Information Extraction
Event-Centric Natural Language Understanding (Part I)

Manling Li
Department of Computer Science
University of Illinois at Urbana-Champaign
What is an event?

- An Event is a specific occurrence involving participants.
- An Event is something that happens.
- An Event can frequently be described as a change of state.

Chart from (Dölling, 2011)
Event Extraction

➢ Supervised Event Extraction
   ▪ Schema-guided Event Extraction
   ▪ Document-level Event Extraction

◼ Cross-domain Zero-shot Transfer for Event Extraction
◼ Cross-lingual Transfer for Multi-lingual Event Extraction
◼ Cross-media Structured Common Space for Multimedia Event Extraction
What is Information Extraction (IE)?

- Extract **structured information and knowledge** from **unstructured data** of heterogeneous data types, in various domains, genres, languages, and data modalities.

In Baghdad, a cameraman **died** when a **combat tank** **fired** on the Palestine Hotel.

- It’s naturally a structure prediction task! Convert unstructured sequences to graphs.
“Old” Days: Supervised Learning with Hand-crafted Features

- **Trigger Labeling**
  - **Lexical**
    - Tokens and POS tags of candidate trigger and context words
  - **Dictionaries**
    - Trigger list, synonym gazetteers
  - **Syntactic**
    - the depth of the trigger in the parse tree
    - the path from the node of the trigger to the root in the parse tree
    - the phrase structure expanded by the parent node of the trigger
    - the phrase type of the trigger
  - **Entity**
    - the entity type of the syntactically nearest entity to the trigger in the parse tree
    - the entity type of the physically nearest entity to the trigger in the sentence

- **Argument Labeling**
  - **Event type and trigger**
    - Trigger tokens
    - Event type and subtype
  - **Entity**
    - Entity type and subtype
    - Head word of the entity mention
  - **Context**
    - Context words of the argument candidate
  - **Syntactic**
    - the phrase structure expanding the parent of the trigger
    - the relative position of the entity regarding to the trigger (before or after)
    - the minimal path from the entity to the trigger
    - the shortest length from the entity to the trigger in the parse tree

*(Chen and Ji, 2009)*
• Reduce feature engineering efforts to some extent (Feng et al., 2016)
• But still rely on human annotated clean training data still fragile to noise in training data
Add symbolic features by concatenating them with embeddings (Nguyen et al., 2016)
Pipelined models suffer from the error propagation problem and disallow interactions among components.

Existing neural models do not explicitly model cross-subtask and cross-instance interactions among knowledge elements.

Example: *Prime Minister Abdullah Gul resigned earlier Tuesday to make way for Erdogan, who won a parliamentary seat in by-elections Sunday.*
Our OneIE framework extracts the information graph from a given sentence in four steps: encoding, identification, classification, and decoding.
Move from Entity-Centric to Event-Centric NLU

2014 Thai coup d'état: Однако протесты и блокада длились уже почти 3 месяца, а военные так и не перешли к действиям. 2013 Egyptian coup d'état: ... General Abdel Fattah el-Sisi announced that he would be calling on the Presidential and Shura Council elections.

Ukrainian crisis: At 09:25, protesters pushed the Berkut back to the October Palace after security forces tried to set fire to Kiev Conservatory, which was being used as a field hospital for wounded protesters.

Sense Embedding

AMR-based Local Context

External Knowledge

ConflictEvent, PoliticalConflict, CivilUnrest

International Intervention

ControlEvent, PreventionAction

CoaTfire

Ukrainian crisis

Chechen-Russian Conflict
We design a set of *global feature templates* (e.g., event_type₁ – role₁ – role₂ – event_type₂: an entity acts a role₁ argument for an event_type₁ event and a role₂ argument for an event_type₂ event in the same sentence). A more comprehensive event schema library is inducted following (Li et al, 2020).

The model learns the *weight* of each feature during training.
Incorporating Global Features

• Given a graph $G$, we generate its global feature vector as $f(G)$, where $f$ is a function that evaluates a certain feature and returns a scalar. For example,

$$f_i(G) = \begin{cases} 1, & G \text{ has multiple ATTACK events} \\ 0, & \text{otherwise} \end{cases}$$

• Next, we learn a weight vector $u$ and calculate the global feature score of $G$ as the dot production of $u$ and $f_G$.

• **Global score** of a graph: local graph score + global feature score:

$$s(G) = s'(G) + u f_G$$

• We assume that the gold-standard graph for a sentence should achieve the highest global score and minimize the following loss function:

$$\mathcal{L}^G = s(\hat{G}) - s(G)$$
Decoding

- We use beam search to decode the information graph
- Example: *He also brought a check from Campbell to pay the fines and fees.*
Experiment Results

- We conduct our experiments on ACE (Automatic Content Extraction) 2005 (F-score, %)

<table>
<thead>
<tr>
<th>Model</th>
<th>ACE05-R</th>
<th>ACE05-E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Entity</td>
<td>Relation</td>
</tr>
<tr>
<td>DyGIE++</td>
<td>88.6</td>
<td>63.4</td>
</tr>
<tr>
<td>DyGIE++*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OneIE</td>
<td>88.8</td>
<td>67.5</td>
</tr>
</tbody>
</table>

- We evaluate the portability of the proposed framework on ACE05-CN (Chinese) and ERE-ES (Spanish).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training</th>
<th>Entity</th>
<th>Relation</th>
<th>Trigger</th>
<th>Argument</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE05-CN</td>
<td>CN</td>
<td>88.5</td>
<td>62.4</td>
<td>65.6</td>
<td>52.0</td>
</tr>
<tr>
<td></td>
<td>CN+EN</td>
<td>89.8</td>
<td>62.9</td>
<td>67.7</td>
<td>53.2</td>
</tr>
<tr>
<td>ERE-ES</td>
<td>ES</td>
<td>81.3</td>
<td>48.1</td>
<td>56.8</td>
<td>40.3</td>
</tr>
<tr>
<td></td>
<td>ES+EN</td>
<td>81.8</td>
<td>52.9</td>
<td>59.1</td>
<td>42.3</td>
</tr>
</tbody>
</table>
Extending from Sentence-Level to Document-Level

- **Multi-Sentence Argument Linking (Ebner et al., 2020)**

When Russian aircraft bombed a remote garrison in southeastern Syria last month, alarm bells sounded at the Pentagon and the Ministry of Defense in London.

The Russians weren’t **bombarding** a run-of-the-mill rebel outpost, according to U.S. officials.

- **Roles are evoked by event triggers, forming implicit arguments**

- **Implicit arguments linked to explicit mentions in text**
  - Representations: Learn span representations for each trigger span and candidate argument span
  - Prune: For each trigger, prune to top-K candidate arguments
  - Linking score: Score representations of implicit arguments with representations of explicit arguments using a decomposable scoring function
Event Extraction by Answering (Almost) Natural Questions (Du and Cardie, 2020)

The input sequences for the two QA models share a standard BERT-style format

[CLS] <question> [SEP] <sentence> [SEP]
Event Extraction

- Supervised Event Extraction
  - Schema-guided Event Extraction
  - Document-level Event Extraction
- Cross-domain Zero-shot Transfer for Event Extraction
- Cross-lingual Transfer for Multi-lingual Event Extraction
- Cross-media Structured Common Space for Multimedia Event Extraction
**Hypothesis**: Both event mentions and types have rich semantics and structures, which can specify their consistency and connections.
Corporate sponsors contributed cash, mattresses, rice to reach remote Orang Asli villages.
How Much Human Effort Can We Save?

Achieved comparable performance as a supervised system when it’s trained on 500 event mentions from 3000 sentences.

- Target Event Ontology: ACE(33 types) + FrameNet (1161 frames) = 1194 types
- Seen types for training: 10 most popular ACE types
- Unseen type: 23 remaining ACE types
Label-aware Classification (Zhang et al., 2020)

- **Conflicts:Attack**
  - **Attacker**
    - PER, GPE, ...
  - **Target**
    - PER, FAC, ...
  - **Place**
    - GPR, FAC, LOC, ...

**Label semantics**
- We select “attack” as the label because we assume that it can represent the overall meaning of this event type.

**Constraints**
- “Attacker” can only be the argument of “Conflict:Attack” rather than “Life:Marry”.

Use a cluster of contextualized embeddings to represent labels and use constraints to regularize the predictions by modeling it as an ILP problem.
The Proposed Framework

Event Types

\( E_1 \): Conflict:Attack
\( E_2 \): Life:Marry

Anchor Words

\( E_1 \): [Attack]
\( E_2 \): [marry, wed]

Anchor Sentences

\( E_1 \): \[s_1^{E_1}, s_2^{E_1}, s_3^{E_1}, \ldots\]
\( E_2 \): \[s_1^{E_2}, s_2^{E_2}, s_3^{E_2}, \ldots\]

Contextualized Representations

\( E_1 \): \[\mathbf{v}_1^{E_1}, \mathbf{v}_2^{E_1}, \mathbf{v}_3^{E_1}, \ldots\]
\( E_2 \): \[\mathbf{v}_1^{E_2}, \mathbf{v}_2^{E_2}, \mathbf{v}_3^{E_2}, \ldots\]

External Corpus

Role Types

\( R_1 \): Attacker
\( R_2 \): Money

Anchor Words

\( R_1 \): [attacker]
\( R_2 \): [money]

Anchor Sentences

\( R_1 \): \[s_1^{R_1}, s_2^{R_1}, s_3^{R_1}, \ldots\]
\( R_2 \): \[s_1^{R_2}, s_2^{R_2}, s_3^{R_2}, \ldots\]

Contextualized Representations

\( R_1 \): \[\mathbf{v}_1^{R_1}, \mathbf{v}_2^{R_1}, \mathbf{v}_3^{R_1}, \ldots\]
\( R_2 \): \[\mathbf{v}_1^{R_2}, \mathbf{v}_2^{R_2}, \mathbf{v}_3^{R_2}, \ldots\]

Constraints associated with the event ontology

Offline Preparation

Online Prediction

Bob [hits] John’s head.
How many anchor sentences do we need?

<table>
<thead>
<tr>
<th>Model</th>
<th>Train types</th>
<th>Test types</th>
<th>Trig Hit@1</th>
<th>Trig Hit@3</th>
<th>Trig Hit@5</th>
<th>Arg Hit@1</th>
<th>Arg Hit@3</th>
<th>Arg Hit@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>0</td>
<td>23</td>
<td>9.6</td>
<td>27.2</td>
<td>42.5</td>
<td>25.9</td>
<td>63.4</td>
<td>80.6</td>
</tr>
<tr>
<td>WSD</td>
<td>0</td>
<td>23</td>
<td>1.7</td>
<td>13.0</td>
<td>22.8</td>
<td>2.4</td>
<td>2.8</td>
<td>2.8</td>
</tr>
<tr>
<td>Transfer-learning (A)</td>
<td>1</td>
<td>23</td>
<td>4.0</td>
<td>23.8</td>
<td>32.5</td>
<td>1.3</td>
<td>3.4</td>
<td>3.6</td>
</tr>
<tr>
<td>Transfer-learning (B)</td>
<td>3</td>
<td>23</td>
<td>7.0</td>
<td>12.5</td>
<td>36.8</td>
<td>3.5</td>
<td>6.0</td>
<td>6.3</td>
</tr>
<tr>
<td>Transfer-learning (C)</td>
<td>5</td>
<td>23</td>
<td>20.1</td>
<td>34.7</td>
<td>46.5</td>
<td>9.6</td>
<td>14.7</td>
<td>15.7</td>
</tr>
<tr>
<td>Transfer-learning (D)</td>
<td>10</td>
<td>23</td>
<td>33.5</td>
<td>51.4</td>
<td>68.3</td>
<td>14.7</td>
<td>26.5</td>
<td>27.7</td>
</tr>
<tr>
<td>Our Approach</td>
<td>0</td>
<td>23</td>
<td><strong>80.5</strong></td>
<td><strong>88.9</strong></td>
<td><strong>93.2</strong></td>
<td><strong>68.5</strong></td>
<td><strong>94.2</strong></td>
<td><strong>96.8</strong></td>
</tr>
<tr>
<td>Frequency</td>
<td>0</td>
<td>33</td>
<td>28.9</td>
<td>53.6</td>
<td>62.7</td>
<td>13.8</td>
<td>33.8</td>
<td>51.0</td>
</tr>
<tr>
<td>Our Approach</td>
<td>0</td>
<td>33</td>
<td><strong>82.9</strong></td>
<td><strong>93.1</strong></td>
<td><strong>96.2</strong></td>
<td><strong>53.6</strong></td>
<td><strong>87.9</strong></td>
<td><strong>92.4</strong></td>
</tr>
</tbody>
</table>

Ten sentences are good enough!!
Event Extraction

- Supervised Event Extraction
  - Schema-guided Event Extraction
  - Document-level Event Extraction

- Cross-domain Zero-shot Transfer for Event Extraction
  - Cross-lingual Transfer for Multi-lingual Event Extraction

- Cross-media Structured Common Space for Multimedia Event Extraction
Cross-lingual Joint Entity and Word Embedding Learning

- Cross-lingual Joint Entity and Word Embedding to Improve Entity Linking and Parallel Sentence Mining (Pan et al., 2019)

- Code-switch cross-lingual entity/word data generation

Use English entities as anchor points to learn a mapping (rotation matrix) $W$ which aligns distributions in IL and English

Example Chinese Wikipedia Sentence:

[[小米科技 | 小米]] 被誉为中国的 [[苹果公司 | 苹果]]。

Our Approach:

zh/小米科技 被誉为 苹果公司的 en/Apple_Inc.

(Xiaomi) (is) (known as) (Chinese)

- Use English entities as anchor points to learn a mapping (rotation matrix) $W$ which aligns distributions in IL and English
Cross-lingual Structure Transfer for Relation and Event Extraction (Subburathinam et al., 2019)
Graph Convolutional Networks (GCN) Encoder

- Extend the monolingual design (Zhang et al., 2018) to cross-lingual
  - Convert a sentence with N tokens into N*N adjacency matrix $A$
  - Node: token, each edge is a directed dependency edge
- Initialization of each node’s representation

$$h_i^{(0)} = x_i^w \oplus x_i^p \oplus x_i^d \oplus x_i^e$$

- Word embedding  POS tag  Dependency relation  Entity type
- At the $k^{th}$ layer, derive the hidden representation of each node from the representations of its neighbors at previous layer

$$h_i^{(k)} = \text{ReLU} \left( \sum_{j=0}^{N} \frac{A_{ij} W^{(k)} h_j^{(k-1)}}{d_i + b^{(k)}} \right)$$
Application on Event Argument Extraction

- Task: Classify each pair of event trigger and entity mentions into one of pre-defined event argument roles or NONE
- Max-pooling over the final node representations to obtain representations for sentence, trigger and argument candidate, and concatenate them
- A softmax output layer for argument role labeling

\[
L^a = \sum_{i=1}^{N} \sum_{j=1}^{L_i} y_{ij} \log(\sigma(\mathbf{U}^a \cdot [h_i^t; h_{ij}^s; h_j^a]))
\]
Cross-lingual Event Transfer Performance

- Chinese Event Argument Extraction (Subburathinam et al., EMNLP2019)
Event Extraction

- Supervised Event Extraction
  - Schema-guided Event Extraction
  - Document-level Event Extraction
- Cross-domain Zero-shot Transfer for Event Extraction
- Cross-lingual Transfer for Multi-lingual Event Extraction
  - Cross-media Structured Common Space for Multimedia Event Extraction
Knowledge is Beyond Just Text

- Multimedia Event Extraction (Li et al., ACL2020)
- We produce and consume news content through multimedia, 33% of news images contain event arguments not mentioned in surrounding texts

TransportPerson_Instrument = stretcher
Last week, U.S. Secretary of State Rex Tillerson visited Ankara, the first senior administration official to visit Turkey, to try to seal a deal about the battle for Raqqa and to overcome President Recep Tayyip Erdogan’s strong objections to Washington's backing of the Kurdish Democratic Union Party (PYD) militias. Turkish forces have attacked SDF forces in the past around Manbij, west of Raqqa, forcing the United States to *deploy* dozens of soldiers on the outskirts of the town in a mission to prevent a repeat of clashes, which risk derailing an assault on Raqqa.
Last week, U.S. Secretary of State Rex Tillerson visited Ankara, the first senior administration official to visit Turkey, to try to seal a deal about the battle for Raqqa and to overcome President Recep Tayyip Erdogan's strong objections to Washington's backing of the Kurdish Democratic Union Party (PYD) militias. Turkish forces have attacked SDF forces in the past around Manbij, west of Raqqa, forcing the United States to deploy dozens of soldiers on the outskirts of the town in a mission to prevent a repeat of clashes, which risk derailing an assault on Raqqa.

Output: Multimedia Events & Argument Roles

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Movement.Transport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>Text Trigger deploy</td>
</tr>
<tr>
<td>Image</td>
<td></td>
</tr>
</tbody>
</table>

Arguments

- Agent: United States
- Destination: outskirts
- Artifact: soldiers
- Vehicle

Vehicle
Last week, U.S. Secretary of State Rex Tillerson visited Ankara, the first senior administration official to visit Turkey, to try to seal a deal about the battle for Raqqa and to overcome President Recep Tayyip Erdogan’s strong objections to Washington's backing of the Kurdish Democratic Union Party (PYD) militias. Turkish forces have attacked SDF forces in the past around Manbij, west of Raqqa, forcing the United States to deploy dozens of soldiers on the outskirts of the town in a mission to prevent a repeat of clashes, which risk derailing an assault on Raqqa.

Output: Multimedia Events & Argument Roles

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Movement.Transport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>Text Trigger</td>
</tr>
<tr>
<td></td>
<td>deploy</td>
</tr>
<tr>
<td>Image</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Arguments</th>
<th>Agent</th>
<th>Destination</th>
<th>Artifact</th>
<th>Vehicle</th>
<th>Vehicle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>United States</td>
<td>outskirts</td>
<td>soldiers</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Vision vs. NLP for Event Extraction

- Vision does not study newsworthy, complex events
  - Focusing on daily life and sports (Perera et al., 2012; Chang et al., 2016; Zhang et al., 2007; Ma et al., 2017)
  - Without localizing a complete set of arguments for each event (Gu et al., 2018; Li et al., 2018; Duarte et al., 2018; Sigurdsson et al., 2016; Kato et al., 2018; Wu et al., 2019a)

- Most related: Situation Recognition (Yatskar et al., 2016)
  - Classify an image as one of 500+ FrameNet verbs
  - Identify 192 generic semantic roles via a 1-word description
Treat Image/Video as a foreign language

<table>
<thead>
<tr>
<th>Text</th>
<th>Image / Video Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Image Region</td>
</tr>
<tr>
<td>Entity</td>
<td>Visual Object</td>
</tr>
<tr>
<td>Relation</td>
<td>Visual Relation</td>
</tr>
<tr>
<td>Entity-Relation Graph</td>
<td>Visual Scene Graph</td>
</tr>
<tr>
<td>Event Trigger</td>
<td>Visual Activity</td>
</tr>
<tr>
<td><strong>Linguistic Structure</strong></td>
<td><strong>Situation Graph</strong></td>
</tr>
</tbody>
</table>
Cross-media Structured Common Space

- Treat Image/Video as a foreign language
  - Represent it with a structure that is similar to AMR graph in text

<table>
<thead>
<tr>
<th>Caption</th>
<th>AMR Graph</th>
<th>Image</th>
<th>Situation Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thai opposition protesters [Attacker] attack [Conflict.Attack] a bus [Target] carrying pro-government Red Shirt supporters on their way to a rally at a stadium in Bangkok [Place].</td>
<td><img src="image" alt="AMR Graph" /></td>
<td><img src="image" alt="Image" /></td>
<td><img src="image" alt="Situation Graph" /></td>
</tr>
</tbody>
</table>

Linguistic Structure, e.g., Dependency Tree

Abstract Meaning Representation (AMR)
Weakly Aligned Structured Embedding (WASE)

-- Training Phase (Common Space Construction)

Caption
Weakly Aligned Structured Embedding (WASE)

-- Training Phase (Common Space Construction)
How to generate situation graph?

- Method 1: Object-based Graph Training
  - Learn to project image to verb embedding, and object to noun
  - Learn to classify each object-image pair to a semantic role
How to generate situation graph?

- Method 2: Role-driven Attention Graph
  - Learn to project image embedding to verb embedding
  - Learn a spatial attention on image for each role
  - Learn to project attended role region to noun embedding
Weakly Aligned Structured Embedding (WASE)

-- Training Phase (Common Space Construction)
How to align the two modalities?

- Prior work aligns image-caption vectors by triplet loss.
- We want to align two graphs, not just single vectors.
- Ontology is shared so the nodes carry similar semantics.
How to align the two modalities?

- Prior work aligns image-caption vectors by triplet loss.
- We want to align two graphs, not just single vectors.
- Ontology is shared so the nodes carry similar semantics.
Weakly Aligned Structured Embedding (WASE)

-- Training and Testing Phase (Cross-media shared classifiers)

Training Phase

**ACE Text Event**
Liana Owen [Participant] drove from Pennsylvania to **attend** [Contact.Meet] the rally in Manhattan with her parents [Participant].

**Alignment**

VOA Image-Caption Pairs

**imSitu Image Event**

destroying [Conflict.Attack]

Item [Target]: ship

Tool [Instrument]: bomb

Testing Phase

**Multimedia News**
For the rebels, bravado goes hand-in-hand with the desperate resistance the insurgents have mounted.....

Cross-media Structured Common Representation Encoder

Cross-media Shared Event Classifier

Contact.Meet  Conflict.Attack

Cross-media Shared Argument Classifier

Contact.Meet  Participant

Conflict.Attack  Instrument
Weakly Aligned Structured Embedding (WASE)

-- Training and Testing Phase (Cross-media shared classifiers)
Weakly Aligned Structured Embedding (WASE)

-- Training and Testing Phase (Cross-media shared classifiers)

Training Phase

ACE Text Event
Liana Owen [Participant] drove from Pennsylvania to attend [Contact.Meet] the rally in Manhattan with her parents [Participant].

Alignment
VOA Image-Caption Pairs

imSitu Image Event
destroying [Conflict.Attack]

Item [Target]: ship
Tool [Instrument]: bomb

Cross-media Structured Common Representation Encoder

Testing Phase

Multimedia News
For the rebels, bravado goes hand-in-hand with the desperate resistance the insurgents have mounted.....

Cross-media Shared Event Classifier
Contact.Meet Conflict.Attack Conflict.Attack

Cross-media Shared Argument Classifier
Weakly Aligned Structured Embedding (WASE)

-- Training and Testing Phase (Cross-media shared classifiers)

**Training Phase**
- ACE Text Event
  - Liana Owen [Participant]
  - drove from Pennsylvania to attend [Contact.Meet] the rally in Manhattan with her parents [Participant].

- Alignment
  - VOA Image-Caption Pairs

- imSitu Image Event
  - destroying [Conflict.Attack]
  - Item [Target]: ship
  - Tool [Instrument]: bomb

**Testing Phase**
- Multimedia News
  - For the rebels, bravado goes hand-in-hand with the desperate resistance the insurgents have mounted....

---

**Cross-media Structured Common Representation Encoder**

**Cross-media Shared Event Classifier**
- Contact.Meet
- Conflict.Attack

**Cross-media Shared Argument Classifier**
- Conflict.Attack Attacker
- Conflict.Attack Instrument
Weakly Aligned Structured Embedding (WASE)

-- Training and Testing Phase (Cross-media shared classifiers)
Weakly Aligned Structured Embedding (WASE)

-- Training and Testing Phase (Cross-media shared classifiers)
Weakly Aligned Structured Embedding (WASE)

-- System Diagram

Multimedia Document

Cross-media Event Coreference

WASE

Event1

Arg1

Arg2

Arg3

Event2

Arg4

Event3

Arg5
<table>
<thead>
<tr>
<th>Training Text</th>
<th>Model</th>
<th>Text-Only Evaluation</th>
<th>Image-Only Evaluation</th>
<th>Multimedia Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Event Mention</td>
<td>Argument Role</td>
<td>Event Mention</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$P$</td>
<td>$R$</td>
<td>$F_1$</td>
</tr>
<tr>
<td></td>
<td>JMEE</td>
<td>42.5</td>
<td>58.2</td>
<td>48.7</td>
</tr>
<tr>
<td></td>
<td>GAIL</td>
<td>43.4</td>
<td>53.5</td>
<td>47.9</td>
</tr>
<tr>
<td></td>
<td>WASE$^T$</td>
<td>42.3</td>
<td>58.4</td>
<td>48.2</td>
</tr>
<tr>
<td></td>
<td>WASE$^T_{\text{att}}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WASE$^T_{\text{obj}}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>VSE-C</td>
<td>33.5</td>
<td>47.8</td>
<td>39.4</td>
</tr>
<tr>
<td></td>
<td>Flat$^T_{\text{att}}$</td>
<td>34.2</td>
<td>63.2</td>
<td>44.4</td>
</tr>
<tr>
<td></td>
<td>Flat$^T_{\text{obj}}$</td>
<td>38.3</td>
<td>57.9</td>
<td>46.1</td>
</tr>
<tr>
<td></td>
<td>WASE$^T_{\text{att}}$</td>
<td>37.6</td>
<td>66.8</td>
<td>48.1</td>
</tr>
<tr>
<td></td>
<td>WASE$^T_{\text{obj}}$</td>
<td>42.8</td>
<td>61.9</td>
<td>50.6</td>
</tr>
</tbody>
</table>

**Note:** The highlighted values indicate the best performance for each category.
## Experiment Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Event Mention</td>
<td>Argument Role</td>
<td>Event Mention</td>
<td>Argument Role</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F₁</td>
<td>P</td>
</tr>
<tr>
<td>JMEEM</td>
<td>42.5</td>
<td>58.2</td>
<td>48.7</td>
<td>22.9</td>
</tr>
<tr>
<td>GAIL</td>
<td>43.4</td>
<td>53.5</td>
<td>47.9</td>
<td>23.6</td>
</tr>
<tr>
<td>WASE att</td>
<td>42.3</td>
<td>58.4</td>
<td>48.2</td>
<td>21.4</td>
</tr>
<tr>
<td>WASE obj</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VSE-C</td>
<td>33.5</td>
<td>47.8</td>
<td>39.4</td>
<td>16.6</td>
</tr>
<tr>
<td>Flat att</td>
<td>34.2</td>
<td>63.2</td>
<td>44.4</td>
<td>20.1</td>
</tr>
<tr>
<td>Flat obj</td>
<td>38.3</td>
<td>57.9</td>
<td>46.1</td>
<td>21.8</td>
</tr>
<tr>
<td>WASE att</td>
<td>37.6</td>
<td>66.8</td>
<td>48.1</td>
<td>27.5</td>
</tr>
<tr>
<td>WASE obj</td>
<td>42.8</td>
<td>61.9</td>
<td>50.6</td>
<td>23.5</td>
</tr>
</tbody>
</table>

The table above shows the evaluation results for different models across different modalities: Text, Image, and Multimedia. The results are presented for three different metrics: Precision (P), Recall (R), and F1-score (F₁).
## Experiment Results

<table>
<thead>
<tr>
<th>Training Text</th>
<th>Model</th>
<th>Text-Only Evaluation</th>
<th>Image-Only Evaluation</th>
<th>Multimedia Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Event Mention</td>
<td>Argument Role</td>
<td>Event Mention</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$P$  $R$  $F_1$</td>
<td>$P$  $R$  $F_1$</td>
<td>$P$  $R$  $F_1$</td>
</tr>
<tr>
<td>VSE-C</td>
<td>33.5</td>
<td>47.8       39.4</td>
<td>16.6      24.7   19.8</td>
<td>30.3      48.9   26.4</td>
</tr>
<tr>
<td>Flatatt</td>
<td>34.2</td>
<td>63.2       44.4</td>
<td>20.1      27.1   23.1</td>
<td>27.1      57.3   36.7</td>
</tr>
<tr>
<td>Flatobj</td>
<td>38.3</td>
<td>57.9       46.1</td>
<td>21.8      26.6   24.0</td>
<td>26.4      55.8   35.8</td>
</tr>
<tr>
<td>WASE_att</td>
<td>37.6</td>
<td>66.8       48.1</td>
<td>27.5      33.2   30.1</td>
<td>32.3      63.4   42.8</td>
</tr>
<tr>
<td>WASE_obj</td>
<td>42.8</td>
<td>61.9       <strong>50.6</strong></td>
<td>23.5      30.3   26.4</td>
<td>43.1      59.2   <strong>49.9</strong></td>
</tr>
</tbody>
</table>
## Cross-Media Coreference Accuracy

<table>
<thead>
<tr>
<th>Model</th>
<th>$P$ (%)</th>
<th>$R$ (%)</th>
<th>$F_1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rule_based</td>
<td>10.1</td>
<td>100</td>
<td>18.2</td>
</tr>
<tr>
<td>VSE</td>
<td>31.2</td>
<td>74.5</td>
<td>44.0</td>
</tr>
<tr>
<td>Flat\textsubscript{att}</td>
<td>33.1</td>
<td>73.5</td>
<td>45.6</td>
</tr>
<tr>
<td>Flat\textsubscript{obj}</td>
<td>34.3</td>
<td>76.4</td>
<td>47.3</td>
</tr>
<tr>
<td>WASE\textsubscript{att}</td>
<td>39.5</td>
<td>73.5</td>
<td>51.5</td>
</tr>
<tr>
<td>WASE\textsubscript{obj}</td>
<td>40.1</td>
<td>75.4</td>
<td>52.4</td>
</tr>
</tbody>
</table>
Compare to Single Data Modality Extraction

- Surrounding sentence helps visual event extraction.

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

- Image helps textual event extraction.

  Iraqi security forces search [Justice.Arrest] a civilian in the city of Mosul.
Why Does Vision Help NLP?

- Various triggers and context can be coherent in visual space.
- Cross-media Common space pushes scattered sentences towards the visual cluster.

Berlin police tweeted that six people were arrested after a joint operation with the Berlin's prosecutor's office.

He was asleep in a suburban Seattle house last week morning when immigration agents showed up to arrest his father.

The man in Kosovo is an ethnic Albanian arrested south of the capital, Pristina.

But shortly after the round table began, Marko Djuric, head of the Serbian government office on Kosovo, was detained by police.
Compare to Cross-media Flat Representation

<table>
<thead>
<tr>
<th>Model</th>
<th>Event Type</th>
<th>Argument Role</th>
<th>Argument Role Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat</td>
<td>Justice.ArrestJail</td>
<td>Agent =</td>
<td>man</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>Justice.ArrestJail</td>
<td>Entity =</td>
<td>man</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Event Type</th>
<th>Argument Role</th>
<th>Argument Role Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat</td>
<td>Movement.Transport</td>
<td>Artifact =</td>
<td>none</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>Movement.Transport</td>
<td>Artifact =</td>
<td>man</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Summary of Event Extraction Methods

<table>
<thead>
<tr>
<th>IE Methods</th>
<th>Supervised Learning</th>
<th>Bootstrapping</th>
<th>Distant Supervision</th>
<th>Open IE/Zero-shot</th>
<th>Schema/Discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach Overview</td>
<td>Learn rules or supervised model from labeled data</td>
<td>Send seeds to extract common patterns from unlabeled data</td>
<td>Project large database entries into unlabeled data to obtain annotations</td>
<td>Open-domain IE based on syntactic patterns</td>
<td>Automatically discover scenarios, event types and templates</td>
</tr>
<tr>
<td>Requirement of labeled data</td>
<td>Large unstructured labeled data</td>
<td>Small seeds</td>
<td>Large seeds</td>
<td>Small unstructured labeled data</td>
<td>Little labeled data</td>
</tr>
<tr>
<td>Quality</td>
<td>Precision: High, Recall: High</td>
<td>Precision: Moderate, Recall: Difficult to measure</td>
<td>Precision: Low, Recall: Moderate</td>
<td>Precision: Moderate</td>
<td>Precision: Moderate</td>
</tr>
<tr>
<td>Portability</td>
<td>Poor</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Scalability</td>
<td>Poor</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Good</td>
<td>Good</td>
</tr>
</tbody>
</table>