1

Low-resource Event Extraction via Share-and-Transfer and Remaining Challenges

Heng Ji[a] and Clare Voss

Abstract

Event Extraction aims to find *who did what to whom, when* and *where* from unstructured data. Over the past decade, research in event extraction has made advances in three waves. The first wave relied on supervised machine learning models trained from a large amount of manually annotated data and manually crafted features. The second wave eliminated this method of feature engineering by introducing deep neural networks with distributional semantic embedding features, but still required large annotated datasets. This chapter provides an overview of a third wave with a *share-and-transfer* framework, that further enhances the portability of event extraction by transferring knowledge from a high-resource setting to another low-resource setting, reducing the need there for annotated data. We describe three low-resource settings: a new domain, a new language, or a new data modality. The first *share* step of our approach is to construct a *common structured semantic representation space* into which these complex structures can be encoded. Then, in the *transfer* step of the approach, we can train event extractors over these representations in high-resource settings and apply the learned extractors to target data in the low-resource setting. We conclude the

[a] Supported by ARL NS-CTA No. W911NF-09-2-0053, DARPA KAIROS Program # FA8750-19-2-1004, U.S. DARPA LORELEI Program # HR0011-15-C-0115, U.S. DARPA AIDA Program # FA8750-18-2-0014, Air Force No. FA8650-17-C-7715, and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract # FA8650-17-C-9116. 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, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

chapter with a summary of the current status of this new framework and point to remaining challenges and future research directions to address them.

1.1 Introduction

An event is a specific incident or situation that can be described as indicating, *who did what to whom, when* and *where.* Understanding events and communicating about them to other people are both fundamental human activities. However, it is typically more challenging to remember event-related information than entity-related information. For example, most people in the United States will know the named entity that is the answer to the question "Who was the president of United States in 2010?", but very few people would be able to recall all the event information needed to answer the question "Who died from coronavirus?" Event extraction is the task that outputs essential elements of information about events in support of downstream applications, such as question-answering and timeline-based summaries.

The task of event extraction, as originally defined in natural language processing research, entails first identifying *event triggers* (the words or phrases that most clearly express event occurrences) and their *arguments* (the words or phrases for participants in those events) in unstructured texts and then classifying these phrases, respectively, for their types and roles. Situated at the end of the Information Extraction (IE) pipeline, event extraction represents the most complex component task. Recognizing the different forms in which an event may be expressed, distinguishing events of different types, and finding the arguments of an event are all challenging aspects of this task. Traditional event extraction techniques have focused on extracting events for only a limited set of pre-defined types from English text documents. However, users of event extraction systems now want to analyze event information from a much wider variety of sources, across:
(1) *multiple domains,* where events may be tracked on different time scales and expressed at different levels of specificity or abstraction, while providing users nonetheless with additional essential elements of information for reasoning about both old (known) and new scenarios;
(2) *multiple languages,* where certain relations and events of primary interest to a given community are reported predominantly in the low-resource language found in sources available to that community; and

(3) *multiple media,* where the content in visual modalities (such as images, videos) may reveal activities or identities of event participants not explicitly described in another modality (such as text or speech).

The most successful approaches to event extraction, originating in the first wave of Information Extraction (IE) research, have been based on supervised learning with hand-crafted symbolic features (Grishman et al., 2005; Ahn, 2006; Ji and Grishman, 2008; Liao and Grishman, 2011; Huang and Riloff, 2012; Poon and Vanderwende, 2010; Riedel et al., 2009; Li et al., 2013; Venugopal et al., 2014; Li et al., 2014; Yang and Mitchell, 2016). More recently in the second wave of IE, the systems with best results have been those augmented with distributional embedding features (Chen et al., 2015b; Nguyen and Grishman, 2015, 2016; Nguyen et al., 2016; Nguyen and Grishman, 2018; Liu et al., 2018; Hong et al., 2018; Chen et al., 2018).

These approaches however incorporate domain-specific and language-specific information into their models, and thus are costly to port beyond their original settings; they require substantial amounts of new annotations for retraining to a new domain or a new language. Compared to other tasks within Information Extraction (IE) pipelines, such as Name Tagging, the annotations needed for retraining models for Relation and Event Extraction are more costly; they require both structured data annotations and a rich label space. Publicly available, gold-standard annotations for event extraction exist for only a few languages and a limited number of event types (e.g., 33 event types defined in ACE[1] for English, Chinese, and Arabic). As a result, event extraction is substantially more challenging when conducted in under-resourced settings with little or no annotated data. In this chapter, we examine three such settings: a new domain, a new language, and a new data modality (image in this chapter) respectively.

In the sections to follow, we present an overview of a new event extraction paradigm based on transferable neural network learning techniques that leverage existing, manually-constructed schemas and annotations for a small set of seen[2] information concepts (e.g., types). Modeled as a bottom-up discovery problem, the idea is to "share-and-transfer": first, by combining symbolic and distributional embedding representations of the extracted information derived from source training data and encoding these via deep neural networks into one shared continuous semantic space, and then, by learning the extraction models over this space from

[1] ACE was the Automatic Content Extraction Program (Doddington et al., 2004)
[2] Here *seen* means seen during training.

source data and applying the learned extractors to target data. In this way, knowledge of how to extract the event structures with their types and argument roles from a high-resource setting is transferred, i.e., becomes available for recognizing unseen content in a low-resource setting.

1.2 Approach Overview

One of the challenges of traditional machine learning has been that models built for one task or domain typically suffer with degraded performance when later applied to new distinct tasks or domains. The models, while perhaps highly accurate within their training (source) domain or task, nonetheless require rebuilding from scratch when the relevant feature space distribution for new (target) data changes.

The objective of our "share-and-transfer" approach is to address the cross-domain, cross-lingual, and cross-modal challenges of porting the development of an event extractor from its source setting (domain, language, or modality) over which it was trained with plenty of annotated data, to a new target setting (domain, language, or modality) for which there is little or no annotated data.

Informally, we describe our approach as leveraging annotated source data from a high-resourced setting, to enable building models for a low-resourced or no-resource target setting. For the share phase, we start by identifying the range of needed source-side data that, once transformed, will serve as the basis for a structured semantic space that will be common to, i.e., be a shared space for, both the source and target representations. We use the original, ample source-side data and its transformed representations in the common space to train the event extractors. By virtue of having (i) constructed the shared space so that it would hold both source and target structures and (ii) trained the extractors over that space, we then run the grounding or transfer phase, (iii) applying these extractors to the target data, once transformed, to obtain the events on the target side.

For each of the settings, the key questions pertaining to the construction of such a common semantic space and the transfer, are:

1. which types of source data and other resources are used in constructing the common space?
2. what types of structured semantic information are used in training for grounding or transfer by way of the common space?

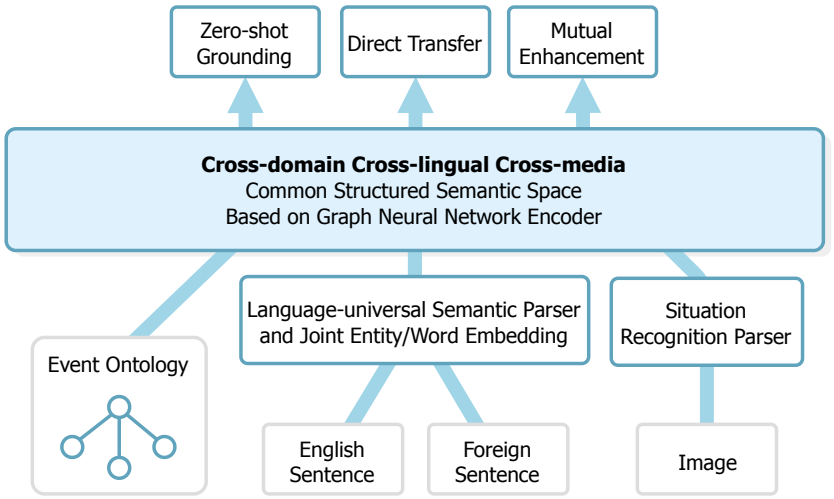

Figure 1.1 Share-and-Transfer Approach Overview

3. how is target data processed for event extraction?

In answer to 1. above, our approach to representation learning for extraction tasks in low-resource settings starts with data sources of structured knowledge that are already or that can be transformed into graph structures, and all of which are either available or can be automatically derived by computational tools in high-resource settings, from: (a) semantic or conceptual networks, such as ontologies, (b) linguistic representations of text data, such as semantic parse trees, and (c) recent, novel computational representations of image data, such as situation graphs.

Then these sources, in answer to 2. above—as features and graph structures that capture structured semantic information from an ontology or parser output on text or images—are encoded into low-dimensional embeddings by way of techniques based on deep learning and nonlinear dimensionality reduction. We describe the ways in which recent methods for embedding individual nodes and (sub)graphs within these sources are transformed as input to shared CNNs or graph-based convolutional neural networks (GCNs) for learning transfer functions. The learned functions preserve their semantic information in constructing a common structured semantic space for each type of transfer setting we consider.

In answer to 3. above, the target datasets (text or image) from the low-resource setting are transformed, by way of embeddings and neu-

ral network encoders, into their vector representations situated in the common semantic space.

The next section provides the descriptions of the transformations of text-based multidomain and multilingual source-side data, respectively, in sections 1.3.1 and 1.3.2. Following these descriptions, the section shifts to detailing the more complex transformations of multimodal data in section 1.3.3. The transformations for the construction of common semantic spaces in each of these three settings take place in advance of the transfer phase that will be described in section 1.4.

1.3 Share: Construction of Common Semantic Space

Traditional methods of processing and understanding unstructured texts have mainly relied on symbolic semantics. For example, to extract the "attacker" of a military attack event from a text passage, there needs to be a way to feed the series of symbolic representations for the words and their relations in each sentence into a knowledge network. The sentence might include a verb that indicates the attack event, a noun phrase in subject position for the candidate attacker entity that may be a person, etc. The symbolic feature engineering to automate the process of determining which words or phrases in each sentence will fill which part or parts of an event structure is costly; it requires substantial linguistic expertise and time.

In recent years, the incorporation of distributional semantic representations in NLP tasks has been a notable game changer. The inspiration for distributional semantics comes from an observation made long ago, now referred to as the distributional hypothesis (Harris, 1954): linguistic items with similar distributions have similar meanings. This hypothesis, together with the advent of dense-vector embeddings for encoding words into a common vector space, has been so successful in monolingual tasks that this combination has led to further research to determine how such spaces might also be used for other applications. One such example has been extracting information from unstructured texts in multiple languages by inducing cross-lingual word embeddings for two or more languages in the same vector space. The success of these applications derives from the mathematical relations between embeddings, specifically where constructed vectors for translations and words with similar meanings are geometrically close in the shared cross-lingual vector space. As described below, our approach takes this methodology further, general-

izing it so that *structured* content of semantic relations within events is preserved in extended embeddings of a shared vector space. That space, when constructed with embeddings for gold standard relation and event annotations from high-resource languages (e.g., English), then supports the training of extractors that can then subsequently be applied directly to low-resource languages. This shared common semantic space also provides much richer representations for words than monolingual embedding due to the lack of monolingual data in low-resource languages.

Previous applications have focused on flat, sequence-level (word or phrase) representations that are not adequate for event extraction. These have, in effect, neglected the value of linguistic signals that are available in gold standard structured annotations (such as ACE and ERE datasets) or that could be derived from linguistic analyses (such as syntactic and semantic parsed datasets). In contrast, we develop a novel multilingual, multimedia common semantic representation that incorporates structure-level, in addition to sequence-level information. In so doing, we treat the end-to-end event extraction in an IE pipeline as an Information Network construction problem over unstructured input texts. In the constructed information network, the extracted entity mentions and event triggers are network nodes, and the extracted relations and event-argument links are network edges labeled with their relation and argument roles, respectively. An information network can be considered as a special form of semantic network (Simmons, 1963; Geldenhuys et al., 1999; Do et al., 2017), where each node and edge belongs to one of a set of pre-defined types. Recent work on multilingual, multimedia common space construction makes use of linear mappings or canonical correlation analysis to transfer features or models across languages or data modalities. Unlike that work, our major innovation is to convert unstructured data into structured representations, as derived from universal semantic parses and enhanced with distributional information, so as to capture individual entities as well as the relations and events involving those entities, so that we can then share the resulting structured representations across multiple languages and data modalities. The structured representations can compress wide contexts in text and capture semantic relations between image regions for better disambiguation.

1.3.1 Multi-domain Common Space Construction

For the first example of building a common space, consider the various linguistic resources available to rapidly move an event extraction system

developed from known domains (e.g., *military action*) to new domains (e.g., *rescue*). Our approach is a novel and highly scalable, zero-shot grounding framework that makes use of a common structured space, as illustrated in Figure 1.2, for classifying event types and event arguments.

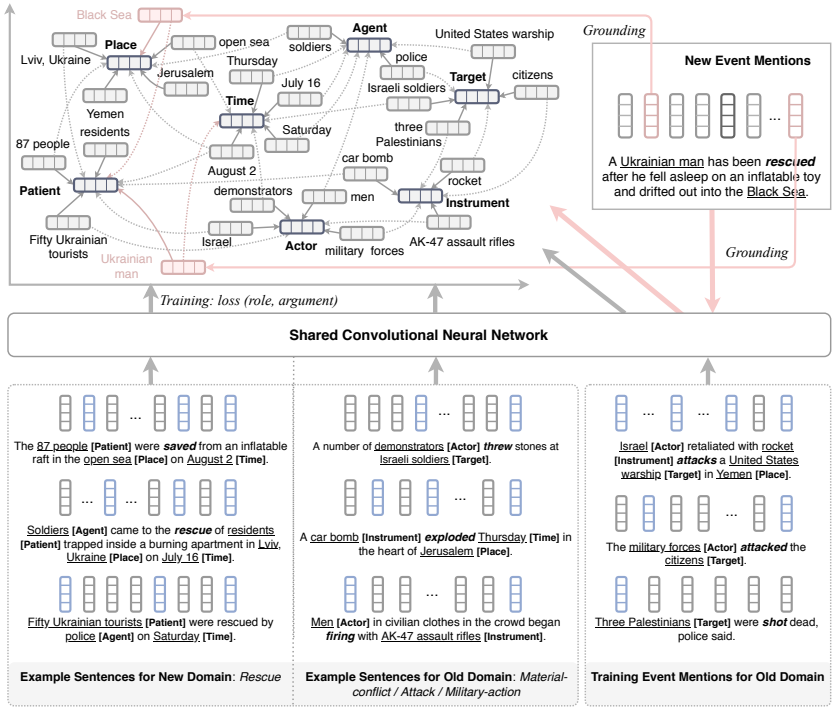

Figure 1.2 Multi-domain common space construction and zero-shot transfer learning.

At training time, the common space is constructed with vector outputs of a shared convolutional neural network (CNN) whose input is embeddings derived from two sets of sentences in the known, i.e, old domain (here, *Attack*) and from the typed event structures of an event ontology. As a result of training, the CNN, shown spanning the middle of the figure, outputs vectors as derived from processing of ontology elements (e.g., such as *Military Action/Actor*) and semantically-analyzed linguistic expressions (parsed fragments). These vectors collectively populate the same common space, shown in the upper left box of the figure. Note that, with this method of construction, all event types and ar-

gument roles that form the event structures in the ontology are now incorporated in the common space, even though many of these typed structures have not been seen before in the training sentences. It will be the common space representation of these unseen typed structures that will provide for the next phase of our approach, zero-shot grounding (as shown in figure 1.1), that results in cross-domain transfer, to extract events from sentences in a new, target domain (such as *Rescue*). We now defer till section 1.4 further discussion of that phase, so that we can introduce two more examples of the share phase where we construct different common structured spaces.

1.3.2 Multi-lingual Common Space Construction

Our next objective is to extend the common space beyond mono-lingual data to include representations from multiple languages for subsequent cross-lingual transfer learning of an event extractor from one language to new language. Support for this extension comes from recent research (Lin et al., 2017) that has found relational facts to be typically expressed by identifiable patterns *within* languages and demonstrated that the consistency in patterns observed *across* languages can be used to improve relation extraction. As shown in Figure 1.3, even for sentences in different languages with different meanings that include quite distinct events and entity mentions, we can see that the sentences' parsed structures can share similar language-universal symbolic features, such as common labeled dependency paths and part-of-speech (POS) tags, as well as distributional features from multilingual embeddings that are readily available for many languages. In particular, for these given parsed sentences with passivized verbs in both languages, the subjects (phrases and nodes colored pink) are both noun arguments of type PER (person) and the oblique objects (phrases and nodes colored blue) are both noun arguments of type FAC (facility).

Our approach to constructing a multilingual structured common space (shown as output layer above bar for encoder that spans the middle of Figure 1.3) consists of three steps: (1) Convert each sentence in any language into a language-universal tree structure based on universal dependency parsing, (2) For each node in the tree structure, create a representation from the concatenation of embeddings corresponding to the node's word, its language-universal POS tag, its dependency role within the sentence parse, and its entity type, so that all sentence nodes, independent of their language, can be encoded for uniform input at the next

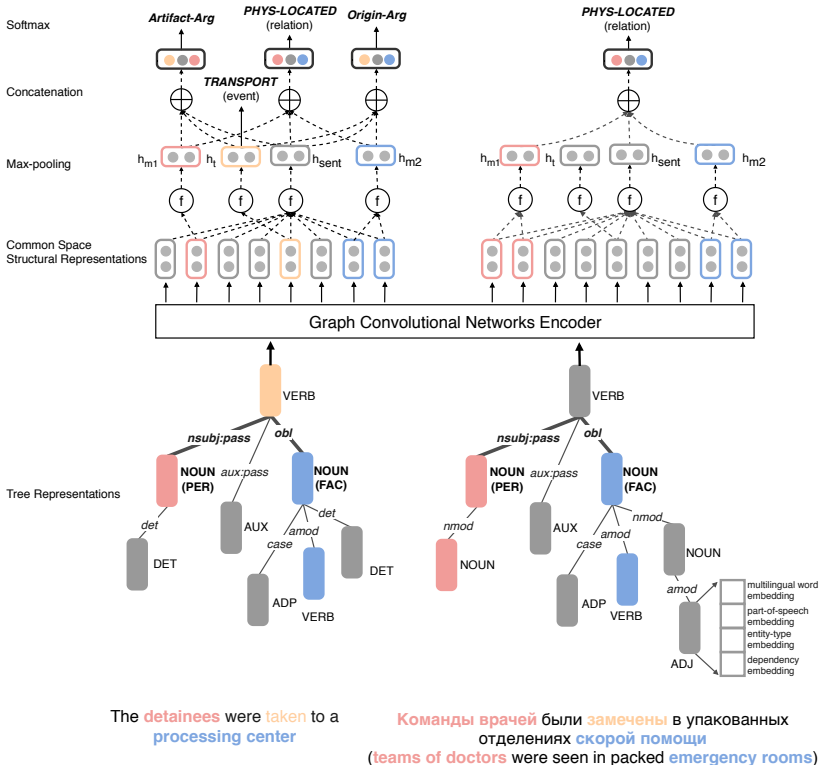

Figure 1.3 Multilingual common semantic space and cross-lingual structure transfer.

step (see lower righthand side of figure for four levels of token-specific embeddings), (3) Adopt graph convolutional networks (GCN) (Kipf and Welling, 2016) to generate contextualized node representations for the common space by leveraging information from each node's neighbors, as derived from the dependency parsing tree.

The graph representation that is obtained from dependency parsing of a sentence with N tokens, in (1) above, is converted into an $N \times N$ adjacency matrix $\boldsymbol{A}$, with added self-connections at each node to capture information about the current node itself, in (2) above. In the matrix, $A_{i,j} = 1$ denotes the presence of a directed edge from node i to node j in the dependency tree. At the k^{th} layer of convolution in the GCN, in (3) above, the hidden representation is derived from the representations of its neighbors at the $(k-1)^{th}$ layer. The final hidden representation

of each node after the k^{th} layer is the encoding of each word $\boldsymbol{h}_i^{(k)}$ in our language-universal common space, and incorporates information about its neighbors up to k hops away in the dependency tree.

In contrast to the common space construction from a high-resource domain in a single language (section 1.3.1), where event structures become available for grounding extractions in a low-resource domain, here the construction of a multi-lingual common space from high-resource languages, makes available structures for training extractors with one or more high-resource languages. The resulting extractors are then applied, in a direct transfer, to a low-resource language.

1.3.3 Multimedia Common Space Construction

We now consider the construction of a third type of common space to bridge the representation of events originating in different resource modalities. By thinking of the content of an image or a video as a foreign language, we further extend our approach to constructing a multimedia structured common space from both text and visual sources.

On the text side, we select Abstract Meaning Representations (AMRs) for their semantic structures that capture whole sentence meanings in rooted, directed, labeled, and (predominantly) acyclic graph structures (Banarescu et al., 2013). AMRs have been developed in conjunction with multi-layer linguistic analyses such as PropBank frames, non-core semantic roles, co-reference, named entity annotation, modality and negation, and include 150 fine-grained semantic roles.

On the vision side, we work with a representation similar to AMR graphs that can encode the semantic structure of the image. Inspired by research on Situation Recognition (Yatskar et al., 2016), we represent an image using a context node that stands in for the event, along with one or more entity nodes that stand in for key arguments of the event, forming what we refer to as a *Situation Graph*. In order to identify candidate arguments for events in images, we seek to extract entities to fill argument roles. In computer vision, the task most similar to entity extraction is *object detection.* We apply a Multi-Layer Perception (MLP) approach to detect visual objects of types defined in OpenImages (Kuznetsova et al., 2018). However, state-of-the-art object detection methods only cover a limited set of object types. Many objects in a scene in an image that may be salient to a human viewer, such as *stone* and *stretcher*, are not included in the ontologies developed in the computer vision research community. Moreover, when there are too many object instances

to be detected in an image (e.g., a big crowd of protesters), the GCN representation tends to lose focus.

To address these issues, we construct a role-driven attention graph for images, where each argument node is derived based on the attention heatmap for each role. This way, the edges of AMR graphs (from text) and situation graphs (from images) both indicate semantic roles of the entities identified respectively within text and image events. The similarity between these two graph structures of verbal and visual information enables us to exploit structure-level alignment, and learn a common embedding space where events, entities, and semantic roles can be represented, independent of the modality they come from.

Figure 1.4 illustrates the steps in the multimedia common space construction. Starting on the left side of the figures, since the image content is relevant to the sentence content, and we can see that the protesters and bus appear in both modalities, we want the embedded nodes for these entities that correspond to each other cross-modally to be close to each other in the common space. More specifically, first, the embedding of the event node from the image, denoted by *throwing*, should be close to the embedding of the corresponding event node from the text *attack*. Then, second, the embedding of the visual entity node, denoted by *man*, should be close to the embedding of the text entity node from the text *protesters*. Furthermore, there should be available external relational knowledge that can be input to the common space to ensure that the embedding of the node representing the visual entity *car* is close to that for the text entity *bus*.

Since the visual *throwing* can be a component activity of the event *attack* mentioned in the text, we would also expect their embeddings to be close to each other within the constructed common space. Given thisdf proximity, this image would be assigned to the sentence and the node *attack* from the text would be classified as a *Conflict.Attack* event type. In completing the representation of the event in the common space, the entity nodes *man* from the image and *protesters* from the text would both be classified as *agent* roles of *Attack* event, and the visual and text entity nodes, respectively *car* and *bus*, would both be classified for that event, as the *target* role. More interestingly, the entity node *Bangkok* from the text that is recognized as a place, even though it does not correspond to any node in the image, would be classified as the *place* role of the *Attack* event. Similarly, vice versa, the entity node *stone* from the image would be classified as the *instrument* role for the event, even

though there is no corresponding text mention of that entity and no other mention fills that role in text.

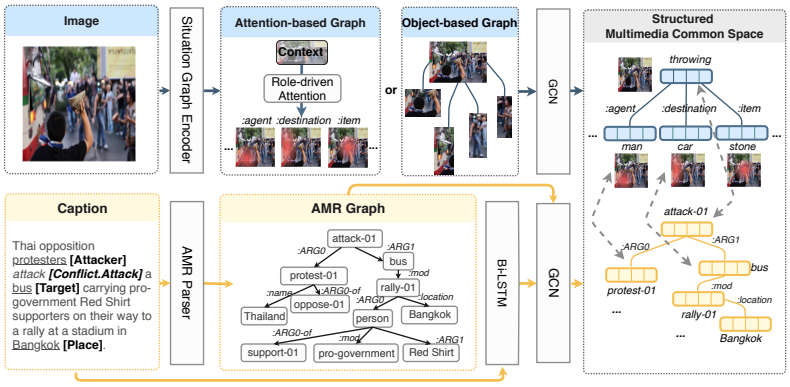

Figure 1.4 Multimedia Structured Common Space Construction

As shown in Figure 1.4, the cross-media common representation encoder is a two-branch representation network, with each branch encoding one modality. We first construct a graph for each modality independently, and apply a GCN to encode the structured information of each graph. We then align them based on the image-caption pairs collected from 17 years of Voice of America news articles.

This approach to constructing a multimedia common space builds on the complementary strengths of modality-specific methods in identifying content that would not otherwise be available to extractors built in the next transfer phase, by way of mutual enhancement. In this way, the textual descriptions of events and arguments that are not typically photographed or not readily visualized can contribute their event structures and roles to the common space, and vice versa, the images of events and objects so typical of their setting that they are not described in text can contribute the entities associated with events of that type to the common space.

1.4 Transfer: From High to Low Resource Setting

Once we have constructed the common semantic space for the selected settings and built network models for our high-resource settings, we

can apply various transfer learning strategies to new, previously unseen data, depending on the amount of available resources in each of the low-resource settings. The pre-processing methods to parse the input data need to be customized for the target setting. In cross-lingual transfer we choose dependency parsing instead of AMR parsing to convert sentences into graphs because the former is available for 76 languages while the latter is only available for English and Chinese. In contrast, in both cross-domain transfer and cross-media transfer we use AMR for English because it's the richest symbolic semantic representation to date.

Cross-domain Transfer: In the cross-domain setting, we apply zero-shot transfer learning (Frome et al., 2013; Norouzi et al., 2013; Socher et al., 2013), which has been very successful in visual object classification, to text event extraction. The basic idea of zero-shot learning is to make use of separate, pre-existing classifiers to build a semantic, cross-concept space that enables the accommodation of new types with no (zero) additional training examples.

We define a similarity metric on the space predicated on these features, and map (*ground*) each event argument candidate to the closest argument role in this space. Consider now again Figure 1.2. Among the three roles of the *rescue* event in the new domain, the argument candidate *(Ukrainian man)* is closer to the embedding representation of the correct role, *patient*, than it is to the event's other roles, *agent* or *time*. Any training data available for a new role can use these instances to train the CNN and the metric. The crucial advantage is that we only have to train it once because this metric is *independent of event types and domains*, supporting event role transfer from old domains (e.g., *military action*) to new ones (e.g., *rescue*) with no additional annotation.

Cross-lingual Transfer: For our cross-lingual setting, we adopt a more straightforward *direct transfer* approach. Using the shared multilingual semantic space, we train event argument extractors with high-resource language training data, and then for transfer, apply the resulting extractors to texts of low-resource languages that do not have any relation or event argument annotations.

We adopt graph convolutional networks (GCN) (Kipf and Welling, 2017; Marcheggiani and Titov, 2017) to encode graph structures over the input data, applying graph convolution operations to generate contextualized word vectors as representations in a latent space. In contrast to other encoders such a Tree-LSTM (Tai et al., 2015), GCNs can cover more complete contextual information from dependency parses because, for each word, it captures all dependency parse tree neighbors of the

word, rather than just the child nodes of the word. Using this shared encoder, we treat the event argument role labeling task as mapping from the latent space to event type and argument role, respectively.

Cross-media transfer: In the cross-media setting, some amount of manual annotations exists for both texts and images. We merge the training data from all data modalities to train the event extractor and argument role labeling component. Using the common representations across modalities as input (described earlier), we train the event classifier and argument role classifier separately.

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 their pairwise similarities. We pair each sentence s with its closest image m, and aggregate the features of each word of s with the aligned representation from m by weighted averaging. We use the aggregated multimedia features to classify each word into an event type and to classify each text-based entity into a role with multi-modal classifiers. Similarly, for each image m we find the closest sentence s, compute the aggregated multi-modal features and feed them into the shared classifiers 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. Figure 1.5 illustrates the cross-media joint inference procedure.

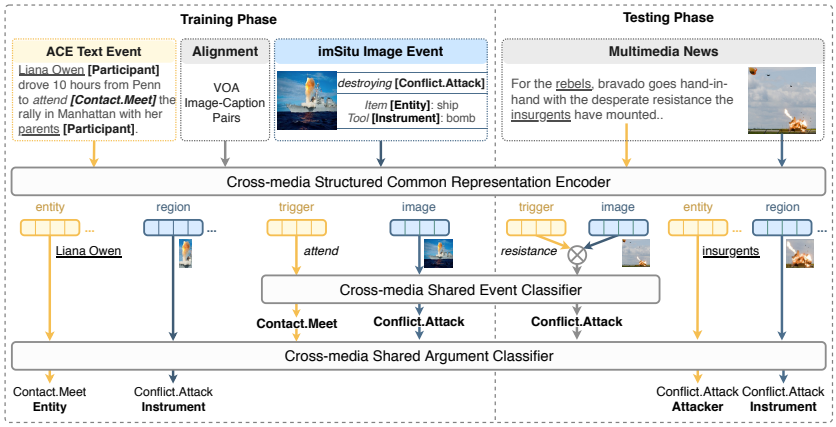

Figure 1.5 Joint Inference based on Multimedia Structured Common Representations

1.5 Transfer Learning Performance

In our preliminary English monolingual event extraction experiments, (Huang et al., 2018), without using any annotations in new domains, our zero-shot learning approach achieves comparable performance to that of a state-of-the-art LSTM model, trained on 3,000 fully annotated sentences for the new domains (Huang et al., 2018). Extensive experiments on cross-lingual relation and event transfer among English, Chinese, and Arabic demonstrate that our approach achieves performance comparable to state-of-the-art supervised models trained on up to 3,000 manually annotated mentions, which costs human annotators about one year to prepare. The event argument role labeling model transferred from English to Chinese achieves similar performance to the model trained from Chinese. We thus find that language-universal symbolic and distributional representations are complementary for cross-lingual structure transfer (Subburathinam et al., 2019). Compared to uni-modal state-of-the-art methods, our cross-media transfer learning approach achieves 4.0% and 9.8% absolute F-score gains on text event argument role labeling and visual event extraction. Compared to state-of-the-art multimedia unstructured representations, we achieve 8.3% and 5.0% absolute F-score gains on multimedia event extraction and argument role labeling, respectively. By utilizing images, we extract 21.4% more event mentions than traditional text-only methods.

1.6 Remaining Challenges

As the most exciting and complex task of IE since it was proposed at the Message Understanding Conference (MUC) (Grishman and Sundheim, 1996), Event Extraction remains a major challenge. The goal of this section is to lay out the current status and potential challenges of this task, and suggest possible future research directions.

1.6.1 Triggers and Arguments in the Long Tail

One of the remaining challenges in labeling event triggers is the infamous *long tail* problem: many event triggers in the test data rarely appear beforehand in the training data. Most triggers are verbs or nouns, but some adverbs, multi-word expressions, and informal metaphors may also function as triggers. Table 1.1 shows some challenging examples. Potential

solutions include significantly expanding trigger candidate lists using linguistic resources such as VerbNet (Kipper et al., 2006), and applying joint entity and word embeddings (Pan et al., 2019) for better semantic matching.

Event Type	**Event Trigger and Context**
Movement.Transport	His men **back** to their compound.
Conflict.Attack	A suicide bomber **detonated** explosives at the entrance to a crowded ...
Movement.Transport	Medical teams **carting** away dozens of wounded victims.
Conflict.Attack	This morning in Michigan, a second straight night and into this morning, hundreds of people have been **rioting** in Benton Harbor.
Election_Person	We've seen in the past in Bosnia for example, you held elections and all of the old ethnic **thugs** get into power because they have organization and they have money and they stop the process of genuine building of democracy.
Sue_Crime	A source tells US Enron is considering suing its own investment bankers for **giving it bad financial advice**.
Employment_End-Position	Today I was **let go** from my job after working there for 4 1/2 years.
Life_Injure	Three boung boys, ages 2, 5 and 10 survived and are in **critical** condition after spending 18 hours in the cold.

Table 1.1 *Examples of Missed, Rare Event Triggers and Arguments*

1.6.2 Ambiguous Argument-Dependent Trigger Words

A challenge at the other end of the distribution arises with frequent support verbs. They typically have multiple senses and are used to indicate various event types in the training data. As a result, they are often mistakenly extracted as triggers when used as abstract words in other contexts. Table 1.2 shows some error examples. Addressing these errors requires us to incorporate the full range of their semantics with associated arguments and contexts. For example, the support verb "get" may indicate a *Transfer.Ownership* event ("Ellison to spend $10.3 billion to **get** his company.") or a *Movement.Transport* event ("Airlines are **getting** flyers to destinations on time more often. ") , as a function of the verb's direct object, what is being got Chen et al. (2015a).

Incorrect Event Type	Event Trigger and Context
Contact.Correspondence	The memorial will take nine months to build. Victims of the regime have been **calling** for reparations for the suffering and loss caused by the Khmer Rouge, including memorials and mental health centers throughout the country.
Life.Die	I miss him to **death**.
Movement.Transport	Stewart has found the road to fortune wherever she has **traveled**.
Movement.Transport	I want to **take** this opportunity to stand behind the Mimi and proclaim my solidarity.
Movement.Transport	He's **left** a lot on the table.
Employment.End-Position	And it's hard to win back that sort of brand equity that she's **lost**.
Conflict.Attack	Still **hurts** me to read this.
Movement.Transport	We happen to be at a very nice spot by the beach where this is a chance for people to **get** away from CNN coverage, everything, and kind of relax.
Movement.Transport	He bought the machinery, **moved** to a new factory, rehired some of the old workers and started heritage programs.

Table 1.2 *Examples of Spurious Event Triggers*

1.6.3 Syntactic Structures

Deep neural models using lexical embedding features have achieved impressive gains on event extraction. However, most of these models are still sensitive to dependency parsing errors and sometimes are biased by embedding features. For example, consider the following sentence: "This supposition is strongly supported by the fact that on April 25, seven days before the **fire**, the **Ukrainian security service**, the SBU, caught a group of Putinite terrorists led by an operative from Crimea attempting to set fire to an Odessa bank using Molotov cocktails." Here the phrase "Ukrainian security service" is mistakenly identified as the *target* of *Conflict.Attack*, as triggered by "fire," because the phrase and the trigger appear close to each other in the sentence. Similarly, consider another sentence: "The EU foreign ministers **met** hours after U.S. President **George W. Bush** gave Saddam 48 hours to leave Iraq or face invasion". Here our model fails to identify the proper multi-clausal syntactic structure of the sentence and so mistakenly identifies "George W. Bush" as the *entity* of the meeting event, rather than as the argument of the subordinating clause.

1.6.4 Localization of Visual Triggers and Arguments

The narrative flow of unstructured text allows event extraction systems to identify trigger and argument boundaries of each event from text. Unfortunately this is not usually the case for images. It's often unclear which region in an image can be said to correspond to the equivalent of an event trigger indicating the image's event type. For example, for an image that shows two people are shaking hands during a meeting, it's not clear whether we should label two hands, two faces or the meeting table as the trigger. Similarly, not all arguments can be precisely localized. The task of object detection in computer vision research suffers from the limited number of object types. An attention-based method is not able to precisely localize the objects for each argument, since there is no supervision on attention extraction during training. Multiple entity arguments often share similar attentions in the same image. When one argument targets too many instances, attention heatmaps tend to lose focus and cover the whole image.

1.6.5 Reasoning with Cross-Event Knowledge

Most current IE methods based on neural networks are limited to sentence-level understanding since they classify single instances with sequence labeling. Progress in multimedia event extraction has helped automate some parts of event understanding, but current automated event understanding is overly simplistic; it is local, sequential and flat. Real events are hierarchical and probabilistic. Understanding them requires knowledge in the form of a repository of abstracted event schemas (complex event templates), scenario models, understanding the progress of time, using background knowledge, and performing global inference.

Table 1.3 provides some remaining error examples from current approaches. In our ongoing work, we are developing approaches to inducing an event schema repository that can be used to automatically discover and verify semantic and structural constraints over extracted events. Previously in (Li et al., 2011), we encode inter-dependency among facts as global constraints in an integer linear programming framework to effectively remove extraction errors. Our systems should now further compare extracted event arguments against background knowledge (e. g., entity profiles, event temporal attributes, schemas and evolving patterns), as extracted from historical data. For example, if we can build a list of instruments employed by policemen when they are the attackers in

The Italian ship was captured by Palestinian terrorists back in 1985 and some may remember the story of Leon Klinghover, he was in a chair and the terrorists shot him and pushed him over the side of the ship into the Mediterranean where he obviously died.
Schema temporal composition: *capture*, *attack* and *die* events happen on the same boat, they share the same year.
Inferred extraction results: *time* of *die* event = 1985
Turkish Armed Forces have taken over the administration. They will reinstate constitutional order, human rights and freedom.
Event interpretation and prediction: Armed forces are unlikely to enforce human rights and freedom in a generic coup schema, unless one is aware of the history of Turkish coups since the 1960s, and the role of the army as the protector of Turkey's democracy. Each time afterward, the military has returned the country to democracy.
Inferred results: *Armed forces* can be predicted to be the *instrument* of *protecting democracy* given background knowledge, deviating from a generic schema.
Then police say the baby's mother pulled out a kitchen knife and on the 911 tape you can hear Williams tape say "go ahead kill me."
Scenario model: in the kitchen environment, *knife* is more likely to be a Die-Instrument argument than in other scenarios.
Fifteen people were killed and more than 30 wounded Wednesday as a suicide bomber blew himself up on a student bus in the northern town of Haifa.
Scenario model: A suicide bomber is both the attacker and the target of an attack event.

Table 1.3 *Error Examples that Require Cross-event Knowledge Reasoning to Address.*

protest events, our system can reason that, for an instance where a woman protester suffered an eye injury, she was not attacked by a Hong Kong policeman because the instrument used for the attack was a slingshot, not an item on the established list. Similarly, when the system has background knowledge of the size of a particular park, then when protestors demonstrate there, the relevant event schema may establish an upper bound on the number of protesters who may populate that location.

1.7 Conclusions and Future Research Directions

In this chapter we provide an overview of our recent research on open-domain Event Extraction from multimedia multilingual sources. We in-

troduce a new cross-domain, cross-lingual, and cross-media structure transfer framework to enable event extraction in new target settings without additional event training data. Moreover, we identify several remaining challenges and possible future research directions. When complex events unfold in an emergent and dynamic manner, the multimedia multilingual digital data from traditional news media and social media often convey conflicting information. To understand the many facets of such complex, dynamic situations, we also need to develop cross-media cross-document event coreference resolution methods for information verification and disinformation detection. Event coreference resolution is a very complex and challenging task itself (Araki et al., 2014). The order of the events depends on local (discourse) cues as well as global understanding of the events unfolding. Cross-media consistency detection and reasoning is often a key to disambiguation. Compared to images, disinformation detection from text at the single-asset knowledge element level can be more challenging, especially when the text is written by a generally trustworthy source. For example, CNN published a news article titled "Police use petrol bombs and water cannons against Hong Kong protesters," but an original video revealed protesters threw bombs at the police. CNN has apologized for its "erroneous" reporting since. Cross-media consistency detection and trustworthy assessment should be added into the research paradigm of events, and the results should be used as feedback to enhance event extraction. Finally, once the quality of event extraction reaches a satisfactory level, we can also leverage KB-to-text generation techniques (Wang et al., 2018) to describe event-centric knowledge bases and construct event timelines or even history books.

References

Ahn, David. 2006. The stages of event extraction. In: *Proceedings of the Workshop on Annotating and Reasoning about Time and Events.*

Araki, Jun, Hovy, Eduard, and Mitamura, Teruko. 2014. Evaluation for Partial Event Coreference. In: *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation.*

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

Chen, Wei-Te, Bonial, Claire, and Palmer, Martha. 2015a. English Light Verb Construction Identification Using Lexical Knowledge. In: *Proceedings of the AAAI-15.*

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

Chen, Yubo, Yang, Hang, Liu, Kang, Zhao, Jun, and Jia, Yantao. 2018. Collective Event Detection via a Hierarchical and Bias Tagging Networks with Gated Multi-level Attention Mechanisms. In: *Proc. EMNLP2018.*

Do, Quynh Ngoc Thi, Bethard, Steven, and Moens, Marie-Francine. 2017. Improving Implicit Semantic Role Labeling by Predicting Semantic Frame Arguments. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing.*

Doddington, George R, Mitchell, Alexis, Przybocki, Mark A, Ramshaw, Lance A, Strassel, Stephanie M, and Weischedel, Ralph M. 2004. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In: *Proceedings of LREC 2004.*

Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. In: *Proc. NIPS2013.*

Geldenhuys, Aletta E., Rooyen, Hendrik O. van, and Stetter, Franz. 1999. Some Approaches to Knowledge Representation. In: *Knowledge Representation and Relation Nets pp 3-22.*

Grishman, R., and Sundheim, B. 1996. Message Understanding Conference-6: A brief history. In: *Proc. The 16th International Conference on Computational Linguistics (COLING1996)*.

Grishman, Ralph, Westbrook, David, and Meyers, Adam. 2005. NYU's English ACE 2005 system description. *ACE*, **5**.

Harris, Zellig S. 1954. Distributional structure. In: *Word, 10(23):146–162*.

Hong, Yu, Zhou, Wenxuan, zhang jingli, jingli, Zhou, Guodong, and Zhu, Qiaoming. 2018. Self-regulation: Employing a Generative Adversarial Network to Improve Event Detection. In: *Proc. ACL2018*.

Huang, Lifu, Ji, Heng, Cho, Kyunghyun, Dagan, Ido, Riedel, Sebastian, and Voss, Clare. 2018. Zero-Shot Transfer Learning for Event Extraction. Pages 2160–2170 of: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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

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

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

Kipf, Thomas N., and Welling, Max. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In: *International Conference on Learning Representations*.

Kipper, Karin, Korhonen, Anna, Ryant, Neville, and Palmer, Martha. 2006. Extensive Classifications of English Verbs. In: *Proceedings of the 12th EURALEX International Congress*.

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

Li, Qi, Anzaroot, Sam, Lin, Wen-Pin, Li, Xiang, and Ji, Heng. 2011. Joint Inference for Cross-document Information Extraction. In: *Proc. 20th ACM Conference on Information and Knowledge Management (CIKM2011)*.

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

Li, Qi, Ji, Heng, Yu, HONG, and Li, Sujian. 2014. Constructing information networks using one single model. In: *Proceedings of EMNLP 2014*.

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

Lin, Yankai, Liu, Zhiyuan, and Sun, Maosong. 2017. Neural relation extraction with multi-lingual attention. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Liu, Xiao, Luo, Zhunchen, and Huang, Heyan. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In: *Proceedings of EMNLP 2018*.

Marcheggiani, Diego, and Titov, Ivan. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Nguyen, Thien Huu, and Grishman, Ralph. 2015. Event detection and domain adaptation with convolutional neural networks. In: *Proc. ACL-IJCNLP2015*.

Nguyen, Thien Huu, and Grishman, Ralph. 2016. Modeling skip-grams for event detection with convolutional neural networks. In: *Proc. EMNLP2016*.

Nguyen, Thien Huu, and Grishman, Ralph. 2018. Graph convolutional networks with argument-aware pooling for event detection. In: *Proceedings of AAAI 2018*.

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

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G., and Dean, J. 2013. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*.

Pan, Xiaoman, Gowda, Thamme, Ji, Heng, May, Jonathan, and Miller, Scott. 2019. Cross-lingual Joint Entity and Word Embedding to Improve Entity Linking and Parallel Sentence Mining. In: *Proc. EMNLP2019 Workshop on Deep Learning for Low-Resource Natural Language Processing*.

Poon, Hoifung, and Vanderwende, Lucy. 2010. Joint inference for knowledge extraction from biomedical literature. In: *Proc. NAACL-HLT2010*.

Riedel, Sebastian, Chun, Hong-Woo, Takagi, Toshihisa, and Tsujii, Jun'ichi. 2009. A markov logic approach to bio-molecular event extraction. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*.

Simmons, Robert F. 1963. Synthetic Language Behavior. In: *Data Processing Management. 5 (12): 11–18*.

Socher, R., Ganjoo, M., Manning, C., and Ng, A. 2013. Zero-shot learning through cross-modal transfer. In: *Proc. NIPS2013*.

Subburathinam, Ananya, Lu, Di, Ji, Heng, May, Jonathan, Chang, Shih-Fu, Sil, Avirup, and Voss, Clare. 2019. Cross-lingual Structure Transfer for Relation and Event Extraction. In: *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP2019)*.

Tai, Kai Sheng, Socher, Richard, and Manning, Christopher D. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Venugopal, Deepak, Chen, Chen, Gogate, Vibhav, and Ng, Vincent. 2014. Relieving the computational bottleneck: Joint inference for event extraction with high-dimensional features. In: *Proc. EMNLP2014*.

Wang, Qingyun, Pan, Xiaoman, Huang, Lifu, Zhang, Boliang, Jiang, Zhiying, Ji, Heng, and Knight, Kevin. 2018. Describing a Knowledge Base. In: *Proc. The 11th International Conference on Natural Language Generation*.

Yang, Bishan, and Mitchell, Tom. 2016. Joint extraction of events and entities within a document context. *arXiv preprint arXiv:1609.03632*.

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