Multimodal Relation Extraction with Efficient Graph Alignment

Changmeng Zheng
Department of Computing, The Hong Kong Polytechnic University
Key Laboratory of Big Data and Intelligent Robot (SCUT), MoE
Hong Kong SAR, China

Junhao Feng
Key Laboratory of Big Data and Intelligent Robot (SCUT), MoE
School of Software Engineering, South China University of Technology
Guangzhou, Guangdong, China

Ze Fu
Key Laboratory of Big Data and Intelligent Robot (SCUT), MoE
School of Software Engineering, South China University of Technology
Guangzhou, Guangdong, China

Yi Cai*
Key Laboratory of Big Data and Intelligent Robot (SCUT), MoE
School of Software Engineering, South China University of Technology
Guangzhou, Guangdong, China
ycai@scut.edu.cn

Qing Li
Department of Computing, The Hong Kong Polytechnic University
Hong Kong SAR, China
qing-prof.li@polyu.edu.hk

Tao Wang
Department of Biostatistics and Health Informatics, King's College London
London, United Kingdom

ABSTRACT

Relation extraction (RE) is a fundamental process in constructing knowledge graphs. However, previous methods on relation extraction suffer sharp performance decline in short and noisy social media texts due to a lack of contexts. Fortunately, the related visual contents (objects and their relations) in social media posts can supplement the missing semantics and help to extract relations precisely. We introduce the multimodal relation extraction (MRE), a task that identifies textual relations with visual clues. To tackle this problem, we present a large-scale dataset which contains 15000+ sentences with 23 pre-defined relation categories. Considering that the visual relations among objects are corresponding to textual relations, we develop a dual graph alignment method to capture this correlation for better performance. Experimental results demonstrate that visual contents help to identify relations more precisely against the text-only baselines. Besides, our alignment method can find the correlations between vision and language, resulting in better performance. Our dataset and code are available at https://github.com/thecharm/Mega.

CCS CONCEPTS

• **Information systems** → l0>499**Multimedia and multimodal retrieval**; • **Computing methodologies** → l0>499**Information extraction**.

KEYWORDS

multimodal relation extraction; graph alignment; multimodal dataset

*Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3476968

ACM Reference Format:
Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021. Multimodal Relation Extraction with Efficient Graph Alignment. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3476968

1 INTRODUCTION

The task of relation extraction (RE) is to identify the semantic relations given two entities in a sentence. RE plays an important role in many applications requiring relational understanding such as question answering[13] and knowledge base population[2]. Most existing relation extraction methods can be divided into two categories: sequence-based models [27, 31] and dependency-based models [6, 33]. Compared to sequence-based models, dependency-based methods can capture the long distance semantic dependency and usually gain better performance.

However, these methods are mainly text-based and suffer sharp performance decline in social media posts when texts lack of contexts. For example, in a sentence "JFK and Obama at Harvard", given two entities "JFK" and "Obama", traditional methods can hardly detect the relation between them is "Alumni" without other supplementary information. As a result, most methods will incorrectly extract the relation "couple" of the two entities since most cases in training corpus are labelled with such tags. We find the image-related information can be a good resource to supplement the missing contexts in relation extraction in social media texts. In the above case, we can easily classify the relation into "Alumni" with an image showing that the two people wear bachelor caps and the same school uniforms.

Utilizing the visual contents to complement textual contexts becomes a research hotspot in recent studies involving multimodal learning [3]. Multimodal named entity recognition is one of the tasks which requires both understanding of vision and language. Zhang et al. [32] propose an adaptive co-attention network which utilizes image-level region features to help extract entities in tweets. Wu et al. [28] consider object-level features as a fine-grained features and provide a new attention method to align visual objects

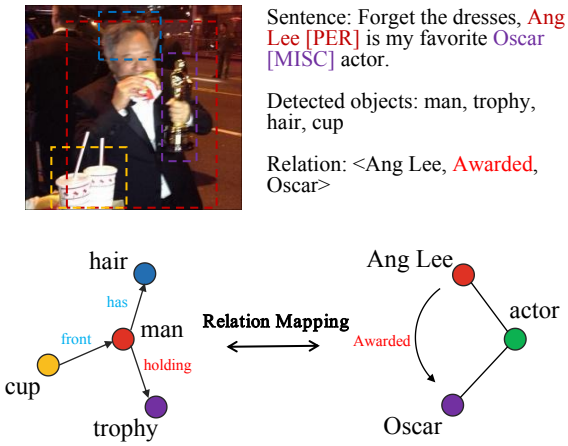


Figure 1: An example of multimodal relation extraction in Twitter. The mappings from visual contents "man holding a trophy" to textual entities "Ang Lee" and "Oscar" will lead to the extraction of textual relation "awarded".

and textual entities. Different from multimodal named entity recognition task, introducing visual information into relation extraction asks models not only to capture the correlations between visual objects and textual entities, but also to focus on the mappings from visual relations between objects in an image to textual relations between entities in a sentence. For example, in Figure 1, our goal is to classify the relation category given the two entities "Ang Lee" and "Oscar". Previous text-based relation extraction models cannot detect the relation "awarded" only from the textual information. However, we can easily gain the correct label from the guidance of "man holding a trophy". The visual relation "holding" may reflect the textual relation "awarded" between the two entities and objects.

In this work, we study **multimodal relation extraction** (MRE), which is the problem of classifying textual relations between two entities with the help of visual contents. Since there is no available dataset for training and evaluating MRE models, we present the MNRE dataset, a manually-labelled dataset for multimodal neural relation extraction. The corpus consists of texts and image posts crawled from Twitter. Four well-educated annotators were asked to tag both the entities and their relations. Due to the noisy nature of social media texts and the limited characters of tweets, MNRE is a challenge dataset to test the multimodal representation, fusion and also reasoning abilities of existing methods.

To learn the mapping from visual relations to textual relations, we propose a **M**ultimodal Neural Network with **E**fficient **G**raph **A**lignment (MEGA) for relation extraction in social media posts. Following the success of dependency-based RE methods [6, 33], we parse the sentences with a dependency tree tool [21]. Considering the scene graphs can be a fine-grained image representation and a parser to analyze the relations with a graph structure, we apply a pretrained scene graph model [25] to extract visual objects and their relations preliminarily. As shown in Figure 1, to capture the relation mapping from visual contents ("man holding a trophy") to textual relations ("Ang Lee is awarded for Oscar"), we propose a graph alignment method that incorporates structural similarity and semantic agreement between visual objects in a image and textual entities in a sentence. Different from previous multimodal methods simply concatenating the graph representations using a graph convolutional network [10, 12], our method can find the most similar nodes between two graphs with structural and semantic features, which resulting in a better alignment for textual and visual relations. The corresponding visual relations can help our model identify textual relations more precisely.

Our main contributions can be summarized as follows:

- We present the multimodal relation extraction (MRE) task which leverages related visual contents to help to extract relations between entities in social media when texts lack of contexts. Since there is no available dataset, we also provide a human-annotated dataset (MNRE) for training and evaluating multimodal relation extraction neural models.
- We propose a multimodal relation extraction neural network with efficient alignment strategy for textual and visual graphs. Compared to previous relation extraction methods, our model can find the correlations between visual objects and textual entities and leverage the visual relations to classify textual relations more precisely.
- We conduct experiment on the MNRE dataset, and the experimental results demonstrate that introducing visual information can supplement the missing semantics of short social media texts. Also, our efficient graph alignment method can improve relation extraction performance with aligned visual and textual relations.

2 RELATED WORKS

2.1 Relation Extraction in Social Media

Relation extraction task are the fundamental process of constructing a knowledge graph. Early researches on relation extraction are based on statistics methods [17, 29]. In recent years, sequence-based methods utilize neural networks and improves RE performance with convolutional neural networks [27], recurrent neural networks [34] and transformers [26].

Dependency-based RE models try to incorporate structural information into predicting relations. Compared to sequence-based methods, dependency-based models are more capable of capturing information from long distance. Peng et al. [20] propose a general framework for cross-sentence n-array relation extraction based on graph LSTMs. Song et al. [24] employ a graph recurrent neural network without changing the input graph structure. Zhang et al. [33] propose a path-centric pruning strategy with graph convolutional networks and Guo et al. [6] improve it with attentive graph weights. Most recently, BERT-based pretraining methods [19, 23] improves RE performance significantly with external training corpus.

Despite the success of using dependency or external information, most existing methods suffer performance decline in social media texts when sentences lack of contexts. However, compared to RE in general or newswire domain, there are few works concerning on the relation extraction on social media [4, 14]. We propose to leverage the image information to supplement the missing semantics in short texts and present a large-scale multimodal relation extraction dataset. Our experimental results demonstrate that introducing

visual information can improve the RE performance with a large margin.

2.2 Multimodal Representation and Alignment

In this paper, we also study multimodal representation and alignment strategy. Similar to the multimodal relation extraction task, multimodal named entity recognition is also a task which requires both understanding of visual and textual information. There are many models proposed to leverage the image-level visual attention for aligning images and texts [15, 32]. However, image-level features cannot help to extract entities with different types since they are trained with only one semantic labels. Wu et al. [28] propose a embedded object-level representations for taking the fine-grained visual objects into consideration. Zheng et al. [35] adopt bi-linear attention networks to align the inner and inter relations between visual objects and textual entities.

Different from multimodal named entity recognition, relation extraction task needs to analyze not only the relation between objects and entities, but also the relation graph which reveals the mapping of visual and textual relations. We build the visual graphs using a pretrained scene graph generator [25]. Inspired by Heimann et al. [8], we assign the graph similarity computed by both structural similarity and semantic agreement. We show this efficient graph alignment strategy will be beneficial to find the mapping from visual relation to textual contents, finally improves the multimodal RE performance.

3 METHODOLOGY

In this section, we introduce the MEGA model for multimodal relation extraction, which is shown in Figure 2. In order to build the model, our work can be summarized as the following steps: (1) First, we extract the textual semantic representations with a pretrained BERT encoder. Besides, we generate the scene graphs from images which provide rich visual information including visual objects features and visual relations among the objects. To represent the semantics of images, we regard the object features in the extracted scene graph as the visual semantic features. (2) Secondly, to acquire the structural representations, we obtain the syntax dependency tree of the input texts which models the syntax structure of textual information. The visual object relations extracted by scene graph can be constructed as a structural graph representation. (3) Thirdly, to make good use of image information for multimodal relation extraction, we respectively align the structural and semantic information of multimodal features to capture the multi-perspective correlation between multimodal information. Then, we effectively merge the two aligned results. (4) Finally, we concatenate the textual representations which represent the two entities and the aligned visual representation as the fusion feature of text and image to predict the relations of entities.

3.1 Semantic Feature Representation

3.1.1 Textual Semantic Representation. In the MNRE dataset, each piece of data contains a text message and an corresponding image from the social media posts, which is used as the input of our model. The input text message is first tokenized into a token sequence s_1. Then, to fit the BERT encoding procedure, we add the token '[CLS]' to the head of the sequence and the token '[SEP]' to the tail as well. In addition, following Soares et al.[23], we augment the s_1 with four reserved word pieces, $[E1_{start}]$, $[E1_{end}]$, $[E2_{start}]$ and $[E2_{end}]$ to mark the begin and end of each entity mentioned in the relation statement and modify s_1 to sequence $\tilde{s_1}$ as shown in Eq.(1),

$$\begin{aligned}\tilde{s_1} =&[w_1, ..., [E1_{start}], w_i, ..., w_{i+n_1-1}, [E1_{end}] \\ &, ..., [E2_{start}], w_j, ..., w_{j+n_2-1}, [E2_{end}], ..., w_l]\end{aligned} \tag{1}$$

where i and j denotes the start position of the first and second entity respectively. n_1 represents the length of the first entity while n_2 denotes the length of the second one. Besides, the token sequence are trimmed to a maximum length l. We pad the sample sequence which has less than l tokens to maximum length by [PAD] token.

Besides, we set a segment sequence to represent the segmentation of the valid tokens and [PAD] tokens. The segment sequence can be denoted as $s_2 = (1, 1, ..., 1, ..., 0, 0)$, where 1 represents the token which is not a padding one, 0 represents the [PAD] token. Therefore, the length of s_2 is l the same as $\tilde{s_1}$.

Following the success of Lample et al. [11], Ma and Hovy [16] , we represent each word in a input text message by combining character embedding into word embedding to obtain its textual features. We fine-tune the pre-trained BERT to get the embedding for each token in the sequence. The two sequences $\tilde{s_1}$, s_2 are fed into BERT to generate the embeddings. After that, each word is further transformed into a vector of d_x dimensions. And we can obtain the textual semantic representation by transforming the whole text message into a matrix $X \in \mathbb{R}^{l \times d_x}$, which is denoted in Eq.(2),

$$X = BERT(\tilde{s_1}, s_2) \tag{2}$$

where *BERT* denotes the BERT Encoder.

3.1.2 Visual Semantic Representation. Object-level visual features are considered as bottom-up manners in several multimodal tasks [1] to represents the image information. Therefore, we obtain the visual semantic feature by extracting the objects representation to represents the semantic of input image. In order to extract the objects from images, the input image is fed into the pre-trained scene graph generation model(with Faster R-CNN[22] as its backbone) to generate the scene graph of input image. An scene graph contains several nodes and edges connecting related nodes. The node contains the object features as its inner information, while the edges model the visual relation such as holding and wearing between different objects. In order to assist the entities relation extraction, we exploit the effective visual information while ignoring the irrelevant ones. Thus, we solely consider the top m salient objects with the higher object classification scores as the valid visual objects for further processing.

The input image is represented as a set of regional visual features in a bottom-up manner contained in the extracted scene graph. Each regional visual feature represents an object in the image with a vector y_i in dimension d_y . We set a confidence threshold to the probabilities of detected objects and obtain the top m objects for each image. Finally, an input image is transformed to a matrix Y. If the number of detected objects in an image is less than m, we would zero-pad Y to the maximum size m.

$$Y = [y_1, y_2, ..., y_m]_{m \times d_y} \tag{3}$$

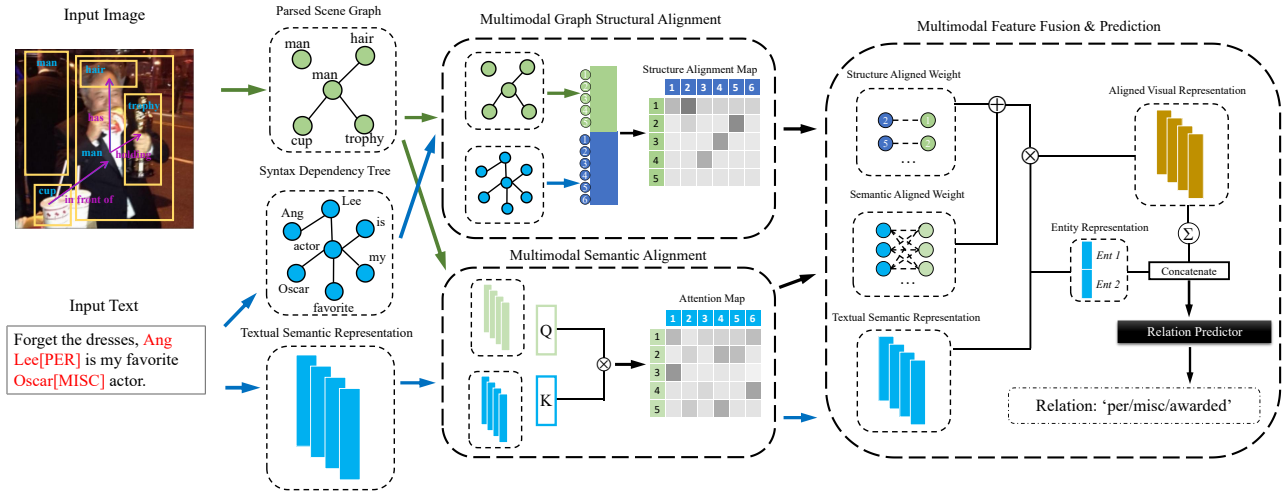


Figure 2: The Overall Framework of Our Proposed MEGA Model. Our Model Introduces Visual Information into Predicting Textual Relations. Besides, We leverages the Graph Structural Alignment and Semantic Alignment to Help Model Find the Mapping From Visual Relations to Textual Contents.

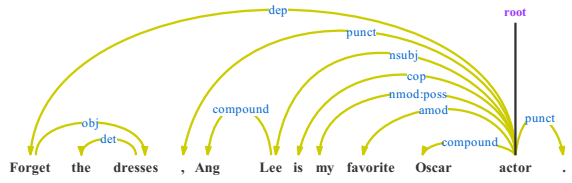


Figure 3: The Input Text Message is Performed by Syntactic Dependency Parsing. The Word *actor* is the Root Node of Dependency Relations while the Words in Blue (e.g., dep, obj) are Dependency Relations. The Direction of Arrow Indicates that There is a Relation Between the Two Words.

3.2 Structural Feature Representation

In some previous works, the structure of the sentences (i.e., dependency trees) can provide important information which is benefit for the relation extraction models. Inspired by this, we generate two unidirectional graphs for the input text and image by using syntax dependency tree and scene graph generation model, which can provide the structural information to help multimodal relation extraction. It is notable that the visual object features plays the role as the node features in the scene graph.

3.2.1 Syntax Dependency Tree. Dependency tree is a structure used to express the dependency between words in a sentence. It has been shown in many previous work that the dependency trees can provide important information/features for the relation extraction. Each dependency corresponding to two words from a sentence can be represented as a triple as Eq.(4):

$$R_{dependency} = (w_g, r_{type}, w_d) \tag{4}$$

where w_g is the governor, w_d is the dependent and r_{type} shows how the dependent modifies the governor. We use ELMo [21], a common dependency tree extraction tool to obtain the dependency tree for the input text after which each word from the text is connected by its governor and obtains its related dependency triple. For example, the sentence *Forget the dresses, Ang Lee is my favorite Oscar actor.* is parsed to obtain the relations between words(e.g., amod, cop), as shown in Figure.3. The words in blue are the dependency relations. The ending of arrow indicates that this word is a dependent as well as a modifier. The word *root* in purple is used to indicate which word is the root node of dependency relations. Since each word is connected directly by another word in the text, the graph representation of the text is generated as G_1, which consists several relation pairs among the words.

$$V_1 = \{t_i | i \in [1, l_0]\} \tag{5}$$

$$E_1 = \{e_i = [t_i^*, t_i] | i \in [1, l_0]\} \tag{6}$$

$$G_1 = (V_1, E_1) \tag{7}$$

t_i represents the node corresponding to the ith token in the original text message which are not padded. t_i^* represents the governor of the ith token. l_0 represents the length of token sequence.

3.2.2 Scene Graph Generation. We obtain m objects and the visual relation between them from the input image by scene graph generation model. Since every relation between two objects is unidirectional, similar to the dependency tree, each object is also pointed by its governors from the image. Therefore, we can obtain the graph representation G_2 of the input image. G_2 consists several relation pairs of objects detected in the image and can be denoted as follows:

$$V_2 = \{o_j | j \in [1, m_0]\} \tag{8}$$

$$E_2 = \{e_{j,j_r} = [o_j, o_{j_r}^*] | j \in [1, m_0], j_r \in [0, r]\} \tag{9}$$

$$G_2 = (V_2, E_2) \tag{10}$$

where o_j represents the node corresponding to the jth object detected in the image. m_0 represents the number of detected objects. $o_{j_r}^*$ denotes the j_rth object which is related to jth object.

$r \in [0, m_0 - 1]$ denotes the dynamic number of objects related to the jth object. After generating G_1 and G_2, we obtain the graph representation of the input text and image.

3.3 Multimodal Feature Alignment

To make full use of the obtained multimodal representation, we align the two graphs above from the structural perspective and use attention mechanism to align the textual and visual features from the semantic perspective.

3.3.1 Graph Structure Alignment. We exploit the node and edge information to extract the structure similarity of multimodal graph representation for structural alignment. First, as shown in Equation (7) and (10), we set $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ as two graphs mentioned above with node sets V_1 and V_2; edges sets E_1 and E_2 respectively. Let n be the number of nodes among two graphs, which means $n = |V_1| + |V_2|$. The steps of structure alignment can be summarized as follows: (1) obtain the node embeddings, conceptually by factorizing a similarity matrix of the node identities; (2) align nodes between two graphs by greedily matching the embeddings with an efficient data structure that allows for fast identification of the most similar embeddings from the other graph.

Following [8], we first set V_1 and V_2 into a union U shown as Eq.(11). In order to extract the node structural identity, we compute the counts of node degrees, including both in and out degrees of k-hop neighbors for each node u in U, which is shown as Eq.(12) and Eq.(13),

$$U = V_1 \cup V_2 \tag{11}$$

$$d_u^k = CountDegreeDistributions(R_u^k) \tag{12}$$

$$d_u = \sum_{k=1}^{K} \delta^{k-1} d_u^k \tag{13}$$

where $k \in [1, K]$, K is a graph diameter set by us and $\delta \in (0, 1]$ is a discount factor. And we compute the similarity between node a and node b in U as Eq.(14),

$$sim(a, b) = exp[-\gamma_s \cdot ||d_a - d_b||_2^2] \tag{14}$$

where γ_s is a scalar parameter controlling the effect of the structural identity. Then, we randomly select $p \ll n$ "*landmark*" nodes chosen across both graphs G_1 and G_2 and compute their similarities to all n nodes in these graphs using Eq.(15). This yields an $n \times p$ similarity matrix C, from which we can extract a $p \times p$ landmark-to-landmark submatrix W_p. Meanwhile, $W_p^\dagger$ is the pseudoinverse of W_p, a $p \times p$ matrix consisting of the pairwise similarities among the landmark nodes (it corresponds to a subset of p rows of C). It is a theorem that[8] given graphs $G_1(V1, E1)$ and $G_2(V2, E2)$ with $n \times n$ structural similarity matrix $S \approx PZ^T$, its node embedding matrix P can be approximated as

$$\tilde{P} = CU\Sigma^{1/2} \tag{15}$$

where $W_p^\dagger = U \sum V^T$ is the full rank singular value decomposition of the pseudoinverse of the small $p \times p$ landmark-to-landmark similarity matrix W_p. Now P and $\tilde{P}$ is the matrix with node embeddings as rows and its approximation. The p-dimensional node embeddings of the two input graphs G_1 and G_2 are then subsets of $\tilde{P}$: $\tilde{P_1}$ and $\tilde{P_2}$, respectively. We use Eq.(16) to obtain $\tilde{P_1}$ and $\tilde{P_2}$, which are the separate representations for nodes in G_1, G_2.

$$\tilde{P_1}, \tilde{P_2} = D(N(\tilde{P})) \tag{16}$$

where D represents the dividing operation of $\tilde{P}$ by the number of $|V_1|$ and $|V_2|$ in order and N is used to normalize the magnitude of the embeddings and make them more comparable based on Euclidean distance. Finally, the last step is to efficiently align nodes using their representations, assuming that two nodes $i \in V_1$ and $j \in V_2$ may match if their embeddings in G_1, G_2 are similar. We find the alignments for each node by computing all pairs of similarities between node embeddings (i.e., the rows of $\tilde{P_1}$ and $\tilde{P_2}$) and choose the top-1 for each node. Here, we define the similarity a_{ij} between the p-dimensional embeddings of nodes i and j as follows:

$$a_{ij} = sim(\tilde{P_1}[i], \tilde{P_2}[j]) = e^{-||\tilde{P_1}[i] - \tilde{P_2}[j]||_2^2} \tag{17}$$

After the structural alignment of multimodal graphs, for each node in the text, its most similar node in structure from the image and their similarity score would be identified effectively. When finishing graph structure alignment, the two graphs are transformed into a feature matrix α,

$$\alpha = \left(a_{ij} \right)_{|V_1| \times |V_2|} \tag{18}$$

where a_{ij} represents the structural similarity between the ith word of the input text and the jth object of the input image. In our model, we only keep the most structurally similar object for each word while the elements corresponding to other objects except the most similar one are all represented by 0 in the matrix.

3.3.2 Semantic Features Alignment. In order to align the semantic of textual and visual information, we implement the guided-attention mechanism to capture the correlation between multimodal semantic features. The input of scale dot-product attention consists of queries and keys of dimension d_{key}, and keys of dimension d_{value}. For simplicity, we set d_{key} and d_{value} to the same number d_a. We calculate the dot products of the query with all keys, divide each by $\sqrt{d_a}$ and apply a softmax function to obtain the attention weights on the values. Given a query $q \in \mathbb{R}^{1 \times d_a}$, n key-value pairs (packed into a key matrix $K \in \mathbb{R}^{n \times d_a}$ and a value matrix $V \in \mathbb{R}^{n \times d_a}$), the semantic aligned feature $y_a \in \mathbb{R}^{1 \times d_a}$ is obtained by weighted summation over all values V with respect to the attention learned from q and K:

$$y_s = A(q, K, V) = softmax(\frac{qK^T}{\sqrt{d_a}})V \tag{19}$$

In practice, to obtain the semantic aligned features of all visual objects $Y_s \in \mathbb{R}^{m \times d_a}$, we compute the attention function on a set of m queries $Q = [q_1, q_2, ..., q_m] \in \mathbb{R}^{m \times d_a}$ seamlessly by replacing q with Q, which represents the visual semantic information guided by the textual features.

After obtaining the multimodal features representation, the input text is transformed into the matrix $X \in \mathbb{R}^{l \times d_x}$ and the input image is transformed into the matrix $Y \in \mathbb{R}^{m \times d_y}$. We employ three learnable matrix $W_k \in \mathbb{R}^{l \times d_a}$, $W_q \in \mathbb{R}^{m \times d_a}$ and $W_v \in \mathbb{R}^{l \times d_a}$ to generate the feature from X and Y for attention mechanism. In detail, the calculation process is shown from Eq.(20) to Eq.(22),

$$K = W_k X + b_k \tag{20}$$

$$Q = W_q Y + b_q \tag{21}$$

$$V = W_v X + b_v \tag{22}$$

where b_k, b_q, b_v are the learnable biases. As a result, we implement the semantic alignment by obtaining the semantic aligned weight β by calculation of Q and K as Eq.(23).

$$\beta = softmax(\frac{QK^T}{\sqrt{d}}) \tag{23}$$

3.3.3 Alignment Fusion. To fully use the structural and semantic alignment information, we integrate the aligned information by Eq.(24) to obtain the aligned visual features.

$$Y^* = (\alpha^T + \beta)V = \alpha^T V + Y_s \tag{24}$$

As we merge the structural and semantic alignment results, the final aligned visual features representation guided by the text is obtained as matrix $Y^* \in \mathbb{R}^{m \times d_a}$.

3.4 Entities Representation Concatenation

To fully exploit the aligned visual information of all objects, we integrate the aligned object features to a vector representation, shown as Eq.(25),

$$\hat{y} = \sum_{i=1}^{m} y_i^* \tag{25}$$

where $y_i^* \in \mathbb{R}^{1 \times d_a}$ represents the ith object feature in matrix Y^*.

Since we need to extract the relation between two entities from the text, we concatenate the representation $v_{[E1_{start}]}$ and $v_{[E2_{start}]}$ of their start position marker in feature V as the textual representation $\hat{v} \in \mathbb{R}^{1 \times 2d_a}$ for multimodal fusion, which is shown as Eq.(26),

$$\hat{v} = [v_{[E1_{start}]}, v_{[E2_{start}]}] \tag{26}$$

We combine the guided visual information and the textual information from the two entities to obtain the final representation for the text and image by concatenating $\hat{v}$ and $\hat{y}$ into z, which is shown as:

$$z = concat(\hat{v}, \hat{y}) \tag{27}$$

Finally, we input z into an MLP to complete the final task of relation classification and obtain the output as shown in Eq.(28),

$$output = softmax(MLP(z)) \tag{28}$$

where $output \in \mathbb{R}^{n_c}$ represents the classification probability of all n_c relation categories.

4 EXPERIMENT SETTINGS

4.1 Dataset

To provide empirical results for the effectiveness of our model, we construct a multimodal neural relation extraction dataset (MNRE) from scratch. The original corpus is built on three sources: two available multimodal named entity recognition datasets - Twitter15[15] and Twitter17[32], and crawling data from Twitter [1]. The posts were selected and filtered by annotators with different topics, such as music, sports and social events. We employed 12 well-educated annotators to label the relations between entity pairs and filter out the wrong samples tagged by automatic NER tools. The dataset contains 15,484 samples and 9,201 images with 23 relation categories. We split the dataset into training, development and testing set with 12247, 1624 and 1614 samples, respectively. The statistics of MNRE compared with a widely-used relation extraction dataset SemEval-2010 Task 8 [9] are listed in Table 1.

[1] https://archive.org/details/twitterstream

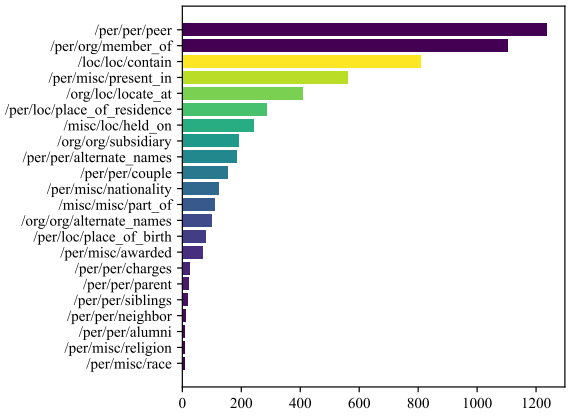


Figure 4: The Distribution of Relation Categories in Our MNRE Dataset.

Table 1: The Statistics of MNRE Dataset Compared to SemEval-2010 Task 8 Dataset. # indicates Numbers.

Statistics	SemEval-2010	MNRE
# Word	205k	258k
# Sentence	10,717	9,201
# instance	8,853	15,485
# Entity	21,434	30,970
# Relation	9	23
# Image	-	9,201

We also show the distribution of relation categories in our MNRE dataset in Figure 4. We start tagging relation types depending on the entity types. For example, the relations between one person and another person can be classified into "alumni", "couple" and "relative" et al. We choose this labeling method since we expect the entity types and visual objects can be aligned and help to understand texts better.

4.2 Baseline Methods

We compare our methods with several relation extraction baselines. To validate the effectiveness of incorporating visual information into text-based RE models, we also provide several variants of the proposed MEGA model.

Glove+CNN Glove+CNN [31] is a classic CNN-based model for relation extraction. We use a improved version of this model [18] which concatenates word embeddings with position embeddings.

PCNN PCNN [30] is a distantly supervised relation extraction model which leverages external knowledge graphs to automatically label sentences with same entities contained.

Matching the Blanks (MTB) MTB [23] is an RE-oriented pretraining model based on BERT. Our method is built on the MTB model which, in turn, is the text-based version of the proposed MEGA model without visual features and the graph alignment strategy. We fine-tune it on our MNRE dataset as a text-based baselines.
BERT+SG The pretrained language model Bert [5] has shown its strong generalization in multiple tasks. We simply concatenate the fine-tuning BERT representations with visual features to show the improvement of introducing visual information. The visual features are extracted by a pretrained scene graph (SG) tool [25].
BERT+SG+Att. A variant of our proposed MEGA model which considers only the semantic similarity between visual graph (scene graph) and textual contents. Here we adopt the attention mechanism to compute the semantic similarity.
MEGA MEGA is our proposed multimodal relation extraction model with efficient graph alignment which considers both structural similarity and semantic agreement between visual and textual graphs.

4.3 Parameter Settings

We implement our model on the open-source and extensible relation extraction toolkit OpenNRE[7]which is based on PyTorch framework. To acquire the textual semantic representation, we initialize the textual representation by pretrained BERT and set the dimension d_x at 768. Besides, the dimension d_y of visual objects features extracted from scene graph is 4096. The latent dimension d_a of semantic alignment is set at 1536. The maximum number of token sequence and objects are 128 and 10 respectively. Our model is trained with Adamw optimizer, where we set the base learning rate at 2e-5 and the batch size at 10. The dropout rate in experiment is 0.5.

5 RESULTS AND DISCUSSION

5.1 Overall Results

We conduct the experiments on the MNRE dataset. Table 2 shows the overall results on the test set of MNRE. We report accuracy, precision, recall and F1 value as the evaluation metrics. Compared to the traditional sequence-based CNN method [18], the distantly supervised RE model PCNN [30] achieves better results in all metrics. Since the MNRE dataset is collected with short social media texts, most words in training or testing set are novel words. In such case, a distantly supervised model will perform better with external KGs. However, the distantly supervised method will suffer the wrong labeling problem and the performance is restricted. Benefiting from the better generalization of pretraining language model representations, the MTB model [23] outperforms PCNN with a higher recall (64.46%) and F1 value (57.81%).

The other part of Table 2 is the performance of our MEGA model and its variants. All the variants of our methods outperform pevious text-based methods, which demonstrate the effectiveness of introducing visual information to supplement the missing text semantics. We use a pretrained scene graph parser to extract the fine-grained visual objects and their relations. Compared to simply concatenation of visual and textual features, a added semantic similarity module with attention mechanism will contribute to an improved recall value. We propose a more efficient alignment method which considers both structural and semantic similarity. Our model can align the visual and textual relations precisely and find more possible textual relations. As a result, the final MEGA model improves the precision and recall value (from 62.65% to 68.44%) in a large margin.

Table 2: The Overall Performance of Our Models and Other State-of-the-art Methods (Acc.: Accuracy, Prec.: Precision). * Indicates the Difference Against the F1 of Our Baseline Variant (MTB) is Statistically Significant by One-Tailed Paired t-test with $p < 0.01$.

Model	Acc.	Prec.	Recall	F1
Glove+CNN [18]	70.32	57.81	46.25	51.39
PCNN [30]	72.67	62.85	49.69	55.49
MTB [23]	72.73	64.46	57.81	60.96
BERT+SG	74.09	62.95	62.65	62.80*
BERT+SG+Att.	74.59	60.97	66.56	63.64*
MEGA	76.15	**64.51**	**68.44**	**66.41***

5.2 Performance on Categories

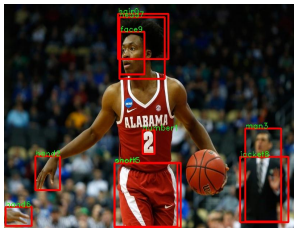
We also report the category results of our MEGA model compared to MTB model [23] in Table 3. Our model gains the highest results on all the six main categories on the MNRE test set. These categories involve the relations of person-to-person, person-to-organization or person-to-misc. Our model achieves relatively higher improvement in relation "Peer" and "Present_in". The two relations cover abundant visual information like "wearing the same uniform" or "appearing in a dance show". Our model introduces the visual information and utilize the mapping from visual relations to textual contents to help model extract relations precisely. However, text-based methods perform poor in these categories due to a lack of text contexts.

Table 3: Our Results on Six Main Categories Compared to MTB [23] on the MNRE Test Set.

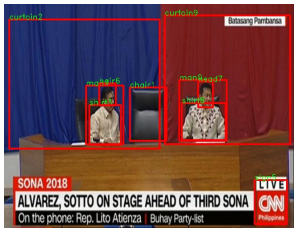
Category	Count	MEGA (Acc.)	MTB (Acc.)
Peer	156	**76.28**	63.46
Member_of	110	**70.90**	63.63
Contain	99	**91.91**	88.89
Present_in	74	**74.32**	51.35
Locate_at	46	**45.65**	41.30
Place_of_residence	29	**37.93**	31.03

5.3 Parameter Sensitivity

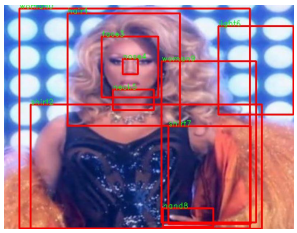
Table 4 describes the results of our proposed MEGA model influenced by choosing different number of aligned relations. Top-1 indicates that for each word in a sentence, the most related visual object will be chosen. As mentioned in Section 3, we leverage the structural and semantic similarity to align the visual and textual features. However, there may be more than one visual relations related to textual contents. For example, we can ensure that two people are



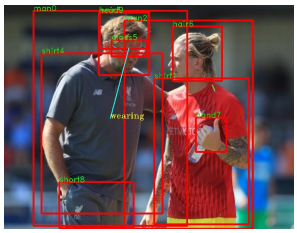
Cavs hosted a workout for Alabama guard Collin Sexton today , per @mcten
Selected objects:
man, shorts, jacket, number
Text-based MTB Model:
<Collin Sexton, place_of_residence, Alabama>
Our MEGA Model:
<Collin Sexton, member_of, Alabama>



Alvarez , Sotto on stage ahead of third State of the Nation Address . # SONA2018
Selected relations:
<man,wearing,shirt>,<man,wearing,shirt>,<man,in front of,curtain>
BERT+SG Model:
<Alvarez,None,Sotto>
Our MEGA Model:
<Alvarez,peer,Sotto>



Cecilia Bolocco with very famous trasparent dress . Vineyard of Sea Festival , 2016
Selected objects:
man,shirt,light
Text-based MTB Model:
<Cecilia Bolocco,None,Vineyard of Sea Festival>
Our Mega Model:
<Cecilia Bolocco, Present_in, Vineyard of Sea Festival>



Loris Karius is reported to feel let down by the signing of Alisson and wants to leave Liverpool .
Selected relations:
<man,wearing,shirt>,<man,wearing,shirt>
BERT+SG Model:
<Loris Karius,None,Alisson>
Our MEGA Model:
<Loris Karius,peer,Alisson>

Figure 5: The Results of Our Method (MEGA) Comparing to Text-based MTB [23] model and BERT+SG Model on the MNRE Test Set. Objects and Relations from Images are Detected in the Left Column, We Present the Relation Extraction Results with Related Objects and Visual Relations in the Right Column. The GroundTruth Labels are in Blue and the Detected Objects or Relations are in Green. Our Model Extracts Relations Precisely with Efficient Alignment between Images and Texts.

alumni with both "person wearing school uniform" and "person at the same school gate" visual contents. We find that in MNRE dataset, choosing the aligned relations with highest confidence will contribute to the best performance.

Table 4: The Performance of Our MEGA Model on the MNRE Test Set Influenced by Different Number of Aligned Relations.

Aligned Relation Num.	Prec.	Recall	F1
Top-1	**64.51**	**68.44**	**66.41**
Top-5	64.13	64.53	64.33
Top-10	62.65	64.89	63.75

5.4 Case Study

Figure 5 shows the case study of comparing our MEGA model with the text-based MTB model [23] and BERT+SG model. With the help of efficient alignment between visual and textual relations, our model performs better in all cases. To evaluate the effectiveness of utilizing visual information, we compare our model with MTB model which only depends on textual information. On the left side of Figure 5, our model extract the relation "member_of" correctly with the guidance of visual objects "man, shorts, number". These objects indicate that the man is a player which is the member of a team. However, without the guidance of visual information, text-based method extracts the wrong relation "place_of_residence". Similarly, our model extracts the relation "present_in" with the guidance of visual objects "man, shirt and light" while the text-based method identifies it as "no relation".

On the right side of Figure 5, we compare our MEGA model with a variant of our model BERT+SG. BERT+SG model simply concatenates the visual and textual representations and ignores the mappings from visual relations to textual contents. For example, the BERT+SG model cannot classify the correct relation "peer", however, our MEGA model finds the two people wearing the same uniform and extract the relation with alignment of visual and textual relations.

6 CONCLUSION

In this paper, we present the multimodal relation extraction (MRE) task which leverages visual information to supplement the missing textual semantics in social media posts. To tackle this problem, we first provide a human-annotated dataset - MNRE which consists of 15000+ sentences with 23 relation categories. Then, we propose a multimodal relation extract neural network with efficient graph alignment (MEGA). MEGA uses graph-structured visual information to guide the extraction of textual relations with considering both structural and semantic graph similarity. The experimental results demonstrate that our model outperforms previous state-of-the-art methods in terms of precision, recall and F1 values.

ACKNOWLEDGMENTS

The work presented in this paper has been supported by Hong Kong Research Grants Council through a General Research Fund (project no. PolyU 11204919), National Natural Science Foundation of China (62076100), and Fundamental Research Funds for the Central Universities, SCUT (D2210010,D2200150,and D2201300), the Science and Technology Planning Project of Guangdong Province (2020B0101100002)

REFERENCES

[1] Peter Anderson, X. He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 6077–6086.

[2] Isabelle Augenstein, Diana Maynard, and Fabio Ciravegna. 2016. Distantly supervised web relation extraction for knowledge base population. *Semantic Web* 7, 4 (2016), 335–349.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[4] Gregory Brown. 2011. An error analysis of relation extraction in social media documents. In *Proceedings of the ACL 2011 Student Session*. 64–68.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[6] Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention Guided Graph Convolutional Networks for Relation Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 241–251.

[7] Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction. In *Proceedings of EMNLP-IJCNLP: System Demonstrations*. 169–174. https://doi.org/10.18653/v1/D19-3029

[8] Mark Heimann, H. Shen, Tara Safavi, and Danai Koutra. 2018. REGAL: Representation Learning-based Graph Alignment. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (2018).

[9] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. 33–38.

[10] Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Aligned Dual Channel Graph Convolutional Network for Visual Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7166–7176.

[11] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, K. Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *HLT-NAACL*.

[12] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10313–10322.

[13] Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-Relation Extraction as Multi-Turn Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1340–1350.

[14] Zuoguo Liu and Xiaorong Chen. 2020. Research on relation extraction of named entity on social media in smart cities. *Soft Computing* 24, 15 (2020), 11135–11147.

[15] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1990–1999.

[16] Xuezhe Ma and E. Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *ArXiv* abs/1603.01354 (2016).

[17] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 1003–1011.

[18] Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. 39–48.

[19] Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3661–3672.

[20] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics* 5 (2017), 101–115.

[21] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*.

[22] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2015), 1137–1149.

[23] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2895–2905.

[24] Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. N-ary Relation Extraction using Graph-State LSTM. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2226–2235.

[25] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased Scene Graph Generation From Biased Training. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 3713–3722.

[26] Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction. In *Proceedings of NAACL-HLT*. 872–884.

[27] Linlin Wang, Zhu Cao, Gerard De Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1298–1307.

[28] Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Multimodal Representation with Embedded Visual Guiding Objects for Named Entity Recognition in Social Media Posts. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1038–1046.

[29] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research* 3, Feb (2003), 1083–1106.

[30] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 1753–1762.

[31] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*. 2335–2344.

[32] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive Co-attention Network for Named Entity Recognition in Tweets.. In *AAAI*. 5674–5681.

[33] Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2205–2215.

[34] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 35–45.

[35] Changmeng Zheng, Zhiwei Wu, Tao Wang, Cai Yi, and Qing Li. 2020. Object-aware Multimodal Named Entity Recognition in Social Media Posts with Adversarial Learning. *IEEE Transactions on Multimedia* (2020).