Coreference by Appearance: Visually Grounded Event Coreference Resolution

Liming Wang[1], Shengyu Feng[1], Xudong Lin[2], Manling Li[1], Heng Ji[1], Shih-Fu Chang[2]
[1]University of Illinois at Urbana-Champaign, [2]Columbia University

Abstract

Event coreference resolution is critical to understand events in growing number of online news with multiple modalities including text, video, speech, etc. However, the events and entities depicting in different modalities may not be perfectly aligned and can be difficult to annotate, which makes the task especially challenging with little supervision available. To address the above issues, we propose a supervised model based on attention mechanism and an unsupervised model based on statistical machine translation, capable of learning the relative importance of modalities for event coreference resolution. Experiments on a video multimedia event dataset show that our multimodal models outperform text-only systems in the event coreference resolution task. A careful analysis reveals that the performance gain of the multimodal model especially under the unsupervised setting comes from better learning of visually salient events.

1 Introduction

With the advance of Internet, news articles nowadays are becoming increasingly multimodal. For example, news about a recent launch of a SpaceX rocket may contain the traditional text article, a drone video of the launch and a narration from the talking head or reporter. Such a diverse source of information provides us a multi-faceted understanding of the newsworthy event, and helps us better follow the content of the article. Event coreference resolution, the task of resolving whether two words or phrases called *mentions* in the article refer to the same real-world event or not, is the first step toward such an understanding.

While human can efficiently and accurately identify the coreference relations when reading an article, this task has shown to be quite challenging to machines. The state-of-the-art coreference resolution system not only accepts textual input, but typically requires over 4000 annotated articles with more than 400 words fully annotated with mention span and coreference relation to train (Lee et al., 2018). Further, since event coreference resolution requires careful reading and deep understanding of the text, it becomes prohibitively expensive to annotate for low-resource languages.

Another issue with the current system is that the model lacks a way to incorporate real-world knowledge typically required for coreference. One potential way to incorporate such knowledge is to consider an alternative modality – videos. A fascinating aspect of multimodal event coreference resolution is that the correlation between the visual modality and the text can be quite weak, and the text tends to contain most of the information for resolving coreference. Nevertheless, the meanings of the event triggers often require additional information sources to resolve. For example, we never know an "attack" is a physical one or verbal one without seeing or imagining a particular scene based on additional contextual constraints, or it is hard to know whether "chant" and "protest" co-refer without knowing the event in the text is a demonstration. But if a video of the relevant events is played, we can immediately understand the relations between the trigger words. In this paper, we answer the following research question: How to leverage weakly aligned multimodal data for event coreference resolution system, especially in an unsupervised setting?

2 Approach

2.1 Notations and Definitions

A *multimodal document* $S = (M, V)$ is defined as the tuple of two sequences: the sequence of *textual event mentions* $M = (m_0, \cdots, m_I)$, including the artificial root mention m_0, and the sequence of *visual event mentions* $V = (v_1, \cdots, v_J)$, i.e., event-level segments from the related video. Each text mention is represented as a span of con-

Proceedings of the 4th Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2021), pages 132–140
Punta Cana, Dominican Republic, November 10–11, 2021. ©2021 Association for Computational Linguistics

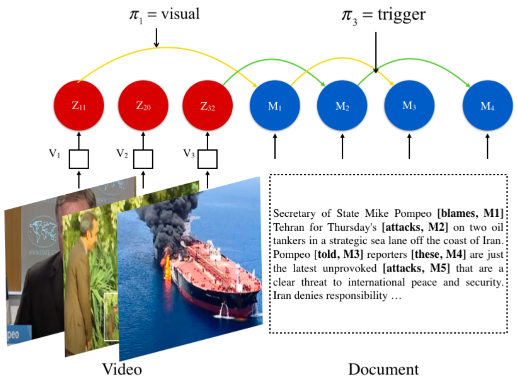

π_1 = visual
π_3 = trigger
Z11
Z20
Z32
M1
M2
M3
M4
V1
V2
V3
Secretary of State Mike Pompeo **[blames, M1]** Tehran for Thursday's **[attacks, M2]** on two oil tankers in a strategic sea lane off the coast of Iran. Pompeo **[told, M3]** reporters **[these, M4]** are just the latest unprovoked **[attacks, M5]** that are a clear threat to international peace and security. Iran denies responsibility …
Video
Document

Figure 1: The proposed unsupervised multimodal coreference resolution model. The bubbles represent the random variables involved in the generative process, the squares represent raw feature inputs and arcs represent the values taken by the antecedent variables

Name	Definition
x_n	n-th token in a document
f_t	t-th frame in a video
m_i	i-th text event mention in a document
v_j	j-th visual event mention in a video
c_i	the position of antecedent for text mention i
MLP	Multi-layer perecptron
$F_v(m_i, v_j)$	Crossmedia coreference score between textual mention m_i and visual mention v_j
$F_c(m_i, m_{i'})$	Textual context score between textual mention m_i and $m_{i'}$
$F_{cv}(m_i, m_{i'})$	Visual context score between textual mention m_i and $m_{i'}$
π_i	mode of i-th text event mention
μ_k	the centroid of k-th visual cluster
$p(m_i \vert m_j, \pi_i)$	probability of features used in mode π_i for m_i given mention m_j is its antecedent
$q(c_i \vert i)$	probability that the antecedent of mention i is at position c_i in attribute mode

Table 1: Definitions of key notations

secutive word tokens $(x_{\text{start}(i)}, \cdots, x_{\text{end}(i)})$, where $\text{start}(i)$ and $\text{end}(i)$ are the indices of the start and end token of the span. The mention spans can be extracted either using a binary classifier as done in previous works (Lee et al., 2017, 2018) or by an end-to-end event extraction system such as OneIE (Lin et al., 2020) as in this work. Similarly, each visual mention is a span of consecutive embeddings of video frames $(y_{\text{start}(j)}, \cdots, y_{\text{end}(j)})$ extracted by human annotators. Inspired by (Ma et al., 2016), the coreference relations between events can be represented by a sequence of latent *antecedent variables* $C = (c_1, \cdots, c_I)$ associated with each event mentions, where $c_i = j$ means that mention i is the parent of (textual or visual) mention j in the tree formed by mentions that are coreferent. For within-text coreference, $c_i \in \{0, \cdots, i-1\}$ and for crossmedia coreference, $c_i \in \{1, \cdots, J\}$.

2.2 Supervised model for end-to-end event coreference resolution

In the setting of supervised coreference resolution, the model tries to learn the coreference relation between textual mention pairs $(m_i, m_{i'})$ from labeled mention pairs $\{(m_i, m_{i'}), y^c_{ii'}\}$, where $y^c_{ii'} = 1$ if m_i and $m_{i'}$ are coreferent and 0 otherwise. In the multimodal setting, we allow the model to observe *crossmedia coreference labels* y^v_{ij}'s, where $y^v_{ij} = 1$ if textual mention m_i and visual mention v_j are coreferent.

Our end-to-end model consists of two main components: a crossmedia coreference resolver and a text-only coreference resolver. For both resolver, embeddings from the RoBERTa-large (Liu et al., 2019) transformer is used as input to the textual mention encoder, and embeddings from the penultimate layer of the TSN action recognition model (Contributors, 2020) is used as input to the visual mention encoder. The transformer is frozen during training. To represent the text mentions, instead of concatenating start token, end token and head word token embeddings as in (Lee et al., 2017, 2018), the textual encoder forms the mention embedding by averaging the embeddings within its span:

$$m_i = \frac{\sum_{n=\text{start}(i)}^{\text{end}(i)} x_n}{\text{end}(i) - \text{start}(i)} \tag{1}$$

Similarly, the visual mention encoder compute a fixed-dimensional embedding for each visual mention by averaging over its frame embeddings f_t's:

$$v_j = \frac{\sum_{t=\text{start}(j)}^{\text{end}(j)} f_t}{\text{end}(j) - \text{start}(j)}. \tag{2}$$

This saves memory without losing significant amount of information since most of the triggers consist of only one token.

To compute the probability that a visual mention and a textual mention are coreferent, the crossmedia resolver first encodes each textual and visual mention pair (m_i, v_j) into the common embedding space as $(m^{\mathbb{C}}_i, v^{\mathbb{C}}_j)$. This is done by two three-layer bidirectional LSTMs (Hochreiter and Schmidhuber, 1997) (BiLSTMs) as visual and textual crossmedia encoder:

$$m_i^{\mathbb{C}} = \text{BiLSTM}_i(M) \quad (3)$$

$$v_j^{\mathbb{C}} = \text{BiLSTM}_j(V). \quad (4)$$

The crossmedia coreference score is then simply the dot-product similarity between the embeddings:

$$F_v(m, v) = m^{\mathbb{C}\top} v^{\mathbb{C}} \quad (5)$$

To compute the probability that two textual mentions are coreferent, we first computed the text-only coreference score using a multilayer perceptron (MLP):

$$F_c(m, m') = \text{MLP}_c([m, m', m \odot m']). \quad (6)$$

To fuse visual and linguistic information, we use a special attention mechanism (Bahdanau et al., 2015) similar to the one proposed in (Yu et al., 2019). In particular, the model attends over the visual mention embedding to create a contextualized embedding as additional feature for each textual mention:

$$\alpha_{m_i,v_j} = \frac{\exp(F_v(m_i, v_j))}{\sum_v \exp(F_v(m_i, v))} \quad (7)$$

The text resolver then utilizes the contextualized embeddings as additional cues to resolve coreference between the mention pairs:

$$\beta_{m_i} = \max_j \alpha_{m_i,v_j} \quad (8)$$

$$\beta_{m_i,m_{i'}} = \max_j \alpha_{m_i,v_j} \alpha_{m_{i'},v_j} \quad (9)$$

$$F_{cv}(m_i, m_{i'}) = \text{MLP}_{cv}([\beta_{m_i}; \beta_{m_{i'}}; \beta_{m_i,m_{i'}}]). \quad (10)$$

Intuitively, the value of β_{m_i}'s scores how “visual” a particular mention is and $\beta_{m_i,m_{i'}}$ scores how likely the two mentions refer to the same visual event. The final text coreference score is then the weighted sum of the textual and viusal context score:

$$F(m_i, m_{i'}) = \lambda_v \cdot F_{cv}(m_i, m_{i'}) + (1 - \lambda_v) \cdot F_c(m_i, m_{i'}). \quad (11)$$

The score can be then interpreted as the logit of the probability $p(y^c_{ii'} | m_i, m_{i'})$ and learned by maximizing the standard binary cross-entropy loss. We will refer to this later as the *weight fusion* model.

2.3 Amodal SMT: Unsupervised model for multimodal event coreference resolution

The supervised multimodal model requires a large amount of labeled training data, which makes it unsuitable for real-world multimodal documents where coreference labels are mostly unavailable. Therefore, it is desirable to design an unsupervised coreference resolution algorithm to learn coreference relations without of the need of such labels. Instead of maximizing the likelihood of the coreference labels as in the supervised case, our unsupervised multimodal coreference resolution model instead tries to maximize the conditional likelihood $p(M|V)$, with independence assumptions based on (Ma et al., 2016):

$$p(M|V) = \prod_{i=1}^{I} p(\pi_i|i) p(c_i|i, \pi_i) \cdot p(m_i|v_{c_i}, m_{c_i}, \pi_i), \quad (12)$$

where $\Pi = (\pi_1, \cdots, \pi_I)$ are discrete variables called *mode* variables. This likelihood resembles that of the classical statistical machine translation (SMT) model (Brown et al., 1993) where the text mentions form the source language sentence the entire multimoal document is the target language sentence. Therefore, we call the model *amodal SMT*. In each mode, the model will resolve the current mention with a different subset of linguistic features. This is motivated by the observation that while resolving proper-nominal mention pairs such as “the Great War” and “World War I” mostly requires semantic features such as embedding similarity, resolving pronoun-nominal mention pairs such as “the Great War” and “it” relies more on syntactic and discourse features such as part-of-speech tags, parse trees and sentence distance. The need of distinct features for different coreferent types is also demonstrated in previous works (Haghighi and Klein, 2009; Ratinov and Roth, 2012; Lee et al., 2013; Ma et al., 2016).

The overview of our unsupervised model is shown in Fig. 1. Our unsupervised model has three modes, one *visual* mode and two *textual* modes, one called *trigger*-matching mode and the other called *attribute*-matching mode. In visual mode, π_i = visual, the antecedent of mention i is restricted to only visual event mentions. In textual modes, the antecedent of event mention i is restricted to only previous textual event mentions as in (Ma et al., 2016). Further, in trigger-matching

Mode	Feature	Description
trigger	Head Lemma	lemmatized head word extracted using NLTK and AllenNLP
	Number	consider only mention pairs with the same number
	Event Type	consider only mention pairs with different types
	GloVe	consider only mention pairs with cosine similarity ≥ 0.5
	String Match	consider mention pair if normalized edit distance ≥ 0.5
attribute	Mention Type	same as in (Ma et al., 2016)
	Number	same as in (Ma et al., 2016)
	Event Type	same as in trigger mode
	Arguments	consider only events with non-overlapping argument entity types or sharing argument of the same type
	Sentence distance	same as in (Ma et al., 2016)
visual	Head Lemma	same as in trigger mode
	Number	consider only pairs with the same number
	Event Type	consider only mention pairs with different types
	Is Visual	consider mention whose type of event appears in at least one video

Table 2: Features for different modes of amodal SMT

mode with π_i = trigger, the model will use primarily semantic features such as the embedding of the trigger. Finally, in attribute-matching mode with π_i = attribute, it will use mostly syntactic and discourse features. More details about the features can be found in Table. 2. Therefore, we have:

$$p(m_i|v_{c_i}, m_{c_i}, \pi_i) \\ =: \begin{cases} p(m_i|m_{c_i}, \pi_i), \text{ if } \pi_i \in \{\text{trigger, attribute}\} \\ p(m_i|v_{c_i}), \text{ if } \pi_i = \text{visual} \end{cases} \tag{13}$$

In textual mode, two textual event mentions $(m_i, m_{i'})$ are *coreferent* if they belong to the same text coreference chain specified by the antecedent variables c_i's; in visual mode, however, they are coreferent only if they refer to the same visual mention, i.e., $c_i = c_{i'}$. Further, the modes are assumed to be equally likely and the prior of antecedent is learnable in the attribute-matching mode to model the discourse structure of the document:

$$p(c_i|i, \pi_i) = \begin{cases} q(c_i|i), \text{ if } \pi_i = \text{attribute}, \\ \frac{1}{i}, \text{ if } \pi_i = \text{trigger}, \\ \frac{1}{J}, \text{ if } \pi_i = \text{visual} \end{cases} \tag{14}$$

Since the representation of a visual event is continuous, we assume that the visual features of similar events tend to cluster together and can be well modeled by a Gaussian distribution with mean μ_k and variance σ_k for each event type k. However, the type of a visual event mention is ambiguous given the visual feature alone because events like "attack" and "protest" tend to co-occur in the video, leading to overlapping visual mentions and similar visual features. To model this, we allow each visual mention j to represent a *different* event type for each text mention i it refers to with cluster variables $z_{ji}, i = 1, \cdots, I, j = 1, \cdots, J$. With such a multi-label assumption, the probabilities $p(m_i|v_j)$ become:

$$p(m_i|v_j) = \sum_{z_{ji}=1}^{K} p(z_{ji}|v_j)p(m_i|z_{ji}, \pi_i) \tag{15}$$

$$p(z_{ji} = s|v_j) := \frac{\exp(\frac{-\|v_j - \mu_s\|^2}{2\sigma_s^2})}{\sum_{k=1}^{K} \exp(\frac{-\|v_j - \mu_k\|^2}{2\sigma_k^2})} \tag{16}$$

2.4 Antecedent prediction

We can use the maximum a posteriori (MAP) estimator to find the optimal antecedent for each mention separately as:

$$\arg\max_{c_i} q(c_i|i)p(m_i|m_{c_i}, v_{c_i}, \pi_i). \tag{17}$$

This approach, however, is less robust when the visual event boundaries are noisy or overlapping, as in our case. Therefore, we instead *marginalize* out all possible alignments for each mention pairs to obtain the pairwise score as in the textual mode:

$$\arg\max_j p(c_i = c_j|m_i, m_j, V, \pi_i = \text{visual}) \\ = \arg\max_j \sum_{c=1}^{J} p(m_i|v_c)p(m_j|v_c) \tag{18}$$

2.5 Training

The whole model can be trained using EM algorithm (Dempster et al., 1977) with more details in Algorithm 1. We initialized the visual centroids using K-means algorithm and the other parameters uniformly on their supports. The model is then trained for 10 EM iterations. We experimented with various number of visual clusters, and the result is shown in Figure 2. The subsequent results are all obtained with 15 clusters.

3 Experiments and Results

3.1 Data and Experiments Setup

The video M2E2 dataset (Anonymous, 2021), inspired by the image M2E2 dataset (Li et al., 2020),

Algorithm 1: Learning with EM

```
Initialization: initialize $t_0, q_0, \mathbf{M}_0$
for $\tau = 1$ to $T$ do
    for each document $D$ do
        for $i = 1$ to $I$ do
            Update counts for $p(m'|m, \pi)$ and $q(c|i)$ as in (Ma et al., 2016);
            for $c = 1$ to $J$ do
                $L_{icz} = \frac{p(z_{ci}|v_c)p(m_i|z_{ci})}{\sum_j p(z_{ji}|v_j)p(m_i|z_{ji})}$
                $c(m_i, z) += \sum_{c=1}^{J} L_{icz}$
                $\Delta\mu_z += (L_{icz} - p(z_{ci}|v_c))\frac{v_c - \mu_z}{\sigma_z}$
        Update $p(m'|m, \pi)$ and $q(c|i)$ parameters as in (Ma et al., 2016);
        $\mu_z = \mu_z + \eta\Delta\mu_z$
```

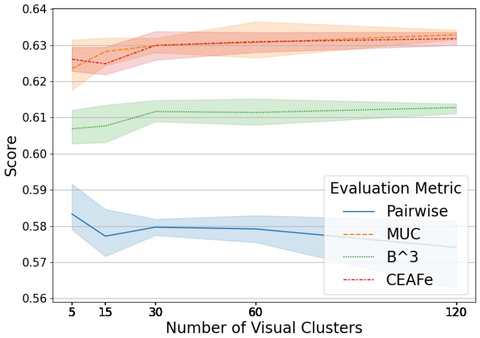


Figure 2: Coreference scores vs. number of visual clusters used by amodal SMT

consists of 860 manually annotated documents, each paired with a 1-2 minutes video released by authentic news outlets such as British Broadcasting Company (BBC), Voice of American (VoA) and Reuters. The videos are found from Youtube channel with event types as searching keywords such as “attack” and “elect”. The textual coreference labels are annotated by two annotators with three rounds of adjudication using the Brat interface (Stenetorp et al., 2011) following ACE2005 guideline. More details about this dataset are available in Table 3 as well as (Anonymous, 2021). We follow the 3:1 random split of training/test sets by (Anonymous, 2021). Three standard metrics for coreference resolution are used to evaluate the models at the cluster level: the MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), Entity-based CEAFe (Luo, 2005).

	Video M2E2
# Train/Test	645/215
# Event mentions	4158
# Event clusters	647

Table 3: Dataset statistics. Only non-singleton clusters are counted

The CoNLL score is calculated as the average of these three scores and we used the CoNLL 2012 scorer (Pradhan et al., 2014). We also used the more straightforward pairwise F1 score and mean average precision (mAP) score, which score the overlap between the set of predicted coreference links and the gold links.

3.2 Baselines

We used the following supervised coreference resolution model as baseline, namely,

- **Text-only model**: an implementation of the state-of-the-art neural coreference resolution model (Lee et al., 2018) by (Cattan et al., 2020), which uses RoBERTa (Liu et al., 2019) to extract contextualized word embeddings. All parameter and optimization settings are made the same as the multimodal supervised model described in Section 2.2. We used ground truth mention spans and within-document antecedent prediction instead of using their cross-document hierarchical clustering algorithm.

In addition, we used the following unsupervised models as baselines, including

- **HDP-LF**: re-implementation of (Bejan and Harabagiu, 2010) with a rich set of semantic and syntactic features;

- **DD-CRP**: re-implementation of (Yang et al., 2015) with additional pairwise features compared to HDP-LF;

- **Text-only SMT**: proposed in (Ma et al., 2016) with features adapted to event coreference resolution for each textual mode described in Section 2.

3.3 Results

The overall results on video M2E2 are shown in Table. 4 respectively. We make the following claims:

Effect of Visual Supervision The supervised multimodal model outperforms the text-only model by 2% in pairwise F1 and 3% in CoNLL (average of the last three metrics), demonstrating the effectiveness of multimodal data for coreference resolution. Further, amodal SMT achieves 3.5% and 2.4% improvement in pairwise and CoNLL score respectively over its text-only counterpart, 5.8% and 7.7% improvement over the HDP-LF method, indicating that video features contain additional information to the text features and can benefit unsupervised coreference resolution. Further, the amodal SMT also outperforms the text-only supervised baselines, though with the help of event type information. However, it does not outperform the DD-CRP approach, possibly due to the lack of global clustering mechanism. A breakdown of pairwise coreference scores across event types is shown in Table 6. We found that the amodal SMT performs significantly better than the text-only approach in visually salient event types such as Demonstrate, TransportPerson and Meet, but perform worse on rare events such as ArrestJail and ReleaseParole or non-visual events such as Die. The amodal SMT also performs worse in visual event such as Attack and Broadcast, probably because the model is confused its visual representation with those of other event types such as Demonstrate as the two events often co-occur in the video. Although the amodal SMT approach underperforms in 7 out of 10 most frequent event types, we found its overall performance superior to the text-only approach because the large improvement in events such as Demonstrate, TransportPerson and Meet (about 18% absolute in average) outweighs the performance drop. On the other hand, for the supervised event coreference, we found a reverse trend that the multimodal model performs superior to the text-only model in 7 out of the 10 most frequent event types, though the improvements are minor for types such as Attack and Broadcast. Furthermore, all the improved event types in the unsupervised setting suffer from performance drop in the supervised setting except Meet event, and all events except Elect that suffer from performance drop in the supervised setting enjoy a performance gain in the supervised setting. This suggests that different aspects of the visual feature is used for the two settings.

Effect of Multi-mode mechanism From Table 4 and Table 5, removing any mode leads to degradation to the model, suggesting the importance of using multi-mode mechanism. In particular, the most performance degradation results from removing trigger mode, followed by removing both textual modes, then by removing visual mode and finally by removing attribute mode. This suggests that semantic information in video is less important than semantic information in text, but more important than discourse information. We also found that HDP does not perform as well as the text and amodal SMTs on video M2E2, probably because of the lack of multi-mode mechanism in the model.

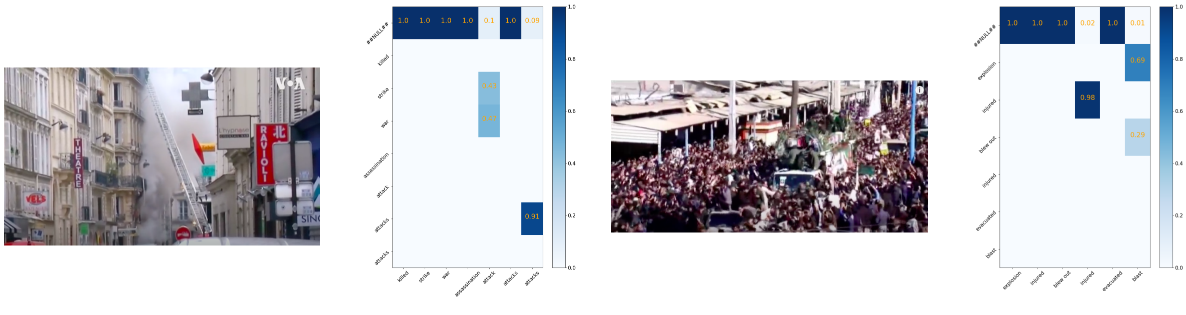
Video M2E2	Pairwise F1	MUC	B^3	CEAFe
(Cattan et al., 2020)	61.0±0.7	66.4±1.0	63.6±0.9	64.3±0.8
Weight Fusion (Ours)	**61.6**±0.5	**67.0**±1.0	**64.4**±1.2	**64.7**±1.8
HDP-LF	51.9±0.0	55.2±0.0	53.7±0.0	54.2±0.0
DD-CRP	56.4±1.7	62.3±0.6	60.7±0.8	**63.3**±1.2
Text-only SMT	54.1±0.0	60.3±0.0	58.7±0.0	59.8±0.0
Amodal SMT (Ours)	**57.7**±0.7	**62.8**±0.4	**60.8**±0.5	62.5±0.4

Table 4: Event coreference resolution performance of all systems on the video M2E2 dataset. Models above the line are supervised models while models below are unsupervised. All results are averages of 4 runs. Note that the confidence interval of text-only SMT is 0.0 since the model is initialized and trained deterministically.

Qualitative analysis To better understand the performance gain of using videos, we look at two typical examples in Fig. 3. we can see that the text-only SMT fails to cluster the trigger word "blew out" with the word "explosion" and "blast", probably because the word embeddings of multi-phrase expressions such as "blew out" and of"explosion" are quite different. However, thanks to the context provided by the video, the amodal SMT is able to identify the two triggers as coreferent, improving the recall of the system. Similarly, we found that while "killed" and "assassination" have similar meaning, the similarity score of their word embeddings is below the threshold we use (0.5), resulting in a miss detection of the coreferent pair. However, with the help of video, the amodal SMT is able to resolve the coreference using their correlations with similar type of videos in the visual mode. Nevertheless, we observe that introducing the visual mode can sometimes lead to false positives.

4 Related Work

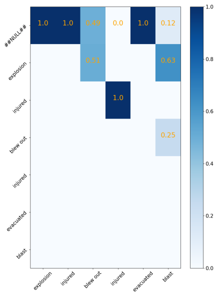
Motivated by event coreference resolution for low-resource languages, unsupervised event coreference resolution algorithms have been proposed (Haghighi and Klein, 2007, 2010; Bejan



(a) Text SMT

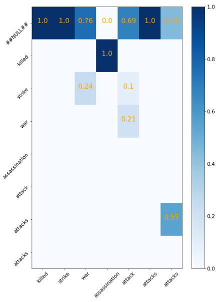
(b) Text SMT

Document	
A powerful **explosion** and fire apparently caused by a gas leak at a Paris bakery Saturday injured several people , **blew out** windows and overturned cars ... Witnesses described the overwhelmingly sound of the **blast** and people trapped inside nearby buildings.	
Text SMT	(explosion, blast)
Amodal SMT	(explosion, blew out, blast)



(c) Amodal SMT

Document	
Hundreds of thousands of mourners have turned out in Iran to receive home the remains of Qasem Soleimani, the general **killed** by a US drone strike in Iraq ... His **assassination** marked a significant escalation between Iran and the US. Iran 's Supreme Leader Ayatollah Khamenei, ...	
Text SMT	()
Amodal SMT	(killed, assasination)



(d) Amodal SMT

Figure 3: Coreference examples when adding a video as an additional modality helps. Only relevant event of the video is shown. Green color denotes true positives detected by the amodal SMT but not by the text-only SMT. The grid plots shows the posterior distribution $p(c_i|M,V), i = 1, \cdots, I$, where the x-axis are the text mentions and y-axis are their antecedent candidates.

Video M2E2	Pairwise F1	MUC	B^3	CEAFe
Full model	**57.7**±0.7	**62.8**±0.4	**60.8**±0.5	**62.5**±0.4
-textual mode	50.7±0.2	53.0±0.2	54.0±0.1	55.2±0.1
Δ	-7.0	-9.8	-6.8	-7.3
-trigger mode	47.7±1.5	53.4±0.8	51.5±0.7	53.0±0.8
Δ	-10.0	-9.4	-9.3	-9.5
-attribute mode	55.1±0.4	61.3±0.3	59.5±0.4	61.7±0.3
Δ	-2.6	-1.5	-1.3	-0.8

Table 5: Ablation study on the effect of removing different modes on amodal SMT. “-textual” means using visual mode only

Event type	Frequency	Text SMT	Amodal SMT	Supervised (Text Only)	Supervised (Multimodal)
Attack	2389	59.0±0.0	58.1±1.3	65.3±1.1	65.7±1.0
Die	1148	39.3±0.0	37.6±1.4	39.9±3.8	47.6±4.2
Demonstrate	874	55.0±0.0	73.6±0.2	69.8±2.1	69.5±2.2
Injure	348	75.0±0.0	70.6±0.0	35.3±11.2	48.5±17.3
TransportPerson	325	32.4±0.0	52.4±1.6	30.7±4.8	25.4±4.5
Broadcast	319	46.2±0.0	43.5±2.4	18.2±0.6	19.3±1.1
Elect	238	50.0±0.0	7.9±0.3	46.4±1.7	45.5±7.6
Meet	214	27.3±0.0	45.6±3.9	21.1±2.4	23.4±2.1
ArrestJail	165	80.0±0.0	38.5±5.2	0.0±0.0	0.0±0.0
ReleaseParole	139	100.0±0.0	66.7±0.0	25.0±0.0	33.7±15.2

Table 6: Pairwise F1 breakdown across the 10 most frequent event types for the text and amodal models. Numbers colored in green denote improvement over the text-only baseline and those in red denote degradation. Frequency for each event type is the number of distinct mention pairs that have at least one of the mention belonging to that event type.

and Harabagiu, 2010; Yang et al., 2015; Ma et al., 2016). Notably, (Bejan and Harabagiu, 2010) proposed a hierarchical Dirichlet process (HDP) to capture recurring surface patterns on the global cluster level such as matching trigger lemma, arguments, neighboring events, etc, with a rich-get-richer mechanism. (Yang et al., 2015) developed a probabilistic model to combine global features with rich pairwise features by learning a distance metric between mentions. (Ma et al., 2016) reformulated text-only coreference resolution as a translation process from each mention to its antecedent with different resolution modes. Previous works on multimodal event coreference resolution focus on the supervised setting with specialized datasets. (Zhang et al., 2015) proposed a pipeline approach based on a pretrained text-only coreference system with additional aligned videos for each event and their visual similarity as visual features for cross-document coreference. However, their cross-document model is not applicable to our within-document coreference setting. (Yu et al., 2019) proposed a supervised end-to-end model to resolve pronouns in short conversational texts, which uses an attention mechanism (Bahdanau et al., 2015) to control the tradeoff between cross-modal and within-modal coreference resolution. Instead of directly applying their model, which is tailored to entity pronoun resolution with images and dialogues and may not be applicable to event coreference resolution with weakly correlated news articles and videos, we adapt their attention mechanism and use crossmedia coreference instead of object detection result as additional supervision. (Li et al., 2020) employed an multimoda event extraction system to perform cross-modal coreference.

5 Conclusions

In this work, we proposed one supervised and one unsupervised model for multimodal event coreference resolution, which outperform the text-only baselines on a realistic multimedia event dataset.

References

Anonymous. 2021. Joint multimedia event extraction from video and article. *Under review*.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, Uppsala, Sweden. Association for Computational Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263 – 311.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and I. Dagan. 2020. Streamlining cross-document coreference resolution: Evaluation and modeling. *ArXiv*, abs/2009.11032.

MMAction2 Contributors. 2020. Openmmlab's next generation video understanding toolbox and benchmark. `https://github.com/open-mmlab/mmaction2`.

Arthur Dempster, Nan Laird, and Donald Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855, Prague, Czech Republic. Association for Computational Linguistics.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1152–1161, Singapore. Association for Computational Linguistics.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393, Los Angeles, California. Association for Computational Linguistics.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. 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 the Annual Meeting of the Association for Computational Linguistics*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Xuezhe Ma, Zhengzhong Liu, and Eduard Hovy. 2016. Unsupervised ranking model for entity coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1012–1018, San Diego, California. Association for Computational Linguistics.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2012. Learning-based multi-sieve co-reference resolution with knowledge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1234–1244, Jeju Island, Korea. Association for Computational Linguistics.

Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Bionlp shared task 2011: Supporting resources. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 112–120, Portland, Oregon, USA. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*.

Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent Bayesian model for event coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:517–528.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5123–5132, Hong Kong, China. Association for Computational Linguistics.

Tongtao Zhang, Hongzhi Li, Heng Ji, and Shih-Fu Chang. 2015. Cross-document event coreference resolution based on cross-media features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 201–206, Lisbon, Portugal. Association for Computational Linguistics.