction (Zhang et al., 2017).

Visual Event Extraction "Events" in NLP usually refer to complex events that involve multiple entities in a large span of time (e.g. protest), while in CV (Chang et al., 2016; Zhang et al., 2007; Ma et al., 2017) events are less complex single-entity activities (e.g. washing dishes) or actions (e.g. jumping). Visual event ontologies focus on daily life domains, such as "dogshow" and "wedding ceremony" (Perera et al., 2012). Moreover, most efforts ignore the structure of events including arguments. There are a few methods that aim to localize the agent (Gu et al., 2018; Li et al., 2018; Duarte et al., 2018), or classify the recipient (Sigurdsson et al., 2016; Kato et al., 2018; Wu et al., 2019a) of events, but neither detects the complete set of arguments for an event. The most similar to our work is Situation Recognition (SR) (Yatskar et al., 2016; Mallya and Lazebnik, 2017) which predicts an event and multiple arguments from an input image, but does not localize the arguments. We use SR as an auxiliary task for training our visual branch, but exploit object detection and attention to enable localization of arguments.

Multimedia Representation Multimedia common representation has attracted much attention recently (Toselli et al., 2007; Weegar et al., 2015; Hewitt et al., 2018; Chen et al., 2019; Liu et al., 2019; Su et al., 2019a; Sarafianos et al., 2019; Sun et al., 2019b; Tan and Bansal, 2019; Li et al., 2019a,b; Lu et al., 2019; Sun et al., 2019a; Rahman et al., 2019; Su et al., 2019b). However, previous methods focus on aligning images with their captions, or regions with words and entities, but ignore structure and semantic roles. UniVSE (Wu et al., 2019b) incorporates entity attributes and relations into cross-media alignment, but does not capture graph-level structures of images or text.

6 Conclusions and Future Work

In this paper we propose a new task of multimedia event extraction and setup a new benchmark. We also develop a novel multimedia structured common space construction method to take advantage of the existing image-caption pairs and single-modal annotated data for weakly supervised training. Experiments demonstrate its effectiveness as a new step towards semantic understanding of events in multimedia data. In the future, we aim to extend our framework to extract events from videos, and make it scalable to new event types. We will also apply our extraction results to downstream applications including cross-media event inference, timeline generation, etc.

Acknowledgement

This research is based upon work supported in part by U.S. DARPA AIDA Program No. FA8750-18-2-0014 and KAIROS Program No. FA8750-19-2-1004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.

Xiaojun Chang, Zhigang Ma, Yi Yang, Zhiqiang Zeng, and Alexander G Hauptmann. 2016. Bi-level semantic representation analysis for multimedia event detection. *IEEE transactions on cybernetics*, 47(5):1180–1197.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proc. ACL-IJCNLP2015*.

Yubo Chen, Hang Yang, Kang Liu, Jun Zhao, and Yantao Jia. 2018. Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms. In *Proc. EMNLP2018*.

Kevin Duarte, Yogesh Rawat, and Mubarak Shah. 2018. Videocapsulenet: A simplified network for action detection. In *Advances in Neural Information Processing Systems*, pages 7610–7619.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives.

Charles J Fillmore, Christopher R Johnson, and Miriam RL Petruck. 2003. Background to framenet. *International journal of lexicography*, 16(3):235–250.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056.

John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. 2018. Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2566–2576.

Yu Hong, Wenxuan Zhou, jingli zhang jingli, Guodong Zhou, and Qiaoming Zhu. 2018. Self-regulation: Employing a generative adversarial network to improve event detection. In *Proc. ACL2018*.

Ruihong Huang and Ellen Riloff. 2012. Bootstrapped training of event extraction classifiers. In *Proc. EACL2012*.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pages 254–262.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Keizo Kato, Yin Li, and Abhinav Gupta. 2018. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–251.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.

Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. 2018. Recurrent tubelet proposal and recognition networks for action detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 303–318.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proc. ACL2013*.

Shasha Liao and Ralph Grishman. 2011. Acquiring topic features to improve event extraction: in pre-selected and balanced collections. In *Proc. RANLP2011*.

Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 3–11. ACM.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018a. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018b. Jointly multiple events extraction via attention-based graph information aggregation. In *Proc. EMNLP2018*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13–23.

Zhigang Ma, Xiaojun Chang, Zhongwen Xu, Nicu Sebe, and Alexander G Hauptmann. 2017. Joint attributes and event analysis for multimedia event detection. *IEEE transactions on neural networks and learning systems*, 29(7):2921–2930.

Arun Mallya and Svetlana Lazebnik. 2017. Recurrent models for situation recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 455–463.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proc. NAACL-HLT2016*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

AG Amitha Perera, Sangmin Oh, P Megha, Tianyang Ma, Anthony Hoogs, Arash Vahdat, Kevin Cannons, Greg Mori, Scott Mccloskey, Ben Miller, et al. 2012. Trecvid 2012 genie: Multimedia event detection and recounting. In *In TRECVID Workshop*. Citeseer.

Wasifur Rahman, Md Kamrul Hasan, Amir Zadeh, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019. M-bert: Injecting multimodal information in the bert structure. *arXiv preprint arXiv:1908.05787*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. 2019. Adversarial representation learning for text-to-image matching. In *The IEEE International Conference on Computer Vision (ICCV)*.

Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. 2018. Learning visually-grounded semantics from contrastive adversarial samples. *arXiv preprint arXiv:1806.10348*.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Mitchell Stephens. 1998. *The Rise of the Image, The Fall of the Word*. New York: Oxford University Press.

Shupeng Su, Zhisheng Zhong, and Chao Zhang. 2019a. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *The IEEE International Conference on Computer Vision (ICCV)*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019b. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. Videobert: A joint model for video and language representation learning. In *The IEEE International Conference on Computer Vision (ICCV)*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Alejandro H Toselli, Verónica Romero, and Enrique Vidal. 2007. Viterbi based alignment between text images and their transcripts. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).*, pages 9–16.

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.

Chuan Wang, Sameer Pradhan, Xiaoman Pan, Heng Ji, and Nianwen Xue. 2016. Camr at semeval-2016 task 8: An extended transition-based amr parser. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1173–1178, San Diego, California. Association for Computational Linguistics.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015a. Boosting transition-based amr parsing with refined actions and auxiliary analyzers. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 857–862, Beijing, China. Association for Computational Linguistics.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015b. A transition-based algorithm for amr parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, Denver, Colorado. Association for Computational Linguistics.

Rui Wang, Deyu Zhou, and Yulan He. 2019. Open event extraction from online text using a generative adversarial network. *arXiv preprint arXiv:1908.09246*.

Rebecka Weegar, Kalle Åström, and Pierre Nugues. 2015. Linking entities across images and text. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 185–193.

Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. 2019a. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 284–293.

Hank Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019b. Univse: Robust visual semantic embeddings via structured semantic representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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 the 57th Conference of the Association for Computational Linguistics*, pages 5284–5294.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542.

Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang. 2015. Eventnet: A large scale structured concept library for complex event detection in video. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 471–480. ACM.

Tongtao Zhang, Heng Ji, and Avirup Sil. 2019. Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence Vol 1 (2): 99-120*.

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 cross-modal integration. In *Proceedings of the 25th ACM International Conference on Multimedia (ACMMM2017)*.

Yifan Zhang, Changsheng Xu, Yong Rui, Jinqiao Wang, and Hanqing Lu. 2007. Semantic event extraction from basketball games using multi-modal analysis. In *2007 IEEE International Conference on Multimedia and Expo*, pages 2190–2193. IEEE.