# Query Prior Matters: A MRC Framework for Multimodal Named Entity Recognition

Meihuizi Jia
jmhuizi24@bit.edu.cn
Beijing Institute of Technology
JD AI
Beijing, China

Xin Shen
u6498962@anu.edu.au
Australian National University
Canberra, Australia

Lei Shen
shenlei20@jd.com
JD AI
Beijing, China

Jinhui Pang*, Lejian Liao
{pangjinhui,liaolj}@bit.edu.cn
Beijing Institute of Technology
Beijing, China

Yang Song, Meng Chen
{songyang23,chenmeng20}@jd.com
JD AI
Beijing, China

Xiaodong He
xiaodong.he@jd.com
JD AI
Beijing, China

## ABSTRACT

Multimodal named entity recognition (MNER) is a vision-language task where the system is required to detect entity spans and corresponding entity types given a sentence-image pair. Existing methods capture text-image relations with various attention mechanisms that only obtain implicit alignments between entity types and image regions. To locate regions more accurately and better model cross-/within-modal relations, we propose a machine reading comprehension based framework for MNER, namely MRC-MNER. By utilizing queries in MRC, our framework can provide prior information about entity types and image regions. Specifically, we design two stages, Query-Guided Visual Grounding and Multi-Level Modal Interaction, to align fine-grained type-region information and simulate text-image/inner-text interactions respectively. For the former, we train a visual grounding model via transfer learning to extract region candidates that can be further integrated into the second stage to enhance token representations. For the latter, we design text-image and inner-text interaction modules along with three sub-tasks for MRC-MNER. To verify the effectiveness of our model, we conduct extensive experiments on two public MNER datasets, Twitter2015 and Twitter2017. Experimental results show that MRC-MNER outperforms the current state-of-the-art models on Twitter2017, and yields competitive results on Twitter2015.

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

## KEYWORDS

multimodal named entity recognition, machine reading comprehension, visual grounding, transfer learning

*Corresponding author.


*MM '22, October 10–14, 2022, Lisboa, Portugal*

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548427

**ACM Reference Format:**
Meihuizi Jia, Xin Shen, Lei Shen, Jinhui Pang*, Lejian Liao, Yang Song, Meng Chen, and Xiaodong He. 2022. Query Prior Matters: A MRC Framework for Multimodal Named Entity Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548427

## 1 INTRODUCTION

Nowadays, multimodal named entity recognition (MNER) has attracted extensive attention of researchers as it extends the traditional text-based NER and alleviates ambiguity in natural language with the help of auxiliary images. Given a sentence-image pair,

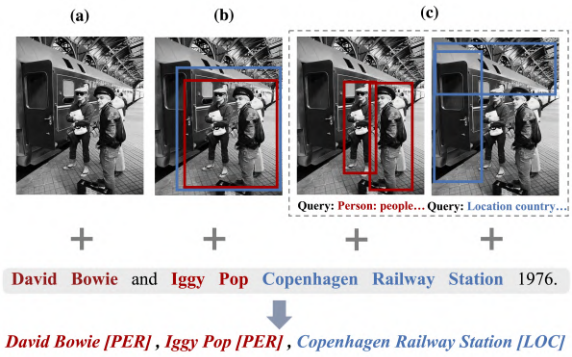


**Figure 1: An example for multimodal named entity recognition with (a) the whole image cue, (b) region cues from VG toolkit, (c) region cues paired with queries for entity types.**

MNER is required to recognize named entities of different types (mainly persons, locations, and organizations labeled as PER, LOC, and ORG respectively) in the sentence with extra image assistance. As shown in Figure 1 (a), two people in the front and train platform in the back of the image provide useful information to recognize named entities "*David Bowie*" (PER), "*Iggy Pop*" (PER), and "*Copenhagen Railway Station*" (LOC).

Existing MNER datasets usually contain few fine-grained annotations in each sentence-image pair, i.e., the relevant image is given as a whole without manually-labeled region signals for a particular

entity type. Therefore, previous works implicitly align contents inside a sentence-image pair and fuse their representations based on various attention mechanisms [1, 4, 17, 19, 32, 37, 40]. However, it is hard to interpret and evaluate the effectiveness of implicit alignments of entity types and their image regions. Recently, researchers [39] exploit a visual grounding toolkit [34] to ground a phrase or sentence on its image region. The grounded regions of different entity types are then bound with the entire sentence and fed into the recognition model together (as shown in Figure 1 (b)). That is, the explicit relations of each type-region pair are still not utilized. In addition, the data in tasks like MNER and visual grounding is biased, which leads to inaccurate results of region detection.

Existing works formalize MNER as a sequence labeling task that integrates image embeddings into a sequence labeling model and assigns type labels to named entities. Recently, the MRC framework is employed in many natural language tasks due to its solid language understanding capability [5, 13, 14]. Similarly, to take advantage of the prior knowledge encoded in MRC queries, e.g., priors of entity types [13], we regard MNER as a machine reading comprehension (MRC) task. MRC queries are expected to (1) ground a specific entity type to its relevant region(s), which achieves explicit alignments of entity types and image regions (as shown in Figure 1 (c)); (2) provide guidance to model text-image and inner-text interactions on cross-modal and within-modal levels. Then named entities and their entity types are converted into answer spans and pre-defined queries accordingly. For example, recognizing entities with type LOC in sentence "David Bowie and Iggy Pop Copenhagen Railway Station 1976" is formalized as extracting answer spans to the query "Location: Country, city..." from the given sentence. Details of transforming different entity types to queries will be described in Section 3.

To tackle the above-mentioned issues and make full use of informative query priors, we propose a **M**achine **R**eading **C**omprehension based framework for **M**ulti-modal **N**amed **E**ntity **R**ecognition (**MRC -MNER**), which consists of two stages, Query-Guided Visual Grounding and Multi-Level Modal Interaction. First, we train a visual grounding model via transfer learning on our newly constructed corpus to achieve MNER adaptation. The model returns the top-$k$ region candidates with their confidence scores for the input query. We then use Multi-Level Modal Interaction to model cross-modal and within-modal relations with three groups of sub-tasks. Specifically, we design two new sub-tasks, namely region weights estimation and existence detection, to facilitate entity span prediction. The former aims to import region candidates dynamically, and the latter provides a global judgement about whether named entities of the given types exist in the input sentence. Finally, the Multi-Level Modal Interaction model is trained to recognize entities with a joint loss under the multi-task scheme.

In summary, the main contributions of this paper are three-fold:

- We propose a novel MRC-based framework for multimodal named entity recognition (MRC-MNER), which unifies visual grounding and MRC by designing queries with prior information of entity types.
- We train a query-guided visual grounding model via transfer learning to extract region candidates, thus achieving fine-grained alignments of regions and entity types. At the same time, we design a Multi-Level Modal Interaction model that conducts region weights estimation, existence detection, and entity span prediction simultaneously.
- We conduct extensive experiments on two public MNER datasets, Twitter2015 [40] and Twitter2017 [17], to evaluate the performance of our MRC-MNER framework. Experimental results show that MRC-MNER outperforms the current state-of-the-art models on Twitter2017 and yields competitive results on Twitter2015.

## 2 RELATED WORK

### 2.1 Multimodal Named Entity Recognition

As a crucial component of natural language processing, named entity recognition (NER) aims to discover named entities in free text and classify them into predefined types [11, 12, 18, 33]. With multimodal data emerging in the various task, images as auxiliary information assist the NER model in better identifying the entities contained in the text. The critical challenge of MNER is aligning and fusing text and image information. [17] proposed a gated mechanism for MNER to model the cross-modal interactions. [37] proposed a multimodal transformer architecture to acquire expressive text-image representation by incorporating the auxiliary entity span detection. [4] integrated both image attributes and image knowledge to improve model performance for MNER. [39] employed a visual grounding toolkit to extract the top-1 image region and proposed a graph fusion approach based on a graph model to obtain text-image representation. [32] proposed a matching and alignment framework for MNER to alleviate the impact of mismatched text-image pairs on encoding. The above methods treat the MNER task as a sequence labeling problem. Due to a lack of prior information of entity types in sequence labeling, image information (whole images, equally regions of images, or the retrieved visual regions) with the entire sentence is fed into the entity recognition model together, which cannot realize the explicit alignment of image and text.

### 2.2 Machine Reading Comprehension

Machine Reading Comprehension (MRC) requires answering specific queries by searching for relevant information in natural language contexts. The task of text span extraction can be executed to two multi-class classification or two binary classification tasks. For the former, the model needs to predict the start and end positions of the answer. For the latter, the model needs to decide whether each token is the start/end position. In previous work, Recurrent Neural Network (RNN) was adopted to encode query and context, multiple RNN and linear projection were stacked to predict answer span in stages [3, 20, 35]. The performance was boosted after the large-scale pre-training model was released [24, 30], such as ELMo [21], BERT [6], RoBERTa [16]. Recently, there is a tendency to employ MRC on variety of NLP tasks, including named entity recognition [13], entity relation extraction [14], sentiment analysis [5]. Our work is inspired by [13], which formalized the task of NER as a single-turn question answering task. Different from this work, we design queries with prior information of entity types to bridge visual grounding and MRC.

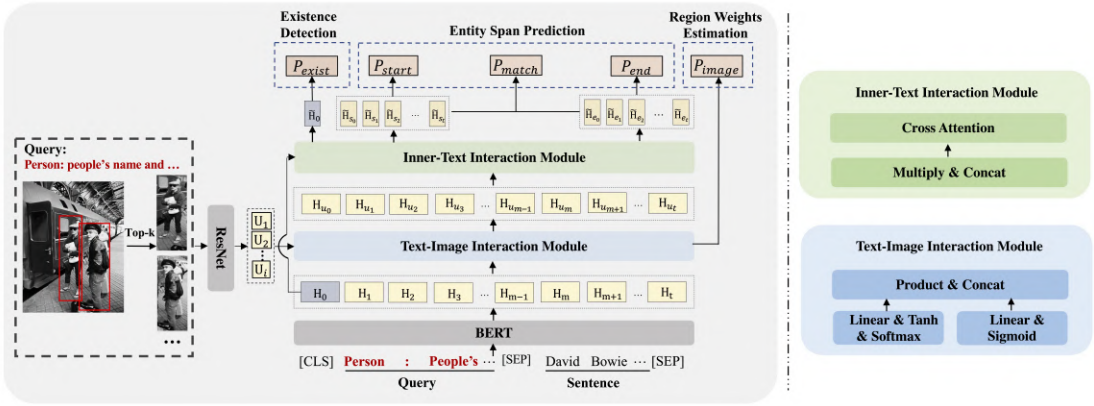


**Figure 2: Overview of our MRC-MNER framework. The details of Multi-Level Modal Interaction are illustrated on the right.**

## 2.3 Visual Grounding

Visual grounding is a vision-language task, which aims to ground a natural language phrase or sentence about an image onto a correct region of the image [34]. That is, the system takes an image and text as input, and outputs the corresponding bounding box (region). Frameworks of visual grounding are either two-stage or one-stage. In the two-stage methods, the first stage is used to propose region candidates through some region proposal methods (e.g., Edgebox [41], selective search [31], and Region Proposal Networks [27]), and then the second stage is designed to rank those candidates based on their similarities with the input text. The one-stage methods utilize one-stage models (e.g., YOLO [25]) combined with extra features to directly output the final region(s). [34] built a one-stage model based on the YOLOv3 object detector [26] by integrating additional spatial features. The model is 10 times faster than state-of-the-art two-stage methods and achieves superior grounding accuracy.

In our work, queries are tokens or phrases related to entity types, and contain few complex spatial and logical information. To better fit the MNER task, we construct a corpus and conduct a domain and task adaptation based on transfer learning to finetune the pre-trained model released by [34].

# 3 METHOD

## 3.1 Task Formalization

Given a sentence $X = \{x_1, x_2, ..., x_n\}$ and its associated image $V$ as input, where $n$ denotes the length of the sentence, the goal of MNER is to find a set of entities from $X$ with the assistance of images and classify each entity into one of the pre-defined types. Previous works on MNER usually formulate the task as a sequence labeling problem. Let $y = (y_1, y_2, ..., y_n)$ denotes a label sequence corresponding to $X$, where $y_i \in$ Y and Y is the pre-defined label set with the BIO tagging schema [29]. Inspired by [13], in this work, we propose to apply the MRC framework to the MNER task, which can take advantage of the query prior information and solid language understanding capability of MRC.

**Data Preparation.** First, we need to transform the MNER data to the form of MRC, i.e., a set of (QUESTION, ANSWER, CONTEXT, IMAGE) quadruples. Each label of the entity type $y_t \in Y_t$ is connected with a natural language query $q_y = \{q_1, q_2, ..., q_m\}$ which we construct, where $m$ denotes the length of the query. The entity $x_{\text{start,end}}$ is a substring in sentence $X = \{x_{start}, x_{start+1}, ..., x_{end-1}, x_{end}\}$, which satisfies start $\leq$ end. At last, we obtain the quadruples $(q_y, x_{\text{start,end}}, X, V)$ which corresponds to (QUESTION, ANSWER, CONTEXT, IMAGE).

**Table 1: Examples for transforming entity types to queries.**

| Entity Type | Natural Language Query |
|---|---|
| PER (Person) | Person: People's name and fictional character. |
| LOC (Location) | Location: Country, city, town continent by geographical location. |
| ORG (Organization) | Organization: Include company, government party, school government, and news organization. |

**Query Construction.** Query plays a significant role in our MRC-MNER since it provides prior information about labels and image regions, so it should be described as generic, precise, and effective as possible. Inspired by [13], we design several different forms of queries. Different from this work, the queries in our model need to unify the two tasks of MRC and VG effectively. The amount of information in the query should be weighted. Less amount of information in query will limit the powerful understanding ability of MRC, while an excess of information will increase the difficulty of VG. Therefore, we design the appropriate query for each entity. Examples are shown in Table 1. In section 4.4, we specially construct relevant experiments for different query construction methods.

## 3.2 Query-Guided Visual Grounding

Visual grounding aims to detect the relevant visual regions given the input query. In this section, we revise a small amount of data on visual grounding using queries that we construct in Section 3.1 and train a well-matched visual grounding model via transfer learning to extract region candidates. As shown in the black dotted box in the left part of Figure 2, top-$k$ image region candidates are retrieved with the assistance of the query "Person: People's name and fictional character...".

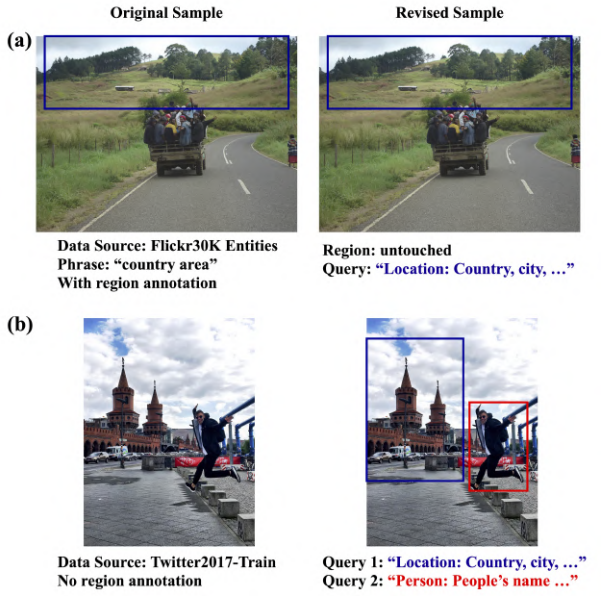


**Figure 3: Illustration of Corpus Construction. (a) Replacing original phrases in the Flickr30K Entities dataset with MRC queries. (b) Labelling existing regions related to PER, LOC, and ORG in the Twitter2015/2017-Train dataset.**

We apply the pre-trained fast and accurate one-stage visual grounding model [34] (denoted as FA-VG) as our based model. In the setting of Phrase Localization task, FA-VG was trained and evaluated on the Flickr30K Entities dataset [22] that augments the original Flickr30K [36] with region-phrase correspondence annotations. However, there are two obstacles: (1) These phrases/queries are from image captions, and not particularly designed for the named entity recognition task. (2) The widely-used MNER datasets (i.e. Twitter2015 [40] and Twitter2017 [17]) have different data domains compared with the Flickr30K Entities dataset. Thus, we utilize transfer learning to overcome above issues.

**Corpus Construction.** To fulfill MNER adaptation of the pre-trained FA-VG model, we construct a corpus consisting of three sets of samples:

(1) Samples from the Flickr30K Entities dataset with phrases highly-related to pre-defined PER, LOC, ORG, and OTHER queries.
(2) Samples from (1) with phrases replaced by MNER queries.
(3) Samples from the Twitter2015/2017 dataset with manually-labeled regions of PER, LOC, and ORG types.

Since only a small part of phrases in the Flickr30K Entities dataset are related to MNER entity types, we first filter the original data to get those highly-relevant samples, and replace phrases in them with MNER queries. Specifically, all phrases in the Flickr30K Entities dataset and four MRC queries in Table 1 are represented by BERT embeddings [6] respectively. Then we calculate cosine similarities between embeddings of phrases and each query, and only keep samples with scores larger than a threshold, e.g., 0.7. To obtain samples in the second set, we make a copy of the first set and conduct query replacement. As shown in Figure 3(a), the original query "country area" is replaced by "Location: Country, city..." that is defined as the MRC query for LOC in Table 1. To take advantage of some in-domain data, we randomly sample 1000+ images from the training set of Twitter2015/2017 dataset, and manually annotate regions related to PER, LOC, and ORG. Take Figure 3(b) as an example. The image is labeled with two pairs of regions and queries: red box with "Person: People's name..." and blue box with "Location: Country, city...". The statistics of the constructed corpus are summarized in Table 2. We split the corpus into training/validation/test set with the ratio of 9:0.5:0.5. During training, all sets of samples are shuffled so that the model can be finetuned to not only maintain the ability of accurate visual grounding, but also adapt to new task and domain.

**Table 2: Statistics of our constructed VG corpus (F.30k and Tw.15/17 denote Flickr30k and Twitter2015/2017, respectively and b.-box denotes bounding box).**

| | |
|---|---|
| Total data volume | 26,311 |
| F.30K data (unmodified) | 12,504 |
| F.30K data + modified query data | 12,504 |
| Tw.15/17 data + query + b.-box | 1,303 |
| LOC query data | 2,983 (F.30K) + 700 (Tw.15/17) |
| ORG query data | 4,191 (F.30K) + 350 (Tw.15/17) |
| PER query data | 4,362 (F.30K) + 253 (Tw.15/17) |
| MISC query data | 968 (F.30K) |

**Training Procedure.** FA-VG uses Darknet-53 [26] with feature pyramid networks [15] to extract visual features for the input image. For the text query, FA-VG embeds it to a 768D real-valued vector using the uncased version of BERT [6]. Then it makes predictions based on the fused representations of image, text, and spatial[1] features. After being finetuned on the constructed corpus, FA-VG achieves a grounding accuracy (IoU>0.5) of 79.96% on the test set[2], which significantly outperforms the pre-trained FA-VG model without MNER adaptation (64.72%). The results show that our query-guided visual grounding module can reach a better localization of regions related to entity types. Finally, FA-VG returns top-$k$ region candidates with their confidence scores $Y_{image}$ from the softmax function of the output layer. Both regions and confidence scores are further utilized in the Multi-Level Modal Interaction model.

## 3.3 Multi-Level Modal Interaction

In this section, we introduce the details of our proposed framework MRC-MNER. Figure 2 illustrates the overall architecture of our model. The input of MRC-MNER contains two modalities, including the text-modality (query and sentence), and image-modality (the detected top-$k$ visual regions from VG in Section 3.2). Without loss of generality, we leverage the popular pretrained language model BERT [6] and ResNet [8] to encode the text and images. Then we devise two novel multi-level interaction modules (text-image interaction and inner-text interaction) to fuse the information from two modalities seamlessly, and enhance the representation for each

[1]The spatial feature captures the coordinates of the top-left corner, center, and bottom-right corner of the grid at $(i, j)$.
[2]IOU (Intersection over Union) is a term used to describe the extent of overlap of two boxes.

token in the sentence. To facilitate the final entity span detection task, we further leverage the multi-task training scheme including two auxiliary sub-tasks of region weights estimation and existence detection. We will explain each module in the following sections.

*3.3.1 Text and Image Representations.* Here, we encode text and images to obtain the text representation and visual representation, respectively. For text representation, we employ the pre-trained BERT model [6] as encoder, and generate an input to feed through BERT by concatenating { [CLS], $q_1, q_2, ..., q_m$, [$SEP$], $x_1, x_2, ..., x_n$, [SEP]} where [CLS] and [SEP] are special tokens. Then BERT outputs a context representation matrix $\mathbf{H} \in \mathbb{R}^{t \times d_t}$ after receiving the combined string, where $d_t = 768$ is the vector dimension of the last layer of BERT and $t = n + m + 3$ is the length of the input to BERT. For visual representation, we use ResNet [8] as the image encoder, which achieves state-of-the-art on various visual tasks[7, 38]. Specifically, we first resize the image to $224 \times 224$ pixels, and obtain its visual representations from a pre-trained 152-layer ResNet. The extracted visual features are represented as $\mathbf{U}_i = ResNet{-}152\,(V)$, where $\mathbf{U}_i \in \mathbb{R}^{d_v}$, and $d_v = 2048$ is the dimension of visual representation. After receiving the text and image representations, we use a linear projection to map them to the same dimension $d = 512$.

*3.3.2 Text-Image Interaction.* As shown in Figure 2, first of all, we use the region candidates and the entire sentence to measure the overall matching degree. The model receives the $k$ region candidates representation $\mathbf{U}$ from ResNet and the "[CLS]" representation $\mathbf{H}_0$ from BERT as input and outputs the relevance between them.

$$\mathbf{Z}_{img} = \tanh\left(\mathbf{W}_g\,[\mathbf{H}_0; \mathbf{U}]^\top\right), \quad \alpha_{img} = \text{softmax}\left(\mathbf{Z}_{img}\right) \tag{1}$$

where $\mathbf{W}_g \in \mathbb{R}^d$, $\mathbf{H}_0 \in \mathbb{R}^d$, and $\mathbf{U} \in \mathbb{R}^{k \times d}$. We use $[;]$ to denote the concatenation of the representations of whole sentence and candidate regions. The concatenation between a matrix and a vector is performed by concatenating each column of the matrix with the vector. And then, we carry out the fine-grained fusion between top-$k$ region candidates and sentence, and obtain the cross-modal representation for each token.

We get the sentence representation $\mathbf{H}$ from the BERT, and follow [32, 37] and use a naive gate mechanism to control the combination of text and top-$k$ region candidates at the token-level.

$$g = \text{sigmod}\left(\mathbf{W}_f\,[\mathbf{H}; \mathbf{U}]\right), \quad \widetilde{g} = g * \alpha_{img} \tag{2}$$

where $\mathbf{W}_f \in \mathbb{R}^{d \times 2d}$. We update the above gate score $g$ using the top-$k$ region weights score $\alpha_{img}$, which aims to weigh the correlation between top-$k$ region candidates and the whole sentence.

Finally, we update region representation $\mathbf{U}$ as follows and obtain the updated sentence representation $\mathbf{H}_u$ that incorporates image information:

$$\widetilde{\mathbf{U}} = \widetilde{g} \odot \mathbf{U}, \quad \mathbf{H}_u = \left(\mathbf{W}_u\left[\mathbf{H}; \widetilde{\mathbf{U}}\right]\right) \tag{3}$$

where $\odot$ is the element-wise product, $\mathbf{W}_u \in \mathbb{R}^{d \times 2d}$ and $\mathbf{H}_u \in \mathbb{R}^{t \times d}$.

*3.3.3 Inner-Text Interaction.* We specially design a global signal to determine whether the sentence contains the entity asked by the current query. This signal interacts with sentence representation and can mutually reinforce each other. We assume that if the sentence contains entities, and the model should be more inclined to tag the entity span; on the contrary, if the model extracts the entity span from the sentence, then the global signal tends to determine the sentence contains entities. We follow [12, 23] and apply a label attention network to update sentence representation with start label and end label information as well as global signal representation. Finally we get label-enhanced contextual information $\mathbf{H}_s$, $\mathbf{H}_e$ and label-enhanced global signal representation $\widehat{\mathbf{H}}_0$. We map the matrices of $\mathbf{H}_s, \mathbf{H}_e, \widehat{\mathbf{H}}_0$ to queries $(\mathbf{Q}_s, \mathbf{Q}_e, \mathbf{Q}_0)$, keys $(\mathbf{K}_s, \mathbf{K}_e, \mathbf{K}_0)$, and values $(\mathbf{V}_s, \mathbf{V}_e, \mathbf{V}_0)$ by using different linear projections. And then, we use the co-attention mechanism to calculate the attention scores, respectively, between the label-enhanced start position information $\mathbf{H}_s$ and global signal $\widehat{\mathbf{H}}_0$, and between $\mathbf{H}_e$ and $\widehat{\mathbf{H}}_0$. Finally, we acquire the updated entity span representation and the global signal representation.

$$\mathbf{C}_s = \text{softmax}\left(\frac{\mathbf{Q}_s\mathbf{K}_0^\top}{\sqrt{d_k}}\right)\mathbf{V}_0, \quad \widetilde{\mathbf{H}}_s = \text{LN}\left(\mathbf{H}_s + \mathbf{C}_s\right) \tag{4}$$

where LN denotes the layer normalization function [2], $\mathbf{Q}_s \in \mathbb{R}^{t \times d}$, $\mathbf{K}_0, \mathbf{V}_0 \in \mathbb{R}^{1 \times d}$, and $\widetilde{\mathbf{H}}_s \in \mathbb{R}^{t \times d}$. The above formula represents the updated entity span start representation. Similarly, we can obtain the updated entity span end representation $\widetilde{\mathbf{H}}_e \in \mathbb{R}^{t \times d}$, and the updated global signal representation $\widetilde{\mathbf{H}}_0 \in \mathbb{R}^{1 \times d}$, which only needs to replace the corresponding $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, respectively.

*3.3.4 Multi-Task Training.* In this section, we design three subtasks to facilitate the MNER problem.

**Region Weights Estimation.** Previous work [39] only extracts top-1 visual region for each entity type, which relies heavily on the accuracy of VG model. Differently, we have two improvements. First, we leverage the VG model to detect the top-$k$ visual regions with their confidence scores. Second, we encourage the model to estimate the weights of different visual regions during entity recognition, which can utilize the image information dynamically. Thus, we devise an auxiliary task named region weights estimation (RWE). Specifically, we obtain the confidence scores $\alpha_{img}$ from Section 3.3.2 and use them as the supervised probability distribution $\mathrm{P}_{image}$ for the detected visual regions. To train this sub-task, we minimize the Mean Square Error between the probability distribution $\mathrm{P}_{image}$ and the confidence scores $\mathrm{Y}_{image}$ of top-$k$ region candidates as follows.

$$\mathcal{L}_{image} = \text{MSE}\left(\mathrm{P}_{image}, \mathrm{Y}_{image}\right) \tag{5}$$

**Existence Detection.** Considering not every utterance contains the specific entity in the dataset, it's necessary to design an auxiliary task to predict whether there exists the specific entity type in the utterance, which can be an effective indicator for the final entity recognition task intuitively. Specifically, the global signal representation $\widetilde{\mathbf{H}}_0$ from Section 3.3.3 is utilized for the existence detection (ED) task. The ED task and the entity span prediction task can share the corresponding mutual information with the co-interactive attention mechanism. Here, we detect of the existence of entity as follows:

$$\mathrm{P}_{exist} = \text{softmax}\left(\widehat{\mathbf{H}}_0\mathbf{W}_{exist}\right) \tag{6}$$

where $\mathbf{W}_{exist} \in \mathbb{R}^{d \times 2}$ and $\mathrm{P}_{exist} \in \mathbb{R}^{1 \times 2}$. The cross-entropy loss is taken as the training objective:

$$\mathcal{L}_{exist} = \text{CE}\left(\mathrm{P}_{exist}, \mathrm{Y}_{exist}\right) \tag{7}$$

where $Y_{exist}$ is the golden label for whether the sentence contains entities asked by a query.

**Entity Span Prediction.** To tag the entity span from a sentence, it is necessary to find the start and end positions of the entity. Two binary classifiers are exploited respectively. One is to predict whether each token is the start index or not, and the other is to predict whether each token is the end index or not. Hence, given the label-enhanced entity span start and end representations, the model then predicts the probability of each token being a start position as follows:

$$\mathrm{P}_{start} = \mathrm{softmax}_{\text{each row}}\left(\widetilde{\mathbf{H}}_s \mathbf{W}_s\right) \tag{8}$$

where $\mathbf{W}_s \in \mathbb{R}^{d\times 2}$ and $\mathrm{P}_{start} \in \mathbb{R}^{t\times 2}$. Similarly, we can predict the probability distribution of the end position index $\mathrm{P}_{end} \in \mathbb{R}^{d\times 2}$.

During training, $\mathrm{Y}_{start}$ and $\mathrm{Y}_{end}$ are two label sequences of length $t$, representing the ground-truth label of each token in sentence $X$, which is the start or end position of any entity. The cross-entropy loss is used for this task:

$$\mathcal{L}_{start} = \mathrm{CE}\left(\mathrm{P}_{start}, \mathrm{Y}_{start}\right), \qquad \mathcal{L}_{end} = \mathrm{CE}\left(\mathrm{P}_{end}, \mathrm{Y}_{end}\right) \tag{9}$$

As there could exist multiple entities from the same query in the sentence, we follow [13] and append a binary classification model to predict the matching probability of start and end positions.

$$\mathrm{P}_{match} = \mathrm{sigmoid}\left(\mathbf{W}_m\left[\widetilde{\mathbf{H}}_s; \widetilde{\mathbf{H}}_e\right]\right) \tag{10}$$

where $\mathbf{W}_m \in \mathbb{R}^{1\times 2d}$. And the cross-entropy loss is used to train this task:

$$\mathcal{L}_{match} = \mathrm{CE}\left(\mathrm{P}_{match}, \mathrm{Y}_{match}\right) \tag{11}$$

where $\mathrm{Y}_{match}$ denotes the ground-truth label for whether each start index should be matched with each end index.

**Joint Training.** There are three sub tasks in our proposed MRC-MNER model: Region Weights Estimation, Entity Span Prediction and Existence Detection. Hence, the whole model can be jointly trained, and the final loss function is defined as follows:

$$\mathcal{L} = \lambda_1\mathcal{L}_{start} + \lambda_2\mathcal{L}_{end} + \lambda_3\mathcal{L}_{match} + \lambda_4\mathcal{L}_{exist} + \lambda_5\mathcal{L}_{image} \tag{12}$$

where $\lambda_1$-$\lambda_5$ are hyper-parameters to control the contributions of each sub-tasks.

# 4 EXPERIMENTS

## 4.1 Experimental Settings

**Datasets.** We evaluate our framework on two widely-used MNER datasets, Twitter2015 [40] and Twitter2017 [17]. Both datasets contain four entity types: Person (PER), Organization (ORG), Location (LOC) and Others (OTHER) for sentence-image pairs. The datasets are separated into training, validation, and test sets with the same type distribution.

**Evaluation Metrics.** Following the previous work [4, 32, 37, 39], we exploit precision (*Pre.*), recall (*Rec.*), and F1 score (*F1*) to evaluate the performance of named entity recognition for overall entity types, and use F1 score (*F1*) only for each type.

**Implementation Details.** We first employ the pre-trained uncased $\mathrm{BERT}_{base}$ model [6] with dimension of 768 and the pre-trained ResNet152 with dimension of 2048 to get the initial representations of text tokens and images, respectively. Then these representations are transformed to 512D with a linear projection. The learning rate and dropout rate are set to 5e-5 and 0.3, which obtains the best performance on the validation set of two MNER datasets after conducting a grid search over the interval [1e-5, 1e-4] and [0.1, 0.6]. For the joint training loss, we set the hyper-parameters $\lambda_1$=$\lambda_2$=$\lambda_3$=$\lambda_5$=0.17 and $\lambda_4$=0.33 by tuning on the validation set. The entire model is trained on one Tesla P40 GPUs with pytorch 1.7.

**Baseline Models.** We compare two groups of baselines with our approach. The first group consists of some text-based NER models that formalize NER as a sequence labeling task: (1) **BiLSTM-CRF**[9], **CNN-BiLSTM-CRF** [18], **HBiLSTM-CRF** [10], which are the NER model with a bidirectional LSTM layer and a CRF layer, and the difference is they use different encoding methods to acquire character-level embedding and fuse with word-level embedding. (2) **BERT** [6], **BERT-CRF**, which use powerful encoder BERT compared with aforementioned methods. (3) **T-NER** [28, 40], which is a NER model designed specifically for tweet. It exploits broadly used features, including the dictionary, contextual and orthographic features. Besides, we compare several competitive multimodal NER models: (1) **GVATT-HBiLSTM-CRF** [17], **GVATT-BERT-CRF** [37], which use different encoder to obtain the text representation, but all propose a visual attention mechanism to combine images with text and acquire text-aware image representation. (2) **AdaCAN-CNN-BiLSTM-CRF** [40], **AdaCAN-BERT-CRF** [37], which exploit different ways to encode the text, but all design an adaptive co-attention mechanism to integrate image and text. (3) **UMT-BERT-CRF** [37], **MT-BERT-CRF** [37], which propose a multimodal interaction module to acquire expressive text-visual representation, but the difference is whether the auxiliary entity span detection is incorporated into multimodal Transformer. (4) **ATTR-MMKG-MNER** [4], which integrates both image attributes and image knowledge into MNER model. (5) **UMGF** [39], which proposes graph fusion approach based on graph model to obtain text-visual representation. (6) **MAF** [32], which proposes a matching and alignment framework for MNER to alleviate the impact of mismatched text-image pairs on encoding. In addition, we also compare MRC-MNER with its two variants: MRC-MNER-Text and MRC-MNER-VG. The former uses text input only, while the latter is equipped with the vanilla VG without transfer learning.

## 4.2 Main Results

Table 3 shows the experimental results of MRC-MNER and our baselines. The upper results are from text-based models, and we notice that *BERT*-based models outperform *LSTM*-based models with a significant margin on both datasets, indicating that the advantage of pre-trained language models on this task. And then, we find that our MRC-MNER-Text (removing image information from MRC-MNER) achieves better performance than sequence labeling models, which verifies the value of prior knowledge in MRC queries and the powerful understanding ability of MRC.

Second, we compare the MNER models with their corresponding uni-modal baselines, such as GVATT-HBiLSTM-CRF vs. HBiLSTM-CRF, AdaCAN-CNN-BiLSTM-CRF vs. CNN-BiLSTM-CRF, and MRC-MNER vs. MRC-MNER-Text. We find that almost all multimodal models can significantly outperform their corresponding uni-modal competitors, indicating that the image information is helpful for the MNER task.

**Table 3: Performance comparison on two MNER datasets. We refer to the results of UMGF from [39] and other results from [32].**

| Methods | Twitter2015 | | | | | | | Twitter2017 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Single Type (*F1*) | | | | Overall | | | Single Type (*F1*) | | | | Overall | | |
| | **PER** | **LOC** | **ORG** | **OTH.** | *Pre.* | *Rec.* | *F1* | **PER** | **LOC** | **ORG** | **OTH.** | *Pre.* | *Rec.* | *F1* |
| BiLSTM-CRF | 76.77 | 72.56 | 41.33 | 26.80 | 68.14 | 61.09 | 64.42 | 85.12 | 72.68 | 72.50 | 52.56 | 79.42 | 73.43 | 76.31 |
| CNN-BiLSTM-CRF | 80.86 | 75.39 | 47.77 | 32.61 | 66.24 | 68.09 | 67.15 | 87.99 | 77.44 | 74.02 | 60.82 | 80.00 | 78.76 | 79.37 |
| HBiLSTM-CRF | 82.34 | 76.83 | 51.59 | 32.52 | 70.32 | 68.05 | 69.17 | 87.91 | 78.57 | 76.67 | 59.32 | 82.69 | 78.16 | 80.37 |
| BERT | 84.72 | 79.91 | 58.26 | 38.81 | 68.30 | 74.61 | 71.32 | 90.88 | 84.00 | 79.25 | 61.63 | 82.19 | 83.72 | 82.95 |
| BERT-CRF | **84.74** | 80.51 | **60.27** | 37.29 | 69.22 | 74.59 | 71.81 | 90.25 | 83.05 | 81.13 | 62.21 | 83.32 | 83.57 | 83.44 |
| T-NER | 83.64 | 76.18 | 59.26 | 34.56 | 69.54 | 68.65 | 69.09 | - | - | - | - | - | - | - |
| **MRC-MNER-Text (Ours)** | 84.72 | **81.13** | 60.07 | **39.23** | **76.35** | **69.46** | **72.74** | **91.33** | **85.23** | **81.75** | **68.41** | **87.12** | **84.03** | **85.55** |
| GVATT-HBiLSTM-CRF | 82.66 | 77.21 | 55.06 | 35.25 | 73.96 | 67.90 | 70.80 | 89.34 | 78.53 | 79.12 | 62.21 | 83.41 | 80.38 | 81.87 |
| AdaCAN-CNN-BiLSTM-CRF | 81.98 | 78.95 | 53.07 | 34.02 | 72.75 | 68.74 | 70.69 | 89.63 | 77.46 | 79.24 | 62.77 | 84.16 | 80.24 | 82.15 |
| GVATT-BERT-CRF | 84.43 | 80.87 | 59.02 | 38.14 | 69.15 | 74.46 | 71.70 | 90.94 | 83.52 | 81.91 | 62.75 | 83.64 | 84.38 | 84.01 |
| AdaCAN-BERT-CRF | 85.28 | 80.64 | 59.39 | 38.88 | 69.87 | 74.59 | 72.15 | 90.20 | 82.97 | 82.67 | 64.83 | 85.13 | 83.20 | 84.10 |
| MT-BERT-CRF | 85.30 | 81.21 | 61.10 | 37.97 | 70.84 | 74.80 | 72.58 | 91.47 | 82.05 | 81.84 | 65.80 | 84.60 | 84.16 | 84.42 |
| UMT-BERT-CRF | 85.24 | 81.58 | 63.03 | 39.45 | 71.67 | **75.23** | 73.41 | 91.56 | 84.73 | 82.24 | 70.10 | 85.28 | 85.34 | 85.31 |
| ATTR-MMKG-MNER | 84.28 | 79.43 | 58.97 | 41.47 | 74.78 | 71.82 | 73.27 | - | - | - | - | - | - | - |
| UMGF | 84.26 | **83.17** | 62.45 | **42.42** | 74.49 | 75.21 | **74.85** | 91.92 | 85.22 | 83.13 | 69.83 | 86.54 | 84.50 | 85.51 |
| MAF | 84.67 | 81.18 | **63.35** | 41.82 | 71.86 | 75.10 | 73.42 | 91.51 | 85.80 | **85.10** | 68.79 | 86.13 | **86.38** | 86.25 |
| **MRC-MNER-VG (Ours)** | 84.88 | 81.43 | 61.06 | 39.93 | 78.08 | 70.75 | 74.22 | 91.83 | 85.84 | 83.09 | 72.11 | 88.59 | 84.16 | 86.32 |
| **MRC-MNER (Ours)** | **85.71** | 81.97 | 61.12 | 40.20 | **78.10** | 71.45 | 74.63 | **92.64** | **86.47** | 83.16 | **72.66** | **88.78** | 85.00 | **86.85** |

**Table 4: Ablation study of MRC-MNER.**

| Methods | Twitter2015 | | | Twitter2017 | | |
|---|---|---|---|---|---|---|
| | *Pre.* | *Rec.* | *F1* | *Pre.* | *Rec.* | *F1* |
| MRC-MNER | 78.10 | 71.45 | 74.63 | 88.78 | 85.00 | 86.85 |
| w/o RWE | 77.24 | 70.86 | 73.91 | 88.11 | 84.26 | 86.14 |
| w/o ED | 77.63 | 70.95 | 74.14 | 88.29 | 84.36 | 86.29 |
| w/o RWE+ED | 76.82 | 70.24 | 73.38 | 87.72 | 83.99 | 85.81 |

**Table 5: Results with different query transformations.**

| **Query transformation** | Twitter2015 | Twitter2017 | Flickr30K |
|---|---|---|---|
| | *F1* | *F1* | *Accuracy* |
| Keyword | 74.03 | 86.11 | 73.37 |
| Rule-based template filling | 74.01 | 86.15 | 70.84 |
| Keyword's Wikipedia | 73.94 | 86.07 | 69.68 |
| Keyword + Annotation | 74.63 | 86.85 | 79.96 |

Finally, we compare MRC-MNER with other MNER models. It is clear to observe that MRC-MNER achieves state-of-the-art performance on Twitter2017 dataset and competitive results on Twitter2015. Particularly, MRC-MNER yields a 0.6 improvement compared with the current best model MAF in terms of overall $F1$ on Twitter2017. At the same time, we present a variant of our model, MRC-MNER-VG, which replaces our query-guided VG model with a VG toolkit. The performance of MRC-MNER-VG drops on all metrics, but its results are still competitive with other models. The above results validate the effectiveness of our framework with query-guided VG and multi-level modal interaction.

## 4.3 Ablation Study

To show the effectiveness of each sub-task in MRC-MNER, we conduct ablation study by removing particular sub-task from it. As shown in Table 4, we can see that all sub-tasks in our MRC-MNER contribute significantly to the final results.

The discussion on the effectiveness of each sub-task is given with respect to two datasets. First, after removing the RWE (Region Weights Estimation) sub-task, the performance significantly drops on all metrics. In particular, $F1$ scores on these two datasets degrade by 0.72 and 0.71, respectively. This shows that the existence of RWE promotes effective interaction between image and text. Besides, removing the ED (Existence Detection) sub-task also damages the performance on all metrics. $F1$ scores on the two datasets decrease by 0.49 and 0.56, respectively. This is because ED provides global information for the entire model, which can help the model determine whether the sentence contains certain entities asked by the query. Finally, after removing both RWE and ED, the performance of the model drops more significantly, indicating that both RWE and ED sub-tasks are essential in our framework. Overall, different sub-tasks of our model can work effectively with each other under multi-task training and enable the model to yield better performance for the MNER task.

## 4.4 Discussions

**Effect of Query Transformations.** To better validate the effect of MRC queries, we explore different ways to transform entity types to queries by utilizing the following expressions: (1) *Keyword:* An entity type keyword. The query for type LOC is "Location". (2) *Rule-based Template Filling:* Phrases generated by a simple template. The query for type LOC is "Please find location". (3) *Keyword's Wikipedia:* The definition of the entity type keyword from Wikipedia. The query for type LOC is "Location is used to denote a region (point, line, or area) on Earth's surface or elsewhere". (4) *Keyword+Annotation:* The concatenation of a keyword and its annotations. The query for type LOC is "Location: Country, city, town, continent by geographical location". Table 5 shows the experimental results on Twitter2015 and Twitter2017 by using different query transformations. Since the queries in our framework bridge the

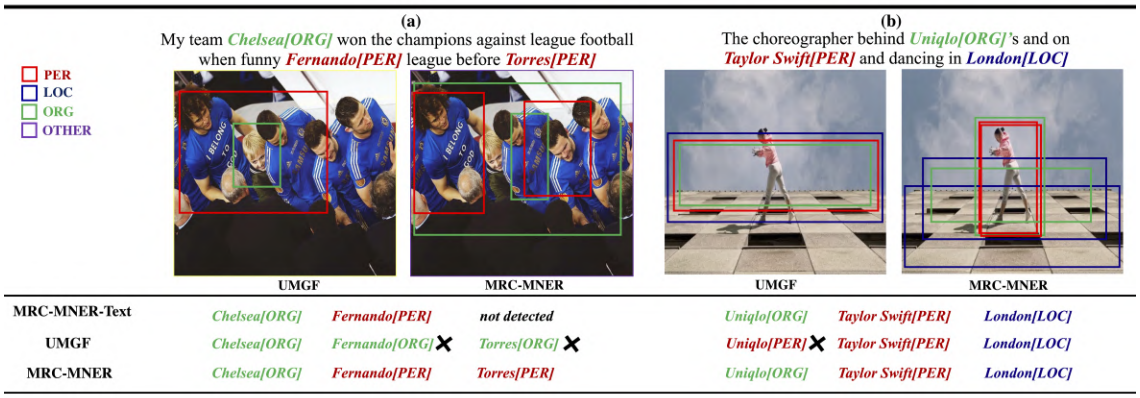


**Figure 4: Example comparison among MRC-MNER, MRC-MNER-Text, and UMGF.**

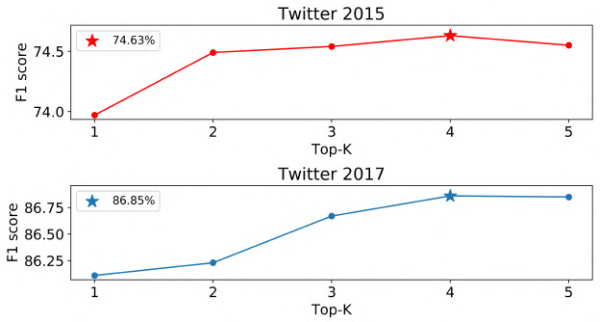


**Figure 5: Results with different numbers of region candidates.**

VG and MRC stages, we also replace original phrases with transformed queries, and provide the results of visual grounding on the Flickr30K Entities dataset. We find that the model with queries from *Keyword+Annotation* achieves the highest $F1$ score. The reasons are that queries constructed by *Keyword* and *Rule-based Template Filling* are relatively simple and contain less information, which results in friendly VG performance but limits the language understanding of MRC. For *Keyword's Wikipedia*, definitions from Wikipedia are relatively general and do not precisely describe the entity types, leading to inferior performance on MNER and VG tasks. Compared with other transformations, the framework with queries constructed by *Keyword+Annotation* achieves better results in both tasks. Therefore, we apply queries from *Keyword+Annotation* to MRC-MNER.

**Effect of the Number of Candidate Regions.** To check the influence of different numbers of region candidates, we set $k$ to several values and depict the results in Figure 5. First, fewer region candidates ($k = 1$) cannot provide sufficient image information for the model. With the increase of region candidates, image information can be supplemented and the model achieves the best result on both datasets when $k = 4$. Then more region candidates ($k > 4$) will bring some noise, leading to the degradation of the performance. The time spends on top-4 region candidates is 1.1 times more than that of top-1 region candidates.

## 4.5 Case Study

Here we conduct further qualitative analysis with two specific examples, in which MRC-MNER recognizes the entities correctly while the baseline models fail. We compare the results from MRC-MNER, MRC-MNER-Text, and the competitive model UMGF. The number of region candidates[3] equals to 2. In Figure 4 (a), the sentence contains three entities "*Chelsea*", "*Fernando*", and "*Torres*" with ORG, PER, and PER types respectively. However, the baseline UMGF mis-recognizes "*Fernando*" and "*Torres*" as ORG. We guess it is because UMGF cannot detect the region of person accurately (red box). Instead, MRC-MNER detects two regions (red boxes) for PER, and extracts a group of people as well as logos on clothing (green boxes) for ORG. This demonstrates the effectiveness of our query-guided visual grounding and multi-level modal interaction stages. Besides, because of the lack of auxiliary image information, MRC-MNER-Text ignores the entity "*Torres*" by mistake.

Figure 4 (b) illustrates a more challenging case, where the entity "*Uniqlo*" is ambiguous in the sentence, and the image cannot provide useful regions about ORG. It can be seen that both UMGF and MRC-MNER cannot locate the relevant visual regions for this entity correctly. However, both MRC-MNER and MRC-MNER-Text can recognize "*Uniqlo*" and label it as ORG. We conjecture that the solid understanding capability of MRC and the guidance of query prior information contribute to the final correct prediction.

## 5 CONCLUSION AND FUTURE WORK

In this work, we propose MRC-MNER, a framework for multi-modal entity recognition using machine reading comprehension. Our model bridges MRC and VG by designing queries with prior information of entity types to facilitate MNER task. We train a query-guided visual grounding model via transfer learning to promote fine-grained text-image alignments, and propose a multi-level modal interaction model to simulate text-image and inner-text relations. We find that the image contains a wealth of information. For future work, we will try to distill more useful parts from the image for multimodal named entity recognition.

[3]Our model achieves the best performance on both datasets when the number of region candidates is 4. However, to avoid a mess, we frame top-2 region candidates in the image.

## 6 ACKNOWLEDGEMENT

This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600.

## REFERENCES


[1] Omer Arshad, Ignazio Gallo, Shah Nawaz, and Alessandro Calefati. 2019. Aiding intra-text representations with visual context for multimodal named entity recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 337–342.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[3] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. 1870–1879.

[4] Dawei Chen, Zhixu Li, Binbin Gu, and Zhigang Chen. 2021. Multimodal Named Entity Recognition with Image Attributes and Image Knowledge. In *Database Systems for Advanced Applications - 26th International Conference (DASFAA)*. 186–201.

[5] Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*. 12666–12674.

[6] 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 (NAACL-HLT)*. 4171–4186.

[7] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. 2020. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 2777–2787.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 770–778.

[9] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[10] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. 260–270.

[11] Fei Li, Zheng Wang, Siu Cheung Hui, Lejian Liao, Dandan Song, and Jing Xu. 2021. Effective named entity recognition with boundary-aware bidirectional neural networks. In *Proceedings of the Web Conference 2021 (WWW)*. 1695–1703.

[12] Fei Li, Zheng Wang, Siu Cheung Hui, Lejian Liao, Dandan Song, Jing Xu, Guoxiu He, and Meihuizi Jia. 2021. Modularized Interaction Network for Named Entity Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*. 200–209.

[13] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2019. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 5849–5859.

[14] 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 Conference of the Association for Computational Linguistics (ACL)*. 1340–1350.

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2117–2125.

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

[17] 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 (ACL)*. 1990–1999.

[18] Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[19] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 852–860.

[20] Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Answering while summarizing: Multi-task learning for multi-hop QA with evidence extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*. 2335–2345.

[21] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational (NAACL)*. 2227–2237.

[22] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision (ICCV)*. 2641–2649.

[23] Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021. A co-interactive transformer for joint slot filling and intent detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8193–8197.

[24] Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. 6140–6150.

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 779–788.

[26] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *AAdvances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015 (NIPS)*. 91–99.

[28] Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing (EMNLP)*. 1524–1534.

[29] Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. In *9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 173–179.

[30] Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9073–9080.

[31] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *International journal of computer vision* 104, 2 (2013), 154–171.

[32] Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. MAF: A General Matching and Alignment Framework for Multimodal Named Entity Recognition. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM)*. 1215–1223.

[33] Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics, (COLING)*. 3879–3889.

[34] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4683–4693.

[35] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2369–2380.

[36] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics (TACL)*. 67–78.

[37] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 3342–3352.

[38] Xuehui Yu, Pengfei Chen, Di Wu, Najmul Hassan, Guorong Li, Junchi Yan, Humphrey Shi, Qixiang Ye, and Zhenjun Han. 2022. Object Localization under Single Coarse Point Supervision. *arXiv preprint arXiv:2203.09338* (2022).

[39] Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 14347–14355.

[40] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*. 5674–5681.

[41] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating object proposals from edges. In *European conference on computer vision (ECCV)*. 391–405.